FGS+: OPTIMIZING THE JOINT SNR-TEMPORAL VIDEO QUALITY IN MPEG-4 FINE GRAINED SCALABLE CODING

Raj Kumar Rajendran, Mihaela van der Schaar, Shih-Fu Chang*

{raj,sfchang}@ee.columbia.edu, mihaela.vanderschaar@philips.com

Columbia University, *Philips Research

ABSTRACT

To enable video transmission over heterogeneous wireless networks, a highly scalable compression and streaming framework that can adapt to large and rapid bandwidth variations in real-time is necessary. MPEG-4 Fine Grained Scalability (FGS) provides fine-grained SNR and temporal scalabilities, but these scalabilities are implemented and performed independently, thereby neglecting the gains that can be made from making joint SNR-temporal decisions to maximize quality. In this paper, a novel Fine Grained SNR-Temporal scalability framework called FGS+ that provides a new level of performance by considering SNR and temporal scalability jointly is presented. This new solution uses the results of our subjective tests, which indicate the levels to which SNR-quality needs to be enhanced before motion-smoothness should be improved. The study also reveals that these SNR-temporal tradeoff points vary among videos, and depend on the characteristics of the video. Based on these observations, our solution uses reference frames, enhanced relative to FGS, for prediction, improving visual quality over MPEG-4 FGS by up to 1.5 dB.

1. INTRODUCTION

Real-time streaming of audiovisual content over wireless networks (GPRS, UMTS, WLANs etc.) is emerging as an important technology area in multimedia communications. Due to the wide variation of available bandwidth over short intervals in wireless sessions, there is a need for scalable video coding methods that allow streaming to flexibly adapt to changing network conditions in real-time. One such technique is MPEG-4 Fine-Granular Scalability (FGS) [1][2], which can adapt in real-time to bandwidth variations while using the same pre-encoded stream.

The key advantages of the MPEG-4 FGS framework -resilience and flexibility- come at the expense of lower video quality. In [1], the FGS performance is compared to a set of non-scalable streams coded at discrete bit-rates covering the same bandwidth range. The results obtained indicated a decrease in coding efficiency of up to 2-3 dB for FGS[1].

In this paper, a novel scalable video-coding framework called FGS+ that improves the FGS coding efficiency is introduced. The basis of this new scheme lies in the realization that in the FGS framework, the SNR and temporal scalabilities are implemented and performed independently, neglecting that SNR-temporal tradeoffs should be made jointly for improved visual quality. First, we present the results of a subjective study that allows us to determine the optimal division of bits between SNR and temporal layers at different bit-rates. Based on this analysis, we conclude that to optimize overall visual quality, certain tradeoffs between the bandwidths allocated to spatial and temporal layers need to be established. For instance, at low transmission bit-rates, the SNR-quality requires relatively larger improvement before motion-smoothness is improved (i.e. at low bit-rates, more bandwidth should be allocated to SNR-quality instead of temporal-quality). We then present an enhanced FGS scalability solution called FGS+ that uses these relationships to produce up to 1.5 dB gain compared with the MPEG-4 FGS framework. We tabulate the gains for various video sequences at different bit-rates.

The paper is organized as follows. Section 2 briefly describes MPEG-4 FGS scalability. Section 3 presents the results of our subjective study on FGS SNR-temporal-quality. Sections 4 and 5 present two variants of the improved FGS+ framework that include performance results and comparisons with MPEG-4 FGS. Section 6 outlines how FGS+ parameters for unseen videos can be chosen, and Section 7 draws conclusions.

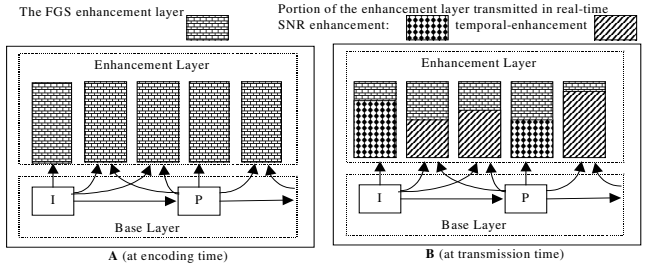

Fig. 1. The FGS SNR-temporal scalability structure (A) and examples of its usage in supporting joint SNR-temporal scalability in a fine-granular way (B).

2. MPEG-4 FGS SNR-TEMPORAL SCALABILITY

This section briefly presents the MPEG-4 FGS framework (the user is referred to [2] for details) whose structure is portrayed in Fig. 1A. Under this framework, video content can be compressed to cover any bandwidth range [R_{min}, R_{max}] by the use of two streams: a base-layer (BL) stream that is always transmitted and an enhancement-layer (EL), which is transmitted only as bandwidth allows. The base-layer bit-rate R_{BL} is chosen for coding the base-layer (BL) such that the available bandwidth is almost certainly lower than R_{min} at all times (i.e. $R_{BL} \leq R_{min}$). The FGS enhancement-layer (EL), which is progressively coded using a fine-grained approach based on bit-plane DCT coding, improves upon the base-layer video, and supports *both* SNR and temporal scalability through a *single* pre-encoded stream. MPEG-4 FGS scalability therefore provides flexibility in supporting (Fig 1B):

- Temporal scalability by increasing only the frame-rate.
- SNR scalability while maintaining the same frame-rate.
- Both SNR and temporal scalabilities.

[1] Although this type of comparison may seem to be unfair to FGS - the non-scalable streams are optimized for particular bit-rates whereas FGS covers the same range of bandwidth with a single enhancement-layer - this comparison provides insight into the theoretical limits of the FGS rate-distortion performance if the base-layer and enhancement-layer are independently coded and optimized.

However, no automatic mechanism for performing the optimal SNR-temporal tradeoff is standardized in MPEG-4.

3. JOINT SNR-TEMPORAL FGS+ ENCODING

As mentioned in the previous section, the MPEG-4 FGS server has two degrees of freedom: the temporal-quality *or* the SNR-quality can be enhanced at each transmission bit-rate. However, SNR and temporal-quality are just two components of the overall video quality, and hence, only certain combinations of SNR and temporal quality will lead to maximal overall quality (as depicted in Fig. 2A).

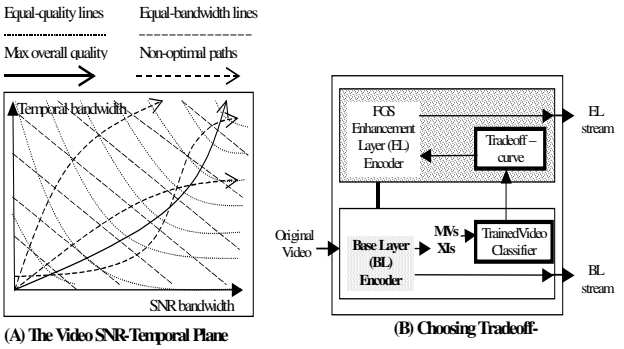

Fig. 2: (A) SNR-Temporal tradeoffs at various bit-rates. (B) SNR-temporal tradeoff for video-classes based on BL data

Since no standardized objective measure for determining the overall (SNR-temporal) quality of a coded sequence exists, we conducted a subjective quality assessment for evaluating the performance of various FGS SNR-temporal tradeoffs at four transmission bit-rates R_t = (200,300,500,1000kbps)[2] for four 10-second CIF sequences from the MPEG-4 test suite. The four sequences were selected to cover a wide range of video characteristics. Their relative motion-vector magnitudes (MV) and I-frame TM-5 image-complexity values (Xi) are shown in Table 1.

	Foreman	*Stefan*	*Coastguard*	*Mobile*
Motion Vector size in bits (MV)	7,200	13,100	10,100	3,800
TM5 I-frame Complexity (Xi)	267,163	519,229	354,381	990,749

Table 1: Relative values of motion-vectors and TM-5 complexity in the four videos used in testing

At each bit-rate R_t, four videos with different frame-rates (5, 7.5, 10, 15 fps) and corresponding[3] SNR-quality were produced. The videos were synchronized and shown simultaneously to the twelve viewers who, at each bit-rate, selected the encoding perceived to have the best overall quality. The results, representing the average frame-rate chosen at each bandwidth are shown in Fig. 3. The following conclusions can be drawn from the figure:

1. Until the SNR-quality improves to an acceptable level, the users prefer that the additional bandwidth be used to enhance the SNR-quality. Once SNR-quality has improved adequately, the preference is for improved temporal-quality (i.e. motion-smoothness).

The curves vary substantially for different videos, and are correlated to the texture-complexity and motion-vector magnitude of the videos. For instance, spatially complex sequences such as Mobile require more bits to produce an image of satisfactory SNR-quality, while videos with large motion-magnitude, such as Coastguard, require more temporal-quality at a lower bandwidth.

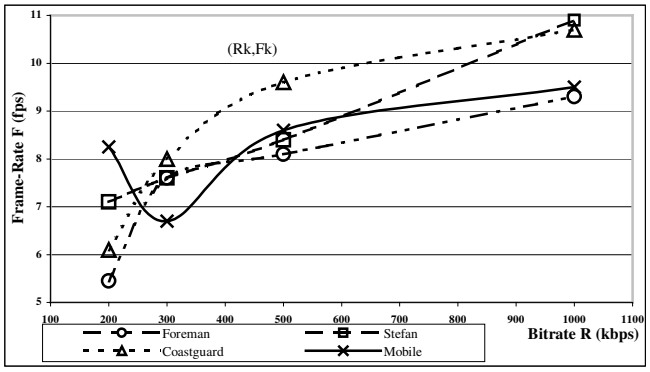

Fig. 3: SNR-Temporal preference curves for different videos.

Since the study indicates a clear preference for a specific allocation of bandwidth between SNR and temporal-quality at each bit-rate, the server can make optimal rather than ad-hoc choices about bandwidth allocation. Additionally, since this allocation path is pre-determined for a particular video, the encoder/transmitter and decoder/receiver can follow the same path in the SNR-temporal plane (Fig. 2A) and will know the exact number of bits allocated to each frame at any bandwidth, as depicted in Fig.2B.

4. FGS+ FOR B-FRAMES

This section describes our enhanced FGS+ framework, which uses the results of the study in Section 3 to provide "extended" reference frames for temporal B-frames in the enhancement-layer (EL).

Recall that in MPEG-4 FGS, the EL frames are solely predicted from the base-layer. This ensures that complete reference-frames are always received at the decoder, independent of bandwidth. However, these small references-frames reduce coding efficiency as temporal correlation between frames is only minimally exploited. In our schemes, the path the transmitter will take in the SNR-temporal-quality plane is known a-priori (Fig. 2A).

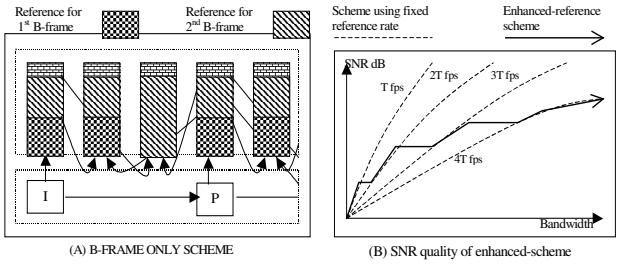

Fig. 4: (A) The proposed scheme using optimized reference and (B) SNR-quality as bandwidth increases

This path pre-determines the level to which the quality of each frame is improved before a new temporal-enhancement frame is introduced. Therefore, the newly introduced temporal frame can use extended-frames as reference, thereby producing smaller residues and improving compression efficiency. MPEG-4 FGS and the FGS+ scheme are illustrated in Fig. 1B and Fig. 4A. Note that while the temporal B-frames introduced first in FGS+ use reference frames extended by a small fraction of the bits present in the EL,

[2] The base-layer bit-rate R_{BL} was 100 kbps.
[3] Bits were divided among frames to maintain uniform SNR-quality.

the second temporal B-frame, introduced at a higher bandwidth R_t uses reference frames extended by a larger number of bits.

Subsequently, this section answers the two questions that need to be addressed in implementing the GFS+ framework:

- At which bandwidths R_k should temporal-frames be introduced (i.e. the frame-rate be increased)?
- When a temporal-frame is introduced, what should be the optimal size of the extended-reference?

4.1 Introducing temporal-frames

We illustrate the procedure for introducing temporal-frames through an example. Assume that the base-layer frame-rate is F_{BL} and that for the video being transmitted, the SNR-temporal subjective study has indicated that users prefer videos with frame-rates of F_i until each enhancement-layer frame improves to a quality corresponding to B_i bits (R_i–R_{BL}=B_i*F_i), followed by frame-rates of F_{i+1} until each frame improves to B_{i+1} bits etc., where F_0=F_{BL} and R_0=R_{BL}[4]. Let us examine what happens when the current bandwidth changes to R_N when the system is operating at a bandwidth R_C corresponding to a frame rate of F_C and a quality of B_C bits/frame. In the MPEG-4 FGS scheme (Fig. 1B), at the bandwidth R_N, the transmitter can choose the best SNR-quality of (R_N-R_{BL}/F_{BL} bits/frame, the best temporal-quality (determined by the number of FGS-temporal frames compressed at encoding time) or make any other SNR-temporal tradeoff in between. However, in the enhanced scheme (Fig. 4A), the transmitter takes a pre-determined path with respect to dividing additional bandwidth between SNR and temporal-quality. When the bandwidth changes to R_N, the transmitter first identifies k, such that F_k*B_k < R_N-R_{BL}< F_{k+1}*B_{k+1} and allocates (R_N-R_{BL})/F_k bits to each enhancement-layer frame. If F_k > F_C, the transmitter will introduce new frames to increase the frame rate from F_C to F_k.[5]. For instance, if R_N for the Coastguard video changes to 300kbps, based on the results depicted in Fig. 3, F_N should be changed from 5 fps to 8 fps, yielding B_N = (R_N –R_{BL})/F_N = 25 kb given R_{BL} = 100kbps.

Note that in the new scheme, at frame-rate F_k, the encoder and decoder are aware[6] that all frames have qualities corresponding to at least B_k, such that the newly introduced temporal frames do not need to be predicted solely from the base-layer, but can be predicted from the extended reference frames of size B_{BL}+B_k. Adopting higher quality extended reference frames for FGS+ leads to an improved compression efficiency as will be shown in Section 4.3.

4.2 Determining the size of the reference

The user-study indicates the bandwidths at which temporal-quality should be improved by introducing new B-frames. However we do need to determine the optimal amount of extended-reference B_{BL}+B_i to use at each bandwidth R_i. One approach to determine these optimal B_{BL}+B_i is empirical: plot the compression performance as a function of the extended-references B_i at each bandwidth R_i, and choose the extension B_{OPT} that provides the best performance.

The compression performance (PSNR) of the encoded video for the Coastguard sequence is presented in Table 2. It tabulates the PSNR improvement of each frame as a consequence of using varying amount of extended-references. At each bit-rate R_i, the performance improves as we extend the reference-frames by adding more EL bit-planes (BP). However, if we use too many bit-planes as reference, the performance starts degrading because at that point, not all the additional bits used in enhancing the reference are transmitted; thus the decoder makes predictions based on incompletely constructed references. As R_i increases, the number of bit-planes than can be used to improve the reference without a loss in performance increases. This is clearly seen in Table 2, where at 300 kbps performance peaks at 2 bit-planes while at 1500 kbps performance falls off only after 4 bit-planes. So, for this video sequence, at a R_i of 300 kbps, the optimal performance improvement is obtained by using 2 bit-planes of enhancement, while at a R_i of 1000 kbps, the optimal choice is 3 bit-planes.

Kbps	*BP=0*	*BP=1*	*BP=2*	*BP=3*	*BP=4*	*BP=5*	*BP=6*
300	28.5	28.7	**29.3**	28.1	27.7	27.4	27.3
500	30.3	30.5	**31.1**	30.9	30.0	29.7	29.4
1000	34.2	34.3	34.8	**35.3**	34.8	33.8	33.4
1500	37.0	37.3	38.2	**38.8**	**38.8**	37.3	36.7

Table 2: The SNR in dB for the Coastguard video as a function of the bit-rate and the number of bit-planes used as reference.

4.3 Performance

The improvement in performance of the B-frame only FGS+ scheme over traditional MPEG-4 FGS is presented in Table 3. The videos had a base-layer bandwidth R_{BL} or approximately 100 kbps, and contained one B-frame per P-frame[7].

Kbps	*Foreman*	*Stefan*	*Coastguard*	*Mobile*
300	0.48 (2)	0.31 (2)	0.63 (2)	0.19 (2)
500	0.85 (3)	0.65 (2)	0.62 (3)	0.74 (2)
1000	0.98 (3)	0.90 (2)	0.67 (3)	0.94 (3)
1500	1.09 (4)	1.22 (3)	1.15 (4)	1.28 (3)

Table 3: The maximum improvement in performance at different bit-rates for the B-frame case.

The number of FGS bit-planes used to enhance the base-layer was chosen by the empirical method outlined above and is given in parenthesis. The performance improvement ranges from 0.19 dB to 1.28 dB. Note that at low bit-rates spatially simple videos such as Coastguard benefit the most, while at higher bit-rates the spatially most complex sequence such as Mobile derive the largest improvement.

5. FGS+ FOR ALL-FRAMES

In the scheme depicted in Fig. 4B, only the temporal B-frames in the enhancement-layer use extended-reference for prediction. However, performance can be further improved by additionally using extended-references for the base-layer P-frames (Fig. 5A). Unlike the B-frame only case, depicted in Fig. 4A, the subjective study does not provide a guideline for the amount of enhancement to be used for P-frames prediction, as base-layer P-frames are present at all bandwidths. Therefore, we need a different approach to determine the amount of extended-reference to use for the P-

[4] Note that for the base-layer, a different number of bits BBL is used for each frame to ensure constant quality among frames.

[5] If frame-rates are changed abruptly, there will be a vertical drop in SNR-quality rather than the horizontal change depicted in Fig 4(B).

[6] If the SNR-temporal tradeoffs made by the encoder at different bit-rates are not known at the decoder, the encoded stream needs to indicate the size of the reference frames used for FGS+ prediction.

[7] It should be noted that the P-frame performance does not change in this scheme, so the average B-frame improvement in performance is twice what is shown in Table 3.

frames. If we choose the amount of enhancement to be B_T bits/frame corresponding to a bandwidth R_T, there are always B_T bits available for each I and P frame at bandwidths higher than R_T. The performance at bandwidths higher than R_T is enhanced due to the larger references used in predicting the P-frames, while at bandwidths lower than R_T, performance degrades, as all B_T bits are not transmitted resulting in incomplete references. This problem is compounded by prediction-drift, where the error in a P-frame is transmitted to all subsequent P-frames within the GOP and stops only at the next I-frame [3]. Hence, the advantages and drawbacks of choosing an enhancement of B bits/frame must be balanced, as illustrated in Fig. 5B.

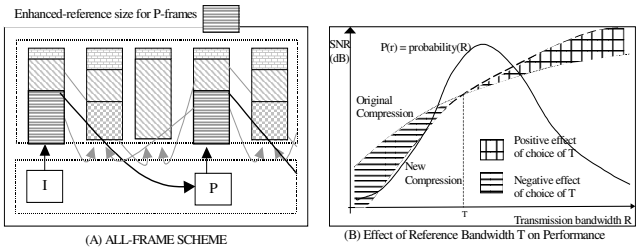

Figure 5: (A) All-frame scheme. (B) Effect on performance of choice of reference bandwidth T

One solution to finding the optimum R_T is to calculate the probable magnitude of improvement given the choice of transmission bandwidth R_T, and probability of transmission at bandwidth r to be p(r). Namely we maximize the expectation of quality improvement by choosing the optimal reference frame.

$$\max_T \left[\left\{ F(T) = \sum_r p(r)\left[PSNR_{new}(r) - PSNR_{old}(r)\right] \right\} \right]$$

After an optimum R_T and the corresponding enhanced-reference level B_T for the P-frames is chosen, all P-frames are predicted from BL frames enhanced to B_T bits by the addition of EL bit-planes as shown in Fig. 5A.

5.1 Performance

The additional improvement in performance gained by using 3 bit-planes of enhanced references for the P-frames is shown in Table 4. For the complex sequence Mobile, the performance degrades marginally at low bit-rates indicating that all 3 additional bit-planes used in creating the enhanced-reference were not transmitted at 300 kbps. Note also that the complex sequences Stefan and Mobile benefit the most at high bandwidths. The results show that at 3 bit-planes the P-frame scheme adds –0.03 dB to 0.59 dB over the B-frame only scheme. Thus using both the B-frame and all-frame schemes result in overall improvements of 0.16 dB to 1.61 dB.

Kbps	*Foreman*	*Stefan*	*Coastguard*	*Mobile*
300	0.18	0.08	0.03	-0.03
500	0.28	0.15	0.16	0.04
1000	0.34	0.36	0.28	0.20
1500	0.31	0.59	0.31	0.30

Table 4: The improvement in performance in SNR dB of the All-Frame FGS+ over B-frame only FGS+ when 3 EL bit-planes are used to enhance the reference

6. CHOOSING PARAMETERS

The bandwidths R_i at which new temporal frames are introduced, the number of bit-planes B_i to use in enhancing the EL temporal references, and the number of bit-planes B_T to use in enhancing the BL P-frame references vary, depending on the characteristics of a video. The characteristics that influence R_i significantly are TM-5 complexity (Xi) and motion-vector magnitude (MV), while Xi largely influences the choice of B_i and B_T. Good choices of these values for unseen videos can be made by classifying videos based on values such as Xi and MV, then using the characteristic value of the class. Since Xi and MV are calculated in the BL by the encoder and decoder, no additional cost is incurred in this process (see Fig 2B).

More detailed algorithms for choosing the appropriate parameters based on observing many video characteristics will be presented in future work. We will adopt content-based classification methods that predict adequate parameter classes based on the video features (such as MV & Xi). Similar approaches have been used in predicting video traffic classes on rate-distortion curves in [4].

7. CONCLUSIONS

This paper has presented a fine-grained scalable encoding scheme referred to as FGS+ that produces gains of up to 1.5 dB over MPEG-4 FGS. In introducing fine-grained scalability MPEG-4 FGS paid a 2-3 dB cost. We show how more than half of that loss can be regained through an innovative view of video quality: we consider SNR-temporal-quality jointly rather than independently and determine their optimal combination at each bandwidth through a subjective study. This new solution uses the results of our subjective tests, which indicate the levels to which SNR-quality needs to be enhanced before motion-smoothness should be improved for optimal visual quality. Based on this observation, our solution uses improved base-layer frames to predict temporal-scalability frames instead of the base-layer frames in the original FGS scheme. Furthermore, the study revealed that different SNR-temporal tradeoffs videos should be made based on sequence characteristics at various transmission bit-rates. Since our tradeoff experiment is resolution dependent, we plan to conduct further studies that include resolution scalability along with temporal and SNR scalabilities.

References

[1] H. Radha, M. van der Schaar, Y. Chen, "The MPEG-4 Fine-Grained Scalable Video Coding method for Multimedia Streaming over IP", IEEE Trans. on Multimedia, March 2001.

[2] M. van der Schaar, H. Radha, "A hybrid temporal-SNR Fine-Granular Scalability for Internet Video", IEEE Trans. on CSVT, March 2001

[3] M. van der Schaar, H. Radha, "Motion-Compensation based Fine-Granular Scalability (MC-FGS)", MPEG-4 Contribution M6475, October 2000.

[4] S.-F. Chang and P. Bocheck, "Principles and Applications of Content-Aware Video Communication," IEEE ISCAS, Geneva, Switzerland, May 2000.