FGS+: A FINE-GRANULAR SPATIO-TEMPORAL-SNR SCALABLE VIDEO CODER

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

Columbia University, Philips Research USA

ABSTRACT

To enable universal multimedia access, a framework that allows a video stream to be delivered to displays with different viewing resolutions under varying network conditions in real-time with minimal processing is necessary. Existing scalable coders such as MPEG-4 are limited in that they use multiple loops to provide resolutions and implement Spatial, SNR and Temporal scalabilities independently neglecting gains that can be made by considering these scalabilities jointly.

We present a new Spatio-Temporal-SNR scalable coder called FGS+ that provides Spatial scalability in addition to Fine-Grained SNR and Temporal scalabilities and a new level of performance by considering Spatial, SNR and Temporal scalability jointly. We present experimental results that show the performance of the FGS+ coder to be comparable to other schemes while providing enhanced flexibility.

1. INTRODUCTION

Real-time streaming of audiovisual content over wireless networks is emerging as an important area in multimedia communications. Due to the wide variation in the kinds of devices on which videos may be viewed and of network bandwidth, there is a need for scalability in video coding methods that allows streams to be flexibly adapted for different display resolutions and network condition in real-time.

Standardized codecs such as MPEG-2 and MPEG-4 implement coarse Spatial scalability and a variant called frequency scalability as a partial solution to satisfying the resolution requirements of different clients. MPEG-4 also implements Fine-Grained SNR Scalability (FGS) which was proposed as a solution to the problem of transmitting video over networks with fast-varying bandwidth conditions. However MPEG-4 suffers from various limitations:

- It implements Spatial scalability in a multiple-loop architecture, reducing efficiency and increasing coding complexity.
- It does not allow Spatial scalability and SNR scalability to be used simultaneously.
- It considers SNR, Spatial and Temporal scalabilities independently rather than jointly.

Progressive Fine Grain Scalability (PFGS) [1] is one attempt to more efficiently use the correlation present in videos by providing the decoder multiple references. A Spatial scalability scheme called FGSS [2] has been proposed on top of PFGS, but provides only two resolutions.

The second author is with Philips Research, USA

In this paper, FGS+, out novel scalable video framework that improves FGS coding efficiency in addition to providing Spatial scalability is introduced.The basis of the new scheme lies in the realization that in the FGS framework, the SNR, Spatial and Temporal scalabilities are implemented and performed independently, neglecting that Spatio-Temporal-SNR tradeoffs should be made jointly for improved visual quality. We present the results of two subjective studies that determine the optimal division of bits between the SNR, Spatial 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 the Spatial, SNR and Temporal layers need to be established.

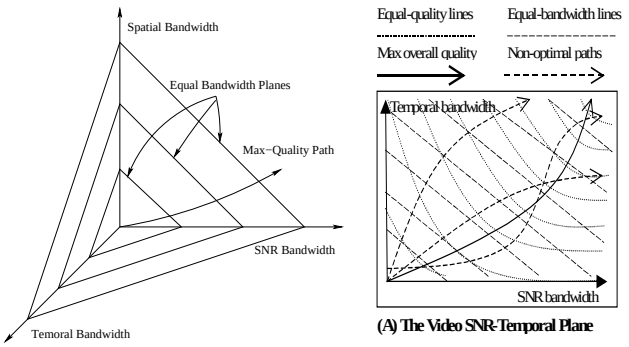
The rest of the paper is organized as follows. Section 2 presents our approach to maximizing overall video quality and Section 3 describes FGS+. Sections 4 and 5 outline two distinct problems of bandwidth allocation that come up when using such a flexible coder, and provide analytical and subjective solutions to them. Section 6 provides experimental results of our implementation of resolution scalability and Section 7 draws conclusions.

2. JOINT SPATIO-TEMPORAL-SNR VIDEO QUALITY

The notion of overall video quality is an important consideration when coders provide multiple dimensions of flexibility and merits some discussion. When bandwidth conditions change, a server with a Spatio-Temporal-SNR scalably coded stream has the flexibility to respond by changing the streaming bandwidth in three different dimensions: SNR, Resolution, or Temporal bandwidths as illustrated in Fig. 1(A) where the parallel planes indicates points of equal total bandwidth. At each bandwidth there is some point that produces the best overall quality. This is illustrated in Fig. 1(B) which is a slice through the SNR and Temporal dimensions. Equal quality lines are concave, i.e., allocating all the bandwidth to SNR quality or Temporal smoothness will produce poorer perceived quality than some balanced combination. The points on each equal-bandwidth plane that correspond to the best quality (equivalently the points that are tangent to a quality curve) trace the path the server should take in the Spatio-Temporal-SNR dimensional bandwidth space as available bandwidth varies.

Unfortunately no standardized objective methods for determining overall video quality exist today. Peak Signal-to-Noise ratio (PSNR) can be used to measure SNR quality, frame-rate or frame-difference statistics can be used to measure Temporal smoothness and pixel-count can be used to measure Spatial quality independently, but methods to combine these independent valuations do not exist. A measure that takes into account the sensitivity of the human visual system to different spatial and temporal frequencies has been proposed in [3]. A second measure that computes a set of impairments in the video, then weights these impairments by correlating them to the choices of human viewers has also been

proposed [4], but both these measures need extensive use and test before they can be reliably used. Therefore overall quality, for the moment, needs to be evaluated subjectively. We present the results of such tests in Sections 3 and 5.



(A) Video Quality Dimensions (B)The SNR Temporal Slice

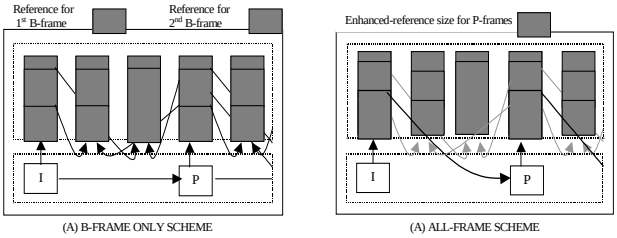
Fig. 1. The Joint Spatio-Temporal-SNR Video Quality Space

3. FGS+

This section briefly presents the MPEG-4 FGS framework (the user is referred to [2] for details) and our extension to it called FGS+. FGS provides Fine-Grained SNR scalability by dividing the video into two streams, a low bitrate base-layer that can always be transmitted and an enhancement layer which can optionally be transmitted to enhance the SNR quality of the base layer. The enhancement-layer improves upon the base-layer video, and is progressively coded using a fine-grained approach based on bitplane DCT coding.

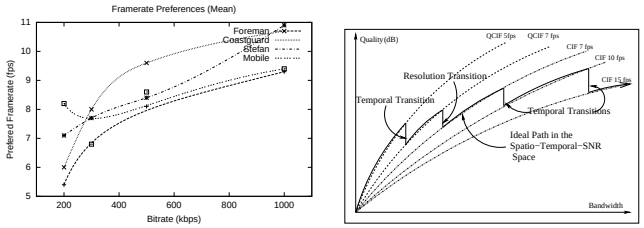
3.1. Enhancing efficiency

MPEG-4 FGS, however, suffers from a 1-2 dB loss in coding efficiency since it does not take advantage of the temporal correlation present in the enhancement layer. FGS+ partially solves this problem by using the knowledge that only one path in the Spatio-Temporal bandwidth plane produces the best overall quality (Fig. 1). This path implies that a certain amount of the enhancement layer will definitely be present before new temporal frames are introduced to improve temporal smoothness. Therefore the newly introduced frame can use a reference extended by the amount of the enhancement layer known to be present as pictured in Fig. 2(A). A variant that uses enhanced references for all frames, and produces further improvements in performance at high bitrates is pictured in Fig. 2(B). The reader is referred to [5] for more details. Since there exist no quantitative techniques that can measure joint spatio-temporal video quality, we conducted subjective tests to determine the ideal path in the Spatio-Temporal bandwidth plane. The results are charted in Fig. 3(A) for four videos that were chosen to represent four broad classes of videos based on complexity (Xi) and motion-vector magnitude (MV). The results show a need for more temporal smoothness as SNR quality improves and also that this need for improved temporal smoothness varies for different classes of video.



(A) FGS+ for B-Frames only (B) FGS+ for All-Frames

Fig. 2. Enhancing FGS efficiency

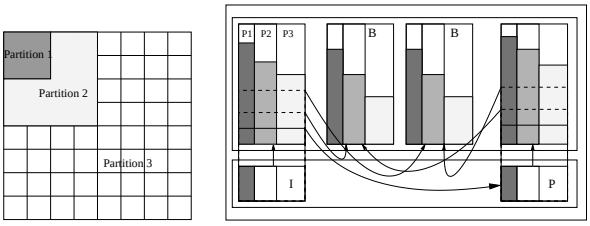


(A)SNR-Temporal Tradeoffs (B)Spatio-Temporal Tradeoffs

Fig. 3. Spatio-Temporal-SNR Tradeoffs

3.2. Spatial Scalability

In addition to improved SNR Fine-Grained scalability, FGS+ provides Spatial scalability. Our implementation allows multiple resolutions, and achieves efficiency in coding by taking advantage of the enhanced references mentioned in Section 3.1. It is based on partitioning the DCT coefficients. While many partitioning schemes are possible the octave partitioning scheme pictured in Fig. 4(A) corresponds to three well known resolutions: the top-left 2x2 coefficients of Partition 1 correspond to a QQCIF resolution, the remaining top-left 4x4 coefficients of Partition 2 correspond to the QCIF resolution, and the rest of the coefficients that make up Partition 3 correspond to the full CIF resolution. Other partitioning schemes can provide finer-grained resolution scalability if needed. Unless mentioned otherwise, we use this octave partitioning scheme for the rest of the paper.



(A)Octave Partitioning (B)FGS+

Fig. 4. FGS+ Spatial Scalability

In our scheme the base-layer contains only the[1] coefficients that correspond to the lowest resolution the stream supports (which

[1] sufficiently quantized to meet any base-layer bandwidth restrictions

in our example are the top-left 2x2 coefficients from each block that constitute Partition 1 and the QQCIF resolution). The residue of the first partition and the coefficients of the other partitions are transmitted as enhancement layers, one for each resolution. Each of these enhancement layers is bit-plane coded in the traditional manner of FGS enhancement layers. [2] Our example scheme would produce a base-layer and three enhancement layers for Partition 1, 2 and 3 corresponding to the QQCIF, QCIF and CIF resolutions.

The server, based on the resolution and bandwidth requirement of each client, sends the base-layer and only the required parts of each enhancement stream. If for example a client requires to display the stream at a QCIF resolution, the server would send the base-layer and the first two enhancement layers corresponding to the QQCIF resolution and the QCIF resolution. The third stream would not be transmitted at all. The decoder would then reconstruct an image that would only contain the coefficients of Partition 1 and Partition 2, which would be the top-left 4x4 coefficients which corresponds to a video at the QCIF resolution. The structure of FGS+ Spatial scalability is pictured in Fig. 4(B).

3.2.1. Spatial Scalability Implementation Issues

Implementing Spatial Scalability entails some tradeoffs. Allowing only the frequencies corresponding to the lowest resolution in the base-layer ensures that no redundant frequency information is present if the server decides to use the lowest resolution. However, a base-layer with such a small number of coefficients reduces the performance at higher bitrates since temporal correlation is not effectively exploited. Allowing frequencies other than that of the lowest resolution into the base-layer will improve performance at higher bitrates but will cause drift in the lower frequencies. Pure DCT domain implementations such as the DCT Pyramid, a Layered Coding technique that follows a filter-bank approach have been suggested [6]. However motion-estimation (ME) in the DCT domain is difficult [7] and and been found to be much less efficient than ME in the spatial domain.

Therefore the Partition scalability scheme is a practical alternative. We briefly outline a couple of issues that need to be addressed when scalability is provided in such a manner. First, FGS coding seperates the coefficients of each block into individual bitplanes, then run-length and entropy codes each individual block in each bitplane. It also uses a special End-Of-Block symbol (EOB) to avoid the last run of zeros of each block of each bitplane. Since FGS+ divides each 8x8 block into partitions and individually FGS codes each partition into separate streams, a special End-Of-Partition symbol (EOP) is used to signal the end of each partition similar to the EOB symbol of FGS. Since each block is divided into N partitions, there will be N times as many EOP symbols as there are EOB symbols in FGS thus reducing the efficiency of the run length coding. A solution to this and other inefficiencies are provided in Section 6. Second, it has to be realized that the Spatial scalability provided is a *texture* scalability, and that the amount of motion-vector (MV) information does remains constant. This does not present a problem until approximately 32 kbps, as long as 16x16 MVs at 1/2 pel accuracy are used. For rates below that, a hierarchical scheme that sends coarse MVs in the base-layer and refinements in the different partitions is needed.

[2] each partition is bit-plane coded, then run-length coded, and finally entropy coded

3.3. Navigating the Spatio-Temporal-SNR Space

With FGS+'s Spatial scalability, the server has three degrees of freedom: it can send the bits required for a required resolution; it can change the overall bitrate and SNR quality in response to fluctuations in network bandwidth by simple truncation (since each partition is encoded as an independent Fine-Grained Scalable stream); it can change temporal resolution by choosing to send or not send frames. In summary, the server has complete control in the Spatial, SNR and temporal dimension of the stream sent to each client, and can exercise this control, by simple bit truncation.

The question that naturally arises from the realization that the server has these three degrees of freedom is the appropriate division of bandwidth (or quality) among these three dimensions. As mentioned in Section 2, there is one ideal path through the Spatio-Temporal-SNR space and is shown in Fig. 3(B) mapped to PSNR curves at different frame-rates and resolutions. [3] It can be seen that two kinds of transitions are made: temporal-transitions between frame-rates and spatial-transitions between resolutions. The SNR-Temporal tradeoff study of Section 3.1 showed when these temporal transitions are to be made. The Resolution-SNR tradeoff study that is presented in Section 5 will show when the resolution transitions should be made.

The coder and the decoder need to know and agree on when SNR-Temporal transitions need to be made, so that they can use appropriate amounts of enhanced reference. The server needs to know when to make resolution transitions (since it controls the ultimate display resolution) and how to apportion bits among the partitions as will be explained in Section 4.

The results of Figs. 3(A), and 6 also show that the ideal transition points and bit allocations vary for different classes of video. To take advantage of this variation, multiple transitions and bit-allocations mapped to different video-classes should be stored at the coder/decoder and the server, and the appropriate data chosen based on the class of the video. Complexity (Xi) and motion-vector magnitude (MV) are good measures to use in classifying videos as Xi correlates well with the need for SNR and Spatial quality, while MV correlates well to the need for temporal smoothness. If these measures are chosen, the encoder and decoder can periodically measure the Xi and MV of a video, classify that interval of the video, and choose the appropriate transitions. In addition, the coder should embed these Xi and MV values in the stream as out-of-band data, so that the server can also periodically extract the information and make resolution transition and bit-allocation decisions based on the class of that interval of the video.

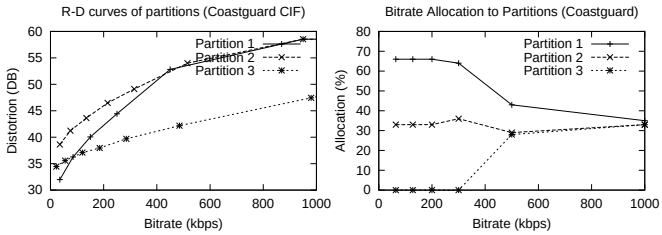
4. RATE-DISTORTION SNR-QUALITY OPTIMIZATION

FGS+ gives the server the choice of determining the amount of each partition to send to the decoder thereby determining the resolution of the output video. When bandwidth is limited, the server needs to truncate the different partitions that constitute a resolution to meet the bandwidth constraint. This question of determining the amount of each partition to retain can be formulated and solved as a constrained rate-distortion problem: given a bandwidth constraint, what allocation among the partitions will result in least distortion? We present the intuition behind the solution. For details on Lagrangian-based rate-distortion optimization the reader is referred to [8].

[3] Note that PSNR measures only SNR quality, not overall video quality. Therefore the ideal path in Fig. 3B shows occasional drops in PSNR.

The solution is to compute the Rate-Distortion curves for each partition. Then, at each bitrate, the partition corresponding to the steepest R-D slope is allocated additional bandwidth since that allocation produces the largest overall reduction in distortion.

This procedure is illustrated in Fig. 5, where the first graph shows the R-D curves for partition 1,2 and 3 from our example 3-resolution octave partitioning scheme for the Coastguard video. At each bandwidth the partition with the largest slope is allocated additional bandwidth. The allocations that resulted from such a search for the Coastguard video at the CIF resolution are shown in Fig. 5(B). Allocations for different videos classes can be precalculated and used as explained in Section 3.3.



(A)Partition R-D curves (B)Partition Bit Allocation

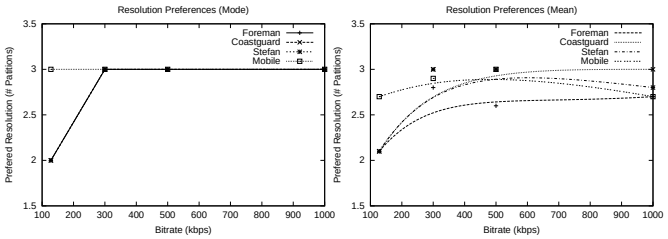
Fig. 5. Bitrate allocation using R-D curves

5. CHOOSING THE OPTIMAL RESOLUTION

Since the FGS+ coder allows the server to change resolutions as bandwidth becomes increasingly available, the question of when to switch to a higher resolution remains (Fig. 3(B)). Rate-Distortion analysis does not provide a solution here, as it cannot account for coding artifacts such as blockiness. Our solution to the problem was to conduct a subjective study.

Four videos representing four different classes were chosen. The classification was based on complexity (Xi) and motion-vector magnitude (MV) as Xi correlates well with the need for SNR and Spatial quality, while MV correlates well to the need for temporal smoothness. For each video, at four different bitrates (128, 300, 500, 1000 kbps) three videos were produced: CIF, QCIF and QQCIF [4]. These three videos were simultaneously displayed at the same size to users who were asked to choose the encoding that they perceived to have the best overall video quality. The mean and the modes of the preferences of 9 users for MPEG test sequences Foreman, Coastguard, Stefan and Mobile are charted in Fig. 6 where the preferences (Y axis) is charted as the number of partitions the users preferred. The results show a clear preference for the CIF video at bitrates above 300 kbps. At lower bitrates users prefer the QCIF resolution except for the Mobile sequence which has the largest complexity (Xi) among the videos, indicating that more complex sequences require quicker transitions to a higher resolution. This variation among video classes can be effectively used by the server as explained in Section 3.3.

[4]Note that, although bitrates lower than 128 kbps are possible, only one resolution is available at those rates, allowing the user no choices



(A)Mode (B)Mean

Fig. 6. Subjective Resolution Preferences

6. RESULTS

The performance of the Spatio-Temporal-SNR coder on the MPEG-4 test sequences Foreman and Coastguard at the CIF resolution with different base-layer rates are shown in Fig. 6(A). The effects of the size of the base layer are clearly seen. The performances of a MPEG-4 Codec that does not have Spatial scalability and of the FGSS coder as reported in [9] are shown in Fig. 6(B). It can be seen that the performance of FGS+ is comparable to that of FGSS.

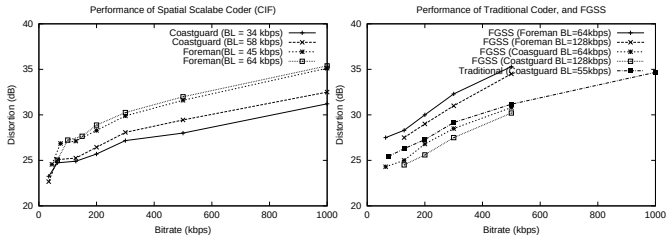
The cost of providing Spatial scalability are charted in Figure 6(C) where FGS+ with scalability is compared to FGS+ without. FGS+ without Spatial scalability is shown with 1,2 and 3 partitions in the base layer. The costs can be seen to extend to approximately 1 dB at high rates and is due to the following:

- Since only the lowest resolution coefficients are present in the base-layer, temporal-correlation is less effectively used.
- There are N times as many EOP symbols, as there are EOB symbols in FGS (approximately 2000 * N extra symbols per frame).
- The probabilities of run-lengths used in FGS entropy coding are optimized assuming 8x8 blocks. They are no longer optimal when partitions are used.
- FGS+ uses a constant amount of enhanced reference from all partitions. However as shown in Section 4 the optimal allocation of bits across partitions is uneven.

The costs due to the last three item can be eliminated: inefficiencies due to the extra EOB symbols can be overcome by performing run-length coding over all blocks in a frame, rather than a block-by-block basis. The cost due to mismatched entropy coding tables can be overcome by simply regenerating the tables for a Scalable coder. The last inefficiency can be avoided by using different amount of enhancement from different partitions according to the R-D Optimized allocations of Section 4. We will provide the results of such additional performance improvements in forthcoming papers.

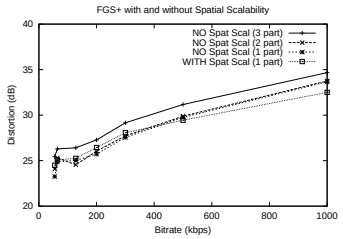
7. CONCLUSION

We have presented a video coding framework called FGS+ that provides spatial scalability in addition to Fine-Grained SNR and Temporal scalabilities. We provide experimental results that show that this new FGS+ coder provides performance comparable to other coders while providing enhanced flexibility which will allow servers to better respond to varying bandwidth conditions and resolution requirement. We then show that with such a flexible



(A)FGS+

(B)Non-Scalable Coder, FGSS



(C)FGS+ with/wo Spat. Scal.

Fig. 7. Performance of different coders

coder, one should make joint rather than independent decisions about Spatio-Temporal-SNR tradeoff to achieve optimum video quality. We provide an R-D optimization technique and the results of two subjective video-quality studies that allows the coding framework to achieve these optimum tradeoffs.

8. REFERENCES

[1] Feng Wu and Ya-Qin Zhang, "A framework for efficient progressive fine granularity sclable video coding," *IEEE trans on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 332–344, 2001.

[2] Ron Yang, Feng Wu, Shipeng Li, Ran Tao, and Yue Wang, "Efficient video coding with hybrid spatial and fine-grain snr scalabilities," *SPIE Visual Communicationa ans Image Processing*, 2002.

[3] Christian J. van den Branden Lambrecht and Olivier Verscheure, "Perceptual quality measure using a spatio-tmporal model of the human visual system," *Proceedings of SPIE*, 1996.

[4] A. Webster, "An objective video quality assessment system based on human perception," February 1993.

[5] Raj Kumar Rajendran, Mihaela Van Der Schaar, and Shih-Fu Chang, "Fgs+: Optimizing the joint snr-temporal video quality in mpeg-4 fine grained scalable coding," *ISCAS*, May 2002.

[6] M. Ghanbari, "Wireless video," 2000.

[7] Shih-Fu Chang and David G. Messerschmitt, "Manipulation and compositing of MC-DCT compressed video," *IEEE Journal of Selected Areas in Communications*, vol. 13, no. 1, pp. 1–11, 1995.

[8] Antonio Ortega, "Optimal bit allocation under multiple rate constraints," in *Data Compression Conference*, 1996, pp. 349–358.

[9] Qi Wang, Feng Wu, Shipeng Li, Yuzhuo Zhong, and Ya-Qin Zhang, "Fine-granularity spatially scalable video coding," *IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2001.