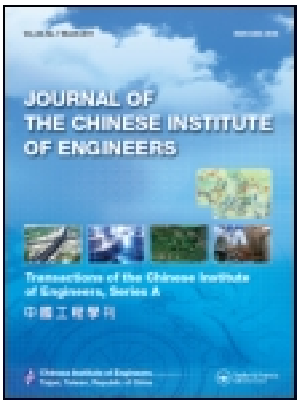

Journal of the Chinese Institute of Engineers
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tcie20

Stack robust fine granularity scalable video coding

Hsiang-Chun Huang [a] & Tihao Chiang [b]

[a] Department and Institute of Electronics Engineering , National Chiao Tung University , Hsinchu 300, Taiwan

[b] Department and Institute of Electronics Engineering , National Chiao Tung University , Hsinchu 300, Taiwan Phone: 886-3-5712171 ext. 54135 Fax: 886-3-5712171 ext. 54135 E-mail:

Published online: 04 Mar 2011.

STACK ROBUST FINE GRANULARITY SCALABLE VIDEO CODING

Hsiang-Chun Huang and Tihao Chiang*

ABSTRACT

A novel scalable video coding technique, namely Stack Robust Fine Granularity Scalability (SRFGS), is presented to provide both temporal and SNR scalability. The SRFGS first simplifies the temporal prediction architecture of RFGS. The approach is further generalized using a reconstructed frame from the previous time instance of the same layer to temporally predict the quantization error of the lower layer. With this concept, the RFGS architecture can be extended to multi-layer stack architecture. The SRFGS can be optimized at several operating points to meet the requirements of various applications, while maintaining the fine granularity and error robustness of RFGS. An optimized macroblock-based alpha adaptation scheme is proposed to improve the coding efficiency. A single-loop enhancement layer decoding scheme is proposed to reduce the decoder complexity. The simulation results show that SRFGS can improve the performance of RFGS by 0.4 to 3.0 dB in PSNR. SRFGS has been reviewed by the MPEG committee and ranked as one of the best algorithms according to subjective testing in the Report on Call for Evidence on Scalable Video Coding.

Key Words: scalable video coding (SVC), advance video coding (AVC), fine granularity scalability (FGS).

I. INTRODUCTION

Scalable video coding (SVC) has received more attention with the rapidly growth of multimedia applications over Internet and wireless channels. For such applications, the video information may be transmitted over error-prone channels with fluctuating bandwidth and can be transported through different networks to diverse devices. To serve multimedia applications in a heterogeneous environment, the MPEG-4 committee has developed Fine Granularity Scalability (FGS) (ISO Standard 14496-2 FDAM4, 2001) that provides a DCT-based scalable approach in a layered fashion. The base layer is coded by a non-scalable MPEG-4 advanced simple profile (ASP) while the enhancement layer is intra coded with embedded bit plane coding to achieve fine granular scalability. The lack of temporal prediction at the FGS enhancement layer leads to inherent robustness at the expense of coding efficiency. Several research works are proposed to improve temporal prediction efficiency while keeping the features of fine granularity and robustness of MPEG-4 FGS, as discussed by Huang *et al*. (2002a). Among these approaches, the Robust FGS (RFGS) multiplies the temporal prediction information by a leaky factor α, where $0 \leq \alpha \leq 1$, to strengthen the error resilience and leads to a good tradeoff between coding efficiency and error robustness. Aside from the SVC technologies that are DCT-based and have temporal prediction feedback loops, there is another effective approach, namely three-dimensional (3-D) subband/wavelet coding using a motion compensated temporal filter (MCTF) (Woods and Chen, 2002). 3-D wavelet coding uses the MCTF to exploit the temporal correlations of neighboring frames and applies the wavelet transform in the spatial domain. In addition, 3-D wavelet coding can be used to generate fully embedded bitstreams in spatial, temporal and SNR resolutions.

*Corresponding author. (Tel: 886-3-5712171 ext. 54135; Fax: 886-3-5724361; Email: tchiang@mail.nctu.edu.tw)

The authors are with the Department and Institute of Electronics Engineering National Chiao Tung University, Hsinchu 300, Taiwan.

To verify the improvement from the new SVC techniques after the MPEG-4 FGS, the MPEG committee issued a Call for Evidence on Scalable Video Coding (CFE on SVC) (ISO Document N5559, 2003). Stack Robust Fine Granularity Scalability (SRFGS) has been shown to improve the temporal prediction efficiency of RFGS and provides temporal and SNR scalability. The SRFGS was reviewed by the MPEG committee in the work "Huang *et al.*, 2003" and ranked as one of the best algorithms according to subjective tests in "ISO Document N5701, 2003". Recently, the scalable extension of H.264/AVC (JVT Document R202, 2006) also utilizes the "stack" concept in the closed-loop hierarchical B pictures (Schwarz *et al.*, 2005) to improve the coding efficiency.

This paper is organized as follows. In Section II, we propose a simplified RFGS architecture. It significantly reduces the complexity of the RFGS architecture while maintaining the same performance. It leads to easier understanding of the basic prediction concept used in the RFGS enhancement layer. Based on the simplified architecture, in Section III, the prediction concept of SRFGS is introduced. Section IV shows the detailed encoder and decoder structures of SRFGS. To optimize the coding efficiency of SRFGS, a novel macroblock-based alpha adaptation and the prediction architecture for the B frames are discussed. Single-loop enhancement layer decoder architecture is proposed to reduce the complexity of the SRFGS decoder. In Section V, the simulation results demonstrate the improvement of SRFGS as compared to RFGS. The comparison with AVC is also shown. Finally, the conclusions are given in Section VI.

II. SIMPLIFIED RFGS PREDICTION SCHEME

Figure 1 shows the original RFGS encoder architecture as proposed by Huang *et al.* (2002a) and Huang *et al.* (2002b). The enhancement layer bitstream is generated with the following process. The motion compensation module of the enhancement layer uses the base layer motion vectors and the high quality reference image *HQRI* stored in the enhancement layer frame buffer to generate the high quality prediction image *ELPI*. The enhancement layer motion compensated frame difference $MCFD_{EL}$ is computed by subtracting *ELPI* from the original signal *F*:

$$MCFD_{EL,\,i} = F_i - ELPI_i = F_i - (HQRI_{i-1})_{mc}, \quad (1)$$

where the subscripts *i* and *i* –1 mean the current frame time *i* and the previous frame time *i* –1, respectively. The subscript *mc* means that $(y)_{mc}$ is the motion compensated version of *y*. The signal $\hat{D}$ is computed by subtracting the reconstructed base layer DCT coefficients $\hat{B}$ from the $MCFD_{EL}$:

$$\hat{D}_i = MCFD^i_{EL} - \hat{B}_i. \quad (2)$$

The signal $\hat{D}$ is entropy encoded to generate the enhancement layer bitstream. Note that for simplicity and also due to the linearity of DCT, in this paper we use same notation for the symbol in spatial and transform domain.

The high quality reference image *HQRI* at the enhancement layer is generated as follows. The first β bit planes of the difference signal $\hat{D}$ are summed with $\hat{B}$. The resultant signal is converted back to the spatial domain using the IDCT transform and summed with *ELPI* to get the enhancement layer reconstructed image *ELRI*.

$$ELRI_i = (HQRI_{i-1})_{mc} + \hat{B}_i + \hat{D}_i. \quad (3)$$

It should be noted that for simplicity we assume all of the bit planes in $\hat{D}_i$ will be used in the enhancement layer prediction loop. The base layer reconstructed signal *B* will be subtracted from the signal *ELRI* to get the signal *D* with only enhancement layer information. The signal *D* will be attenuated by a leak factor α and be added back into the signal *B* before storage in the enhancement layer reference frame buffer. Thus, we have the following relationship:

$$HQRI_i = B_i + \alpha D_i \quad (4)$$

The rationale for performing the attenuation process on the signal *D* is that we want the errors to be attenuated for all the past frames recursively. If the attenuation process is only applied to the first few bit planes of $\hat{D}$, only the errors occurring in the current frame are attenuated. The earlier errors are still accumulated for subsequent frames through the motion prediction loop without attenuation.

Although the RFGS prediction architecture efficiently reduces drift error, it is quite complex. The base layer needs to store the reconstructed DCT coefficient $\hat{B}$. The enhancement layer first subtracts $\hat{B}$ from the prediction error $MCFD_{EL}$ to reduce the entropy in the signal $\hat{D}$, and then it uses $\hat{B}$ to form the *ELRI*. The enhancement layer further accesses the base layer reconstructed image *B* to generate the signal *D* with only the enhancement layer information and to generate the *HQRI* stored in the enhancement layer frame buffer. This prediction scheme increases requirements for both memory and memory access bandwidth. Further, with this complex prediction architecture, the prediction concept of RFGS is difficult to grasp or improve. Thus, we will simplify the prediction scheme while maintaining the same

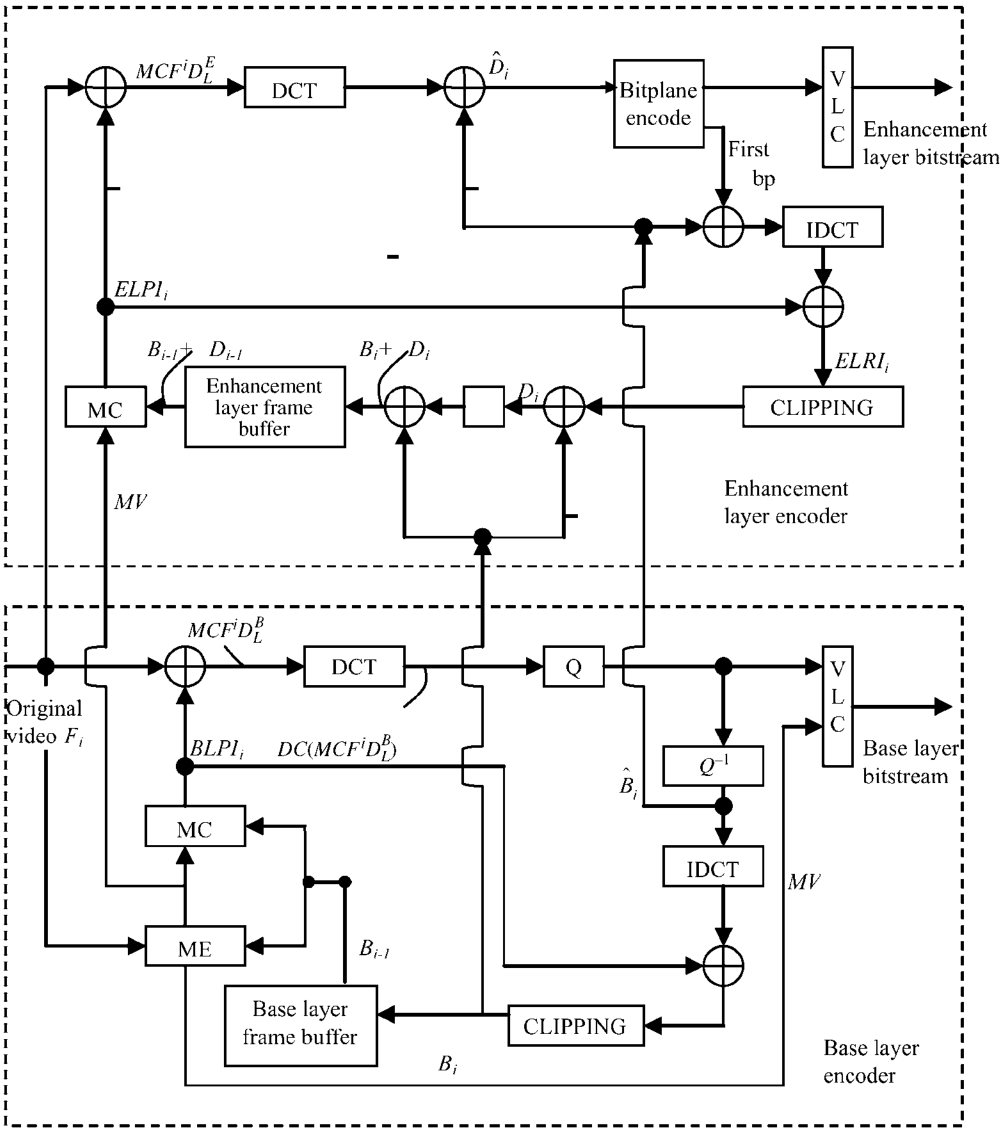

Fig. 1 The original RFGS encoder

coding efficiency. From Eqs. (3) and (4), we can get the following relationship:

$$ELRI_i = (B_{i-1} + \alpha D_{i-1})_{mc} + \hat{B}_i + \hat{D}_i. \tag{5}$$

By grouping the base layer information and the enhancement layer information, Eq. (5) becomes

$$ELRI_i = (B_{i-1})_{mc} + \hat{B}_i + (\alpha D_{i-1})_{mc} + \hat{D}_i = B_i + D_i, \tag{6}$$

where

$$B_i = (B_{i-1})_{mc} + \hat{B}_i \tag{7}$$

and

$$D_i = (\alpha D_{i-1})_{mc} + \hat{D}_i. \tag{8}$$

From (8) we know that the residue D can be derived simply from accumulating the signal $\hat{D}$ in all the previous frames. From Eqs. (1) and (4), we can re-write the derivation of the signal $\hat{D}_i$ in (2) as:

$$\begin{aligned}\hat{D}_i &= MCFD_{EL-i} - \hat{B}_i \\ &= F_i - (HQRI_{i-1})_{mc} - \hat{B} \\ &= F_i - (B_{i-1} + \alpha D_{i-1})_{mc} - \hat{B}.\end{aligned} \tag{9}$$

Again, by grouping the base layer information and the enhancement layer information, Eq. (9) becomes

$$\begin{aligned}\hat{D}_i &= F_i - (B_{i-1})_{mc} - \hat{B} - (aD_{i-1})_{mc} \\ &= F_i - B_i - (\alpha D_{i-1})_{mc}.\end{aligned} \tag{10}$$

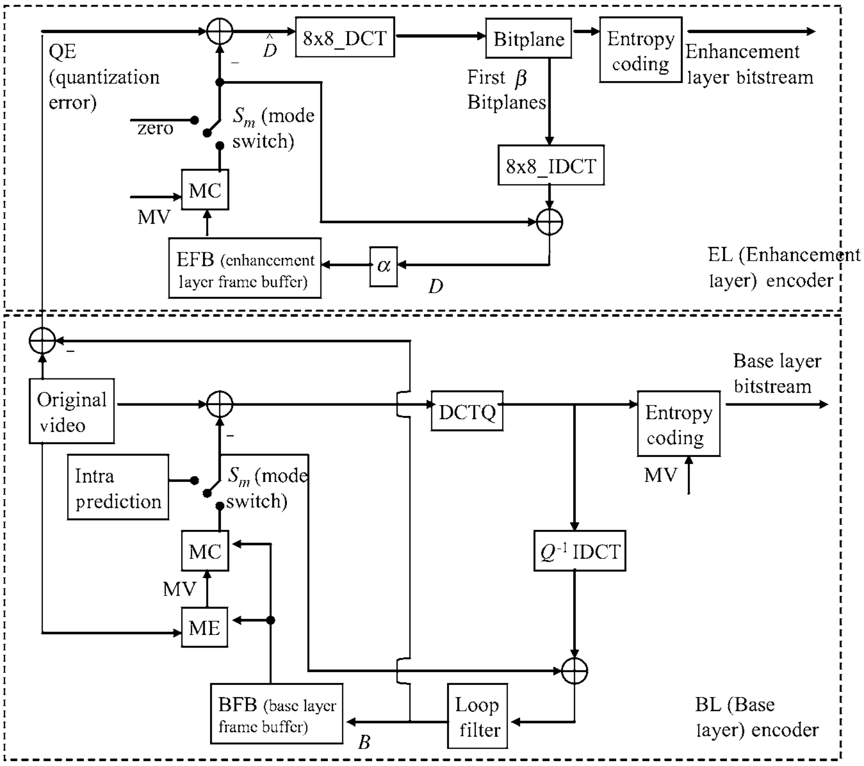

Fig. 2 The simplified RFGS encoder

The difference between the original frame F and the base layer reconstructed image B is actually the quantization error QE at the base layer,

$$QE_i = F_i - B_i. \tag{11}$$

Thus, Eq. (10) becomes

$$\hat{D}_i = QE_i - (\alpha D_{i-1})_{mc}. \tag{12}$$

From (8) and (12), we realize that the only signal that the enhancement layer acquires from the base layer is the base layer quantization error QE, all the other signals can be generated by the enhancement layer itself. With this analysis, we can derive a simplified RFGS prediction scheme as shown in Fig. 2, and it still provides identical functionality with the original RFGS prediction scheme as shown in Fig. 1. In the simplified architecture, the base layer quantization error QE will be predicted with the reference frame stored in the enhancement layer frame buffer EFB. This step performs the Eq. (12) in Fig. 1. The prediction error $\hat{D}_i$ will be transformed and bit plane coded as FGS bitstreams. The first β bit planes will be inversely transformed and added to the prediction to generate the signal D. This step performs Eq. (8) in Fig. 1. The resultant signal D will multiply by α for leaky prediction before it is stored in the frame buffer. The simplified RFGS architecture significantly reduces the complexity of the RFGS. The base layer encoder need not store the reconstructed base layer DCT coefficient $\hat{B}$. The enhancement layer encoder need not access the coefficient and performs the computation with the base layer signal $\hat{B}$ and B. The enhancement layer encoder architecture is just like the base layer encoder replacing the original signal from F with the base layer quantization error QE.

III. ENHANCED PREDICTION ARCHITECTURE USING STACK CONCEPT

With the simplified RFGS architecture, it is also easier to understand the prediction concept within the RFGS structure. In the RFGS structure, the base layer quantization error QE, which is intra coded in the MPEG-4 FGS scenario, is temporally predicted by the previous enhancement layer information to remove the temporal redundancy. The leaky factor α is used to attenuate the drift error on the decoder side when only partial enhancement layer reference information is reconstructed. A smaller leaky factor α leads to less drift. However, smaller α leads to poorer performance when all of the reference enhancement layer information is received but only partial information is used for removing temporal redundancy. The other factor β, which denotes the number of bit planes used in the enhancement layer prediction loop, plays a key role in the RFGS structure, too. A larger β leads to more enhancement layer information used for temporal prediction. With the removal of more temporal redundancy, larger β provides better performance when all the reference bit planes are fully reconstructed. However, larger β may lead to larger drift error at lower bitrate as less of the required reference information is available for motion compensation. In summary, a smaller β reduces the drift at lower bitrate at the expense of coding efficiency because the bit planes after β effectively become intra-coded with poorer coding performance.

To address temporal redundancy removal and drift reduction, a novel architecture, namely Stack RFGS (SRFGS), is proposed. In the SRFGS, the prediction scenario is generalized from that of RFGS as follows: The quantization error of the previous layer is temporally predicted by the reconstructed frame in the previous time instance of the current layer. We utilize this generalized prediction concept and further extend the architecture to multiple layers in SRFGS as illustrated in Fig. 3. At time instant i, the original Frame F_i is predicted by the base layer reconstructed frame of time i-1, which is denoted as B_{i-1}. The quantization error $QE_{A,\,i}$ is computed as the difference between F_i and the reconstructed base layer B_i. The signal $QE_{A,\,i}$ is predicted by the first enhancement layer reconstructed frame at time instant i-1, which is $D_{A,\,i-1}$. At the second layer EL_B, the quantization error $QE_{B,\,i}$ is computed as the difference between $QE_{A,\,i}$ and the reconstructed first enhancement layer $D_{A,\,i}$. The signal $QE_{B,\,i}$ will be predicted by the second enhancement layer reconstructed frame at time i-1, which is $D_{B,\,i-1}$. With this concept, the RFGS enhancement layer prediction scheme is generalized to multi-layer stack architecture. The coding performance of EL_A in SRFGS is the same as the first β bit planes in RFGS, since the temporal redundancy has been removed in both of them. However, the coding performance in EL_B (and all the following layers) of SRFGS is superior to the remaining bit planes of RFGS, because the temporal redundancy is only removed in SRFGS.

IV. THE STACK RFGS SYSTEM ARCHITECTURE

In this section we first describe the encoder and decoder block diagrams of the SRFGS architecture. An optimized macroblock-based alpha adaptation is then introduced to increase the coding performance. The prediction scheme for the B-frame is described, too. We further propose a single-loop enhancement layer decoder architecture to reduce the SRFGS decoder complexity.

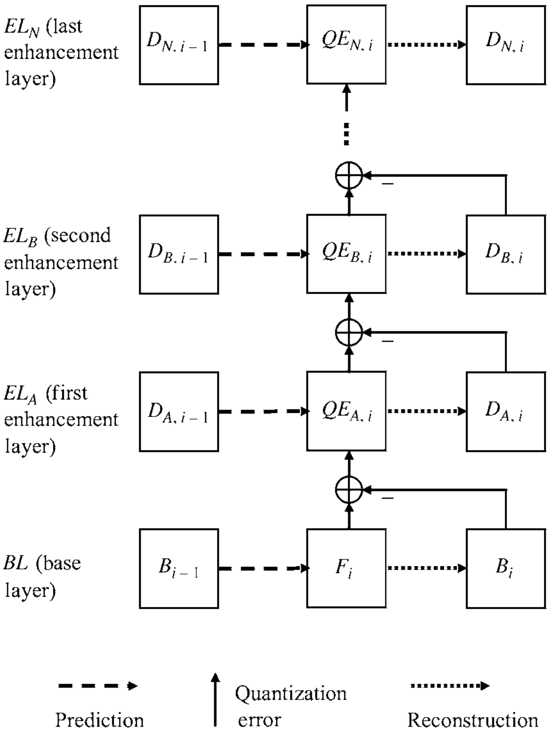

Fig. 3 SRFGS prediction concept

1. Functional Description

Based on the stack concept, the AVC-based SRFGS encoder in Fig. 4 is constructed. The SRFGS base layer prediction scheme is the same as that in RFGS, except that there is no high quality base layer reference in SRFGS. The high quality base layer reference will not be used in the AVC-based SRFGS architecture to prevent drift at low bitrate. The first enhancement layer of SRFGS, as denoted as EL_A, is identical to that in RFGS except in two aspects. First, only the first β_A bit planes are coded and written into the enhancement layer bitstream. Second, the multiplication of the leaky factor α_A is moved after the motion compensation module. All the enhancement layer loops have identical architecture to that in EL_A, except the last enhancement layer loop EL_N. In EL_N, the entire residues are bit plane coded to achieve perfect reconstruction at the decoder.

A scheme similar to the improved motion estimation algorithm by He *et al.* (2001) is utilized in SRFGS. He *et al.* (2001) derive a motion vector that is adequate for both the base and the enhancement layer information. Based on this improved ME algorithm, the base and entire enhancement layer information is embedded into the stack architecture. With the derived motion vector through the improved ME module, the base layer mode decision module selects the best mode using the AVC mode decision

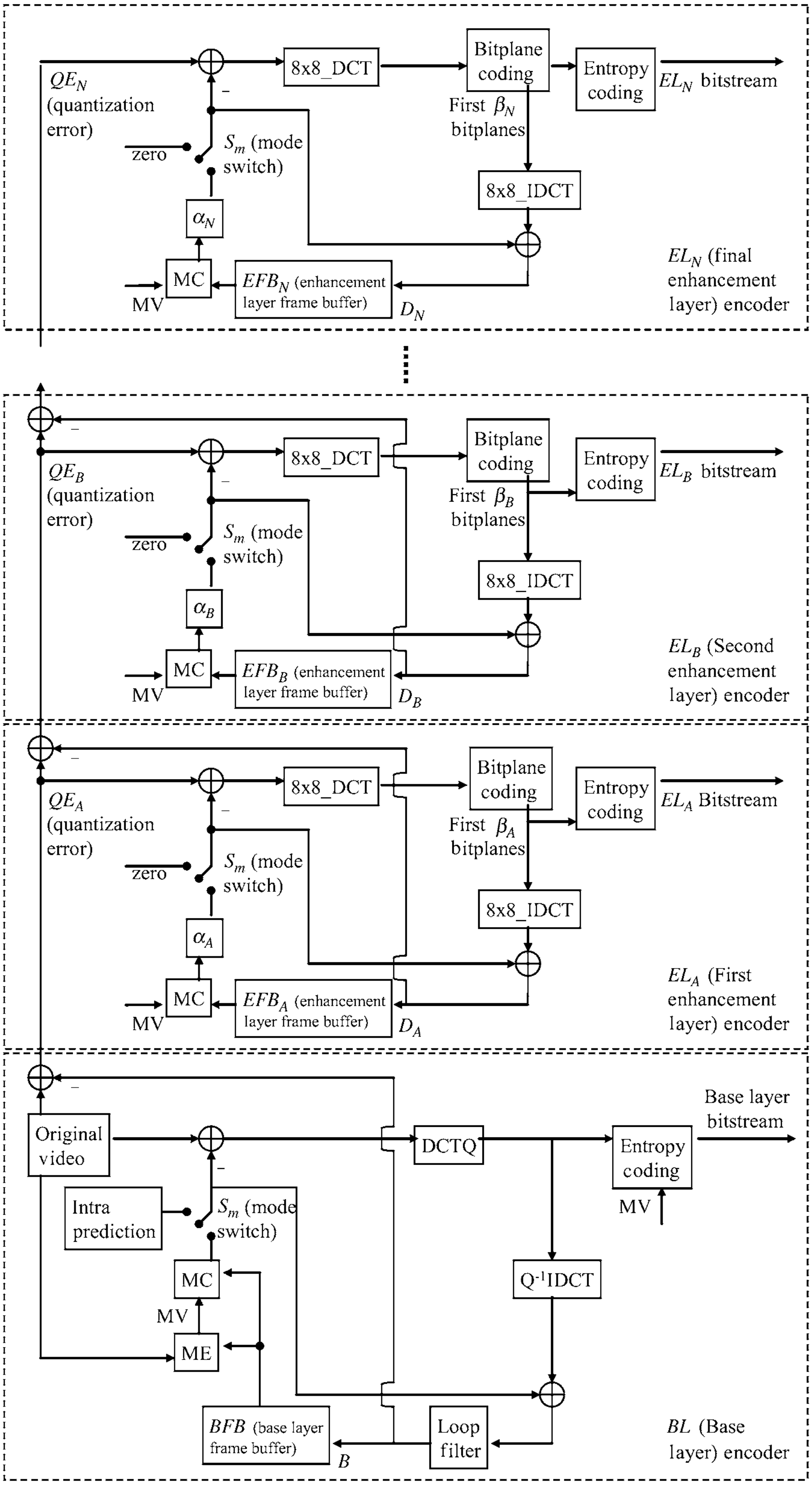
QE_N (quantization error)
8x8_DCT
Bitplane coding
Entropy coding
EL_N bitstream
First β_N bitplanes
zero
S_m (mode switch)
8x8_IDCT
α_N
MC
EFB_N (enhancement layer frame buffer)
MV
D_N
EL_N (final enhancement layer) encoder
QE_B (quantization error)
EL_B bitstream
First β_B bitplanes
α_B
EFB_B (enhancement layer frame buffer)
D_B
EL_B (Second enhancement layer) encoder
QE_A (quantization error)
EL_A Bitstream
First β_A bitplanes
α_A
EFB_A (enhancement layer frame buffer)
D_A
EL_A (First enhancement layer) encoder
Original video
Base layer bitstream
DCTQ
Intra prediction
ME
Q^{-1}IDCT
BFB (base layer frame buffer)
B
Loop filter
BL (Base layer) encoder

Fig. 4 Diagram of the SRFGS encoder framework

algorithm. Consequently, the same coding mode and motion vector are used for the base and entire enhancement layer prediction loops.

On the decoder side, as shown in Fig. 5, the received information of each loop will be decoded by its own loop and summed with the base layer

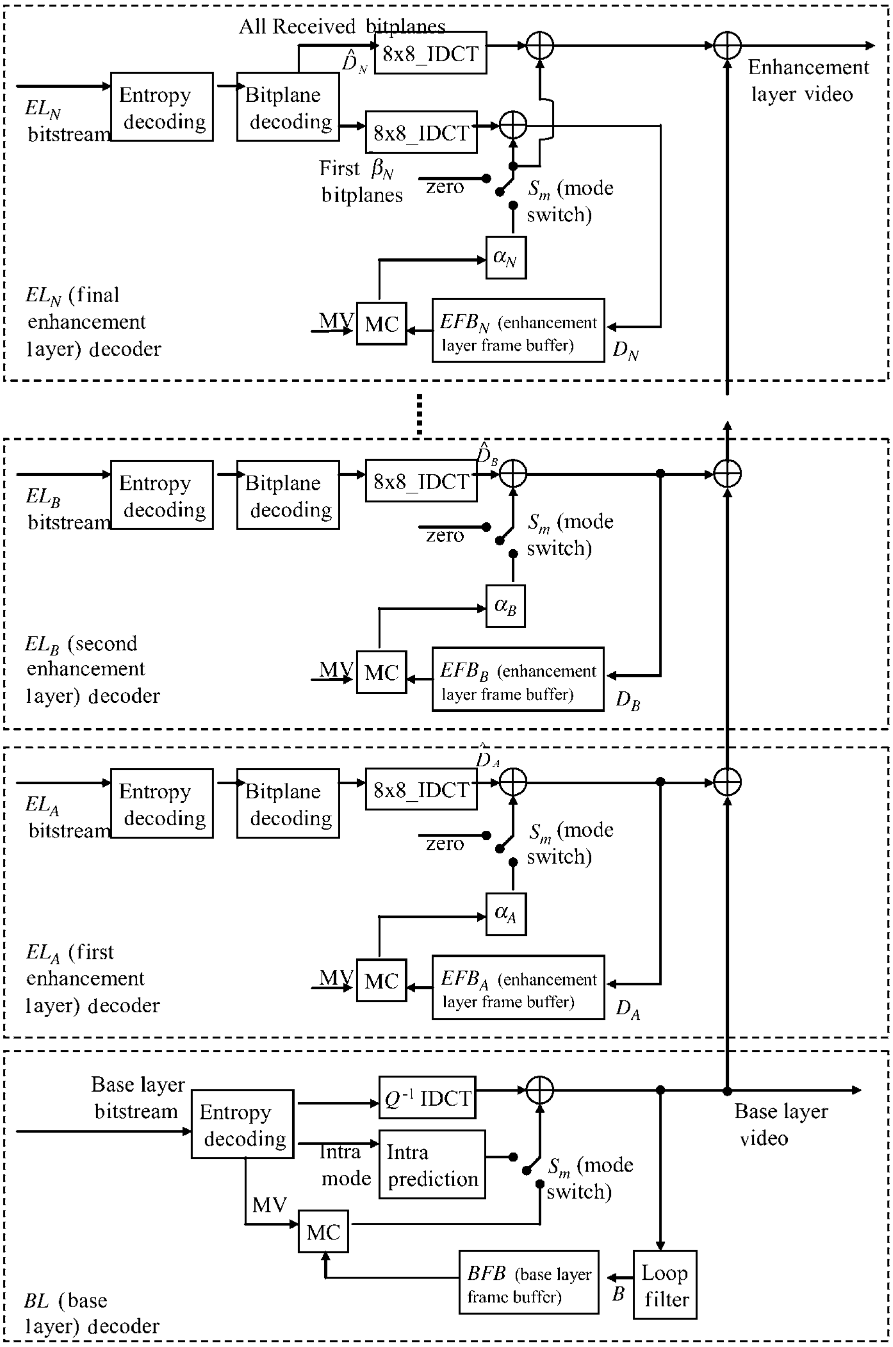

Fig. 5 Diagram of the SRFGS decoder framework

reconstructed image to construct the final image. For each loop, if only a partial bitstream is received, the leaky factor α can attenuate the drift error as in the RFGS case. If there is no information received for a loop, the leaked motion compensated information will directly be stored back to the frame buffer. In the proposed framework, the information of each prediction loop is not used or affected by the information in the other loops. Consequently, if there is any error in a loop, it won't affect the data in the other loops. This intrinsic error localization property of SRFGS offers better performance in an error-prone environment.

More enhancement layer loops usually lead to better coding performance. This sometimes may not be true because the temporal prediction not only reduces the energy of quantization error but also increases the dynamic range with some extra sign bits.

Frame Startcode	EL_A Header	EL_A 1st BP	···	EL_A β_Ath BP	EL_B Startcode & Header	EL_B 1st BP	···	EL_B β_Bth BP	·········	EL_N Startcode & Header	EL_N 1st BP	···	EL_N Last BP

Fig. 6 The SRFGS enhancement layer bitstream format

To overcome this drawback, the size of the enhancement layer loop should be large enough, such that the residue energy reduced from the temporal prediction is larger than the overhead. Note that usually the higher the enhancement layer, the more random the residue. To reduce the same amount of residue energy from the temporal prediction, we need more reference data (larger β) at a higher enhancement layer. Further, static sequences which have more temporal correlation and hence fewer reference data (smaller β) are enough to overcome the overhead. In this case (a static sequence), smaller β also reduces drift error at low bitrate. After determining the size of an enhancement layer based on its position and sequence characteristics, the same process can be used to set the size of the next enhancement layer if the bitrate range of the target application is not fully covered yet.

Figure 6 shows the enhancement layer bitstream format of the SRFGS coding scheme in a frame. Assuming that there are N enhancement layer loops, the bitstream first stored all the β_A bit planes of EL_A, which is the most significant loop. After β_A, we include all the β_B bit planes of EL_B, which is the second most significant loop. Similar processes are applied to code the remaining enhancement layers except EL_N. In EL_N, which is the last significant loop, not only the first β_N bit planes but also all the remaining bit planes are stored in the bitstream. Within each loop, the bit planes are ordered from MSB to LSB. Thus, the SRFGS bitstream is ordered by the importance of the information. With the bitstream, the SRFGS server, operating in similar fashion as the MPEG-4 FGS and RFGS server, can truncate the bitstream at any point to provide the best performance for that bitrate.

2. Optimized Macroblock-Based Alpha Adaptation

In the RFGS architecture, the value of α is adapted at frame level. Each macroblock in the same frame uses the same α. In this paper, we generalize the α adaptation to macroblock level with simple optimization. The optimization is performed such that the handling macroblock has the least prediction error energy. As shown in Fig. 4, the multiplication of α is placed after the motion compensation module. If the handling macroblock is selected as inter mode in the base layer mode decision module, the encoder will sweep the value of α between 0 and 1 to find the optimal value that minimizes the energy of the prediction error. Thus, we can find the best α for the handling macroblock in a very simple way. However, various values of α, coded in the macroblock header, have significant overheads. In our approach, we further define a frame level α named frame_α. The frame_α is adapted at the frame level and uniquely coded at the header for each loop. Each macroblock can select the best α between 0 and frame_α. Thus for each macroblock, only one-bit flag is needed to indicate whether 0 or frame_α is used. In our simulation, this method provides a good tradeoff between energy and overhead reduction.

3. Prediction Scheme of B-Frame

The prediction scheme of B-frame in SRFGS is similar to that in RFGS. In RFGS, the base layer of B-frame is predicted by a high quality reference image that is the sum of the base and enhancement layer reconstructed images, denoted as $B + D$. In the SRFGS structure, the B-frame is predicted by the sum of the base and the entire enhancement layer reconstructed images, which is $B + D_A + \cdots + D_N$. The quantization error, which is the difference between the original and base layer reconstructed frames, is coded as the enhancement layer bitstream. There is no stack architecture in B-frame to reduce the complexity. Since no frame takes B-frame as reference, missing B-frame in the FGS server can support temporal scalability without any drift error for the following frames. The rate control algorithm allocates more bits for the P-frame at low bitrate to provide a better anchor frame. With this bit allocation, we can reduce the drift error of P-frame but also enhance the reference image quality of B-frame. The extra bits at high bitrate will be allocated to B-frames since the information carried by the MSB of B-frame is more important than that carried by the LSB in P-frame for averaged picture quality of reconstructed video.

4. Stack RFGS with Single-Loop Enhancement Layer Decoder

Although the stack architecture improves the enhancement layer coding efficiency, it also significantly

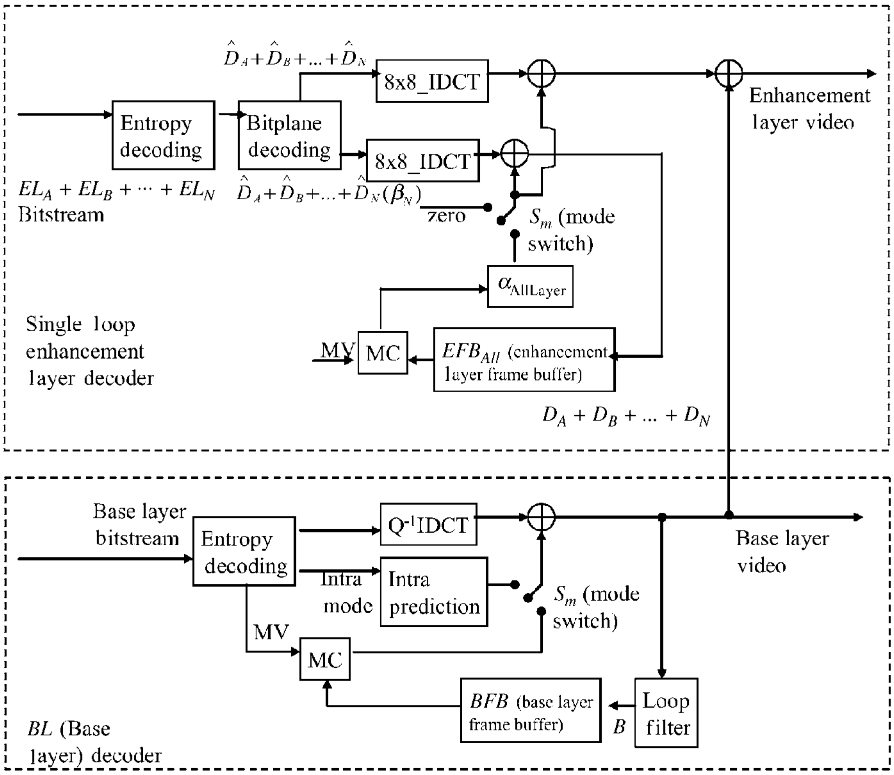

Fig. 7 Diagram of the SRFGS single-loop enhancement layer decoder framework

increases the complexity due to multiple loops. This is critical for a portable client device which is constrained by complexity and power. To address this issue, we propose a novel simplified SRFGS decoder that only requires single-loop enhancement layer decoding. Similar to Eq. (8), at each SRFGS enhancement layer decoder, the reconstructed information at that layer can be derived as:

$$D_{X,\,i} = (\alpha_{X,\,i-1} D_{X,\,i-1})_{mc(mv_{X,\,i-1})} + \hat{D}_{X,\,i}, \qquad (13)$$

where X denotes the enhancement layer X. The signal $(y)_{mc(mv_{X,\,i-1})}$ denotes the motion compensated version of y using the motion vector $(mv_{X,\,i-1})$. In the current SRFGS structure, the motion vector of each layer is identical to that in the base layer. If we further constrain the encoder with the same leaky factor α for each layer, Eq. (13) can be simplified as

$$D_{X,\,i} = (\alpha_{AllLayer,\,i-1} D_{X,\,i-1})_{mc(mv_{AllLayer,\,i-1})} + \hat{D}_{X,\,i}. \qquad (14)$$

That is, the signal D in each layer is attenuated with the same leaky factor $\alpha_{AllLayer}$, and then motion compensated by the same motion vector $(mv_{AllLayer,\,i-1})$. With this constraint, we need not separate the signal D for each layer and can merge them all. Thus, the Eq. (14)) of multiple layers can be merged as:

$$\begin{aligned}&(D_{A,\,i} + D_{B,\,i} + \cdots + D_{N,\,i})\\ &= (\alpha_{AllLayer,\,i-1}(D_{A,\,i-1} + D_{B,\,i-1} + \cdots\\ &\quad + D_{N,\,i-1}))_{mc(mv_{AllLayer,\,i-1})} + (\hat{D}_{A,\,i} + \hat{D}_{B,\,i} + \ldots\\ &\quad + \hat{D}_{N,\,i}).\end{aligned} \qquad (15)$$

This can be further simplified as:

$$\begin{aligned}&D_{AllLayer,\,i}\\ &= (\alpha_{AllLayer,\,i-1} D_{AllLayer,\,i-1})_{mc(mv_{AllLayer,\,i-1})}\\ &\quad + \hat{D}_{AllLayer,\,i},\end{aligned} \qquad (16)$$

where

$$D_{AllLayer,\,i} = (D_{A,\,i} + D_{B,\,i} + \cdots + D_{N,\,i}) \qquad (17)$$

and

$$\hat{D}_{AllLayer,\,i} = (\hat{D}_{A,\,i} + \hat{D}_{B,\,i} + \cdots + \hat{D}_{N,\,i}) \qquad (18)$$

More precisely, for the latest enhancement layer N only the first β_N bit planes are combined with the information in other layers. In the above equation we have not shown this detail for the sake of simplicity. Fig. 7 shows this simplified SRFGS

Table 1 The value of (α, β) used in the simulation

The value of beta is the number of referenced bits.

(α, β)	Tempete	Bus	Container
Stack 0	(0.7500, 24320)	(0.9375, 17067)	(0.7500, 24320)
Stack 1	(0.7500, 78000)	(0.9375, 51200)	(0.7500, 58860)
Stack 2	N/A	N/A	(0.7500, 92160)

decoder. All the enhancement layer decoding loops are merged into a single loop. The entropy and bit plane decoding modules receive and decode the bitstreams for each layer, and merge them, except the bit plane after β_N in layer N, into one transform coefficient. These merged transform coefficients in each block are inversely transformed to the spatial domain. Since the IDCT is a linear process, merging the transform coefficient of each layer before the IDCT leads to identical results when the order is reversed. In this way, we only need one IDCT for all enhancement layers. The resultant spatial domain image is summed with the attenuated prediction image of all enhancement layers to generate the reconstructed signal of all layers. The output signal is the sum of the base layer reconstructed signal and the entire enhancement layer reconstructed signal.

Obviously, the single-loop enhancement layer decoder significantly reduces the decoder complexity with the disadvantage of losing the flexibility to adjust α at each layer. When combined with the macroblock-based alpha adaptation, the collocated macroblocks at different layers need to use the same alpha, which may be 0 or frame_α. Except for the restriction of the alpha selection, the single-loop enhancement layer is identical to the original SRFGS decoder, and the error in each layer is still localized within its own layer although all the layers are merged.

V. EXPERIMENT RESULTS AND ANALYSES

The coding efficiency of the SRFGS is compared with RFGS, AVC and the scalable extension of H. 264/AVC (JVT Document R202, 2006). The test conditions adopt test 1c of CFE on SVC (ISO Document N5559, 2003) specified by the MPEG Scalable Video Coding Ad Hoc Group. The R-D curves of sequences including Tempete, Bus and Container in CIF resolution and YCbCr 4:2:0 format are compared at four bitrates/frame-rates. The frame rate is measured in frames per second. The four bitrates cover 128 kbps/ 15 fps, 256 kbps/15 fps, 512 kbps/30 fps, and 1024 kbps/30 fps. The coding performance of AVC is presented by Rusert and Wien (2003), where RD-optimized and CABAC modules are enabled. Quarter-pixel motion vector accuracy is employed with a search range of 32 pixels. Four reference frames are used. Only one I-frame is used at the beginning. The P-period is 3 in both 15fps and 30 fps. For the scalable extension of H.264/AVC (denoted as “SVC” in the following), the reference software version JSVM_4_6 is used in the simulation. The GOP size is 4 for the Bus sequence, and is 8 for the Tempete and Container sequences. Hierarchical-B GOP structure (Schwarz *et al*., 2005), RD-optimized mode decision, and arithmetic coding are used in the simulation. The bitstream extraction has utilized the quality layer proposed by Amonou *et al*. (2005). The reference frame number is one for the P-frame and is two for B-frame.

For RFGS and SRFGS, the base layer is JM42. The test conditions are identical to those used in AVC except that we have disabled RD-optimization and adopted only one reference frame. At 30 fps, the P-period is 6 for Tempete and Container sequences. The P-period is 4 for Bus sequence. At 15 fps, the P-period is half. The bit plane and entropy coding are as the same as that for the MPEG-4 FGS. In SRFGS, 2 enhancement layer loops are used for Tempete and Bus sequences and 3 enhancement layer loops are used for Container sequence. The detailed α and β used in the simulation are shown in Table 1. Note that regarding the value of β, we use the number of referenced bits instead of the number of referenced bit planes. A simple frame-level bit allocation with a truncation module is used in the streaming server. For various target bitrates, different bit allocations between P and B frames are tested and the one leading to the best RD-performance is used to get the final results. This bit allocation analysis is reasonable because it can be done once in company with the bitstream encoding, and provide the best bit allocation at various operating bitrates during streaming services.

The simulation results are shown in Fig. 8. Two RFGS results are shown, one has a lower reference bitrate (labeled as RFGS_L) and the other a higher reference bitrate (labeled as RFGS_H). The SRFGS has a performance similar to RFGS_L at low bitrate, and provides improvement by 1.7 to 3.0 dB in PSNR at high bitrates, because SRFGS can remove more temporal redundancy at high bitrate than RFGS_L. As compared with RFGS_H, the quality of SRFGS increases by 0.4 to 1.0 dB in PSNR at low bitrate

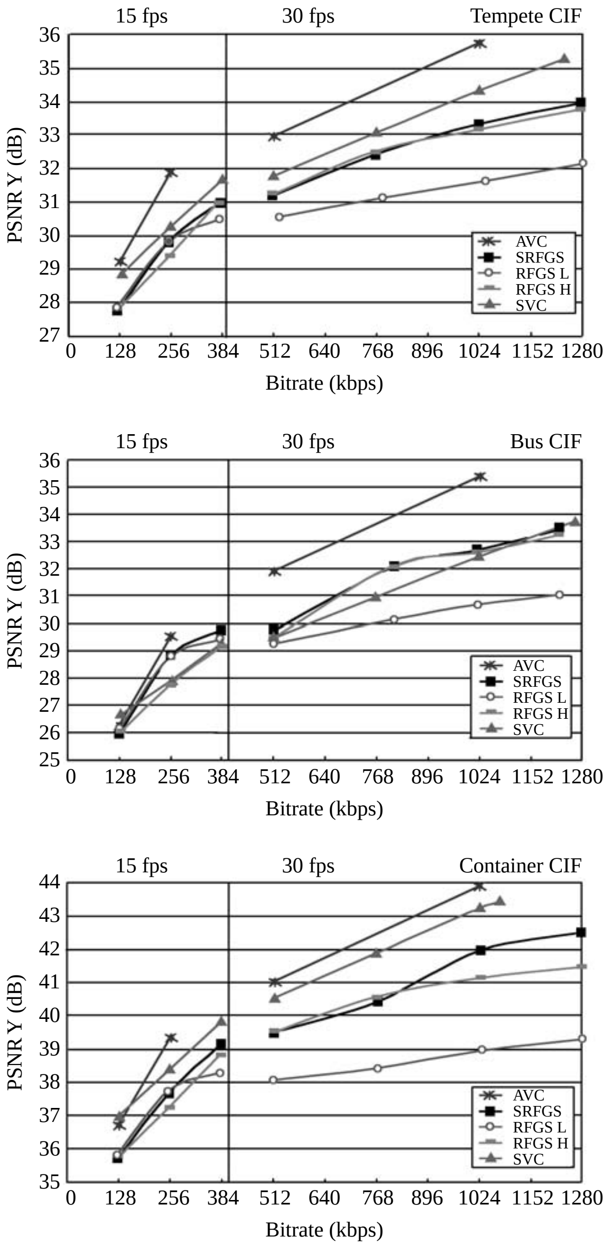

Fig. 8 PSNR versus bitrate comparison between SRFGS, RFGS and AVC coding schemes for the Y component.

because there is more drift error at low bitrate of RFGS_H. At high bitrate, the SRFGS increases 0.8 dB in PSNR for a low motion sequence such as Container and shows similar performance for a high motion sequence, such as Tempete and Bus. For the high motion sequence there is less temporal correlation so the performance of the improved prediction technique in SRFGS decreases. At medium bitrate, SRFGS has at most 0.15 dB PSNR losses than RFGS_H. This comes from the fact that the increased dynamic range and sign bits of each layer in SRFGS slightly lower the coding efficiency. The simulation results show that RFGS can only be optimized at one operating point and SRFGS can be optimized at several operating points, which can provide superior performance at a wider bandwidth. Compared to AVC, SRFGS has 0.4 to 1.5 dB PSNR loss at base layer. This is mainly because the MV in SRFGS is derived from both the base and enhancement layer information as described in Section VI.1. There is a 0.7 to 2.0 dB PSNR loss at low bitrates and a 2.0 to 2.7 dB loss at high bitrates. Compared with SVC, SRFGS has an up to 1.5 dB PSNR loss on Tempete and Container sequences, but has a 0.9 dB PSNR improvement on the Bus sequence. Note that SVC has incorporated the hierarchical-B structure, the RD-optimized mode decision, and the arithmetic coding. These tools can also be integrated in the SRFGS structure to improve the performance.

VI. CONCLUSIONS

In this paper, we have proposed a novel FGS coding technique named SRFGS. Based on RFGS, the SRFGS generalizes its prediction concept and structure to a multi-layer stack architecture. In each layer, the information to be coded is temporally predicted by the information of the previous time instance at the same layer. The stack concept allows the SRFGS to optimize at several operating points for various applications. With the bit plane coding and leaky prediction used in RFGS, SRFGS maintains the features of fine granularity and error robustness. An optimized MB-based alpha adaptation is proposed to improve the coding efficiency. We also propose single-loop enhancement layer decoding scheme to reduce the decoder complexity. The simulation results show that SRFGS provides improvement by 0.4 to 3.0 dB in PSNR over RFGS. Further investigation of the bit allocation for each layer for various types of video content can provide better coding efficiency.

REFERENCES

Amonou, I., Cammas, N., Kervadec, S., and Pateux, S., 2005, “Layered quality optimization of JSVM-3 when considering closed-loop encoding,” *17th meeting of Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG*, Nice, France, JVT-Q081.

He, Y., Yan, R., Wu, F., and Li, S., 2001, “H.26L-based fine granularity scalable video coding,” *58th Meeting of Moving Picture Expert Group (MPEG)*, Pattaya, Thailand, M7788.

Huang, H. C., Peng, W. H., Wang, C. N., Chiang, T., and Hang, H. M., 2003, “Stack Robust Fine Granularity Scalability: Response to Call for Evidence on Scalable Video Coding,” *65th Meeting of Moving Picture Expert Group (MPEG)*, Trondheim, Norway, M9767.

Huang, H. C., Wang, C. N., and Chiang, T., 2002a, "A Robust Fine Granularity Scalability Using Trellis Based Predictive Leak," *IEEE Transaction on Circuits and System for Video Technology*., Vol. 12, pp. 372-385.

Huang, H. C., Wang, C. N., Chiang, T., and Hang, H. M., 2002b, "H.26L-based Robust Fine Granularity Scalability (RFGS)," *61th Meeting of Moving Picture Expert Group (MPEG)*, Klagenfurt, Austria, M8604.

ISO Document N5559, 2003, "Call for Evidence on Scalable Video Coding Advances," *64th Meeting of Moving Picture Expert Group (MPEG)*, Pattaya, Thailand.

ISO Document N5701, 2003, "Report on Call for Evidence on Scalable Video Coding (SVC) Technology," *65th Meeting of Moving Picture Expert Group (MPEG)*, Trondheim, Norway.

ISO Standard 14496-2 FDAM4, 2001, "Streaming Video Profile – Final Draft Amendment (FDAM4)," International Organization for Standardization, Geneva, Switzerland.

JVT Document R202, 2006, "Joint Scalable Video Model JSVM-5," *18th Meeting of Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG*, Bangkok, Thailand.

Rusert, T., and Wien, M., 2003, "AVC Anchor Sequences for Call for Evidence on Scalable Video Coding Advances," *65th Meeting of Moving Picture Expert Group (MPEG)*, Trondheim, Norway, M9725.

Schwarz, H., Marpe, D., and Wiegand, T., 2005, "Comparison of MCTF and Close-Loop Hierarchical B Pictures," *16th Meeting of Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG*, Poznan, Poland, JVT-P059.

Wood, J. W., and Chen, P., 2002, "Improved MC-EZBC with Quarter-pixel Motion Vectors," *60th Meeting of Moving Picture Expert Group (MPEG)*, Fairfax, VA, USA, M8366.

Manuscript Received: Nov. 02, 2005
Revision Received: May 22, 2006
and Accepted: Jul. 20, 2006