FINE-GRANULARITY SPATIALLY SCALABLE VIDEO CODING

Qi Wang[(a)], *Feng Wu*[(b)], *Shipeng Li*[(b)], *Yuzhuo Zhong*[(a)], *Ya-Qin Zhang*[(b)]

(a) Computer Science Department, Tsinghua University, Beijing, China, 100084
(b) Microsoft Research China, Beijing, China, 100080

ABSTRACT

In this paper, we proposed a novel architecture for spatially scalable video coding, namely, Fine-Granularity Spatially Scalable (FGSS) coding. The traditional layered spatially scalable coding provides only coarse scalability in which bit-stream can be decoded only at a few fixed resolution, but not something in between. The proposed FGSS scheme provides fine-granularity property to the spatial scalability. In this scheme, bit plane technique is combined with spatial scalability, thus a fine granularity increase in the image quality from low-resolution to high-resolution can be obtained. In addition, the proposed scheme provides a flexible embedded bitstream that can be decoded up to any point in the enhancement layer bitstream from low-resolution to high-resolution. This feature further enables efficient video streaming over the Internet where the scalable bitstream can adapt to the widely fluctuated bandwidth. The FGSS coding scheme extends new functionalities such as the multi-resolution, fine granularity, channel adaptation and error-recovery properties to scalable video coding, thus can satisfy different user clients with a wide range of channel bandwidth and screen resolution.

1. INTRODUCTION

In the Internet streaming video applications, the servers normally have to serve a large amount of users with different screen resolutions and network bandwidth. When the users' screen resolution is too small and/or when the bandwidth between some users and the server is too narrow to support higher resolution sequences (therefore higher bit rates), the spatial scalability coding is needed to provide different resolutions and widen bit rate range to accommodate different users. Several spatially scalable coding schemes have been proposed and accepted by some major video coding standards, such as H.263 [1], MPEG-2 [2], and MPEG-4 [3]. In general, the traditional spatially scalable coding schemes normally only provide two layers of video, either low-resolution video or high-resolution video.

A typical architecture of the traditional spatially scalable coding is depicted in Figure 1, in which there are two layers. In the base layer, the first frame is an intra (I) frame, and other frames in this Group of Pictures (GOP) are forward-prediction (P) frames. P frames are always predicted from the previous I frame or P frame. In the enhancement layer, only the first frame is a forward-prediction (P) frame predicted from the current up-sampled base layer, and other frames are bi-prediction (B) frames predicted from the enhancement layer of previous P frames (or B frames) and the up-sampled base layer in the current frame.

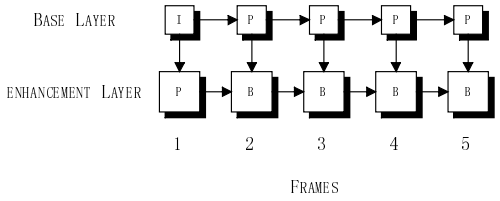


Figure 1 The architecture of the traditional spatially scalable video coding

One of the hardest problem streaming video applications over the Internet facing is that the network bandwidth fluctuates in a wide range from time to time. Since the traditional spatially scalable coding schemes only provide a coarse scalability (normally only two layers), it is very difficult for them to dynamically and precisely adapt to the channel bandwidth fluctuations and provide an optimal (best effort) video quality to each individual receiver. On the other hand, if an enhancement layer is not available in the decoder due to transmission errors or packet losses, all enhancement layers in frames followed cannot be correctly decoded until the next GOP. Moreover, for the low-resolution images, the traditional spatially scalable coding only provides one fixed base layer video to the users. Even these users might access the server with sufficient bandwidth, they could only obtain the lowest picture quality limited by the spatial base layer.

It is well known that bit-plane coding technique can provide fine granularity scalability and generate an embedded bit-stream, which has been extensively used in the SNR scalable coding, such as Fine Granularity Scalable (FGS) video coding in MPEG-4 [4] and EBCOT image coding in JPEG2000 [5]. Naturally, if the same technique is applied to the spatial enhancement layer coding, the enhancement bit-stream generated should have the similar property as in the FGS scheme, i.e., an

*This work has been done while the author is with Microsoft Research China.

embedded bitstream with fine granularity scalability. Furthermore, the low-resolution spatial base layer video can also be enhanced with bit plane coding technique. Obviously, integrating bit-plane technique to spatial scalability coding can offer enhanced flexibility and functionality to the traditional spatial scalability schemes.

In this paper, we propose a flexible and highly efficient spatially scalable video coding architecture, namely, Fine-Granularity Spatially scalable (FGSS) coding. The proposed scheme can provide multiple resolutions, bandwidth adaptation, and error recovery functionalities. Unlike the traditional spatially scalable coding, video quality of low-resolution and high-resolution in the proposed scheme is gradually and smoothly improved while bit rate increases.

The organization of the rest of this paper is as follows. In Section 2, we introduce the basic idea of how to design the FGSS coding architecture. The key techniques used in the enhancement layer coding are described in Section 3. Some experimental results are given in Section 4. Section 5 concludes this paper.

2. THE FGSS CODING ARCHITECTURE

There are two references used in the traditional spatially scalable coding. All base layers are predicted from the reconstructed low-resolution video, and all enhancement layers are bi-direction predicted from the previous high-resolution layer and/or the current low-resolution layer. Since the base layer is a low-resolution video in the spatially scalable coding, if both the base layer and enhancement layers are predicted from the previous low-resolution layer, the coding efficiency of the FGSS scheme will be very low due to the low quality and low-resolution reference used in motion compensation. Therefore, unlike in the FGS scheme[4] (a quality scalable coding), where only one reference and one motion compensation unit are enough, the first key point of the FGSS scheme is that it must be a two-loop coding scheme, i.e., there are two references and two motion compensation units.

In the FGSS scheme, the base layer is still encoded as in the traditional spatially scalable coding. However, the enhancement layer now has to adopt the bit-plane coding technique. Since the enhancement layer in the traditional spatial scalability is always a high-resolution video, if the bit-plane coding technique is directly applied to the enhancement layer coding, only the quality of high-resolution video could be fine-granularly increased. In many situations, users would also hope to obtain different quality video in the low-resolution level as well. For an example, users with low-resolution screen could obtain a higher quality video if they access a server with sufficient bandwidth.

On the other hand, when the bit plane coding technique is applied to the spatial enhancement, the enhancement bit-stream is fine-granularity scalable and can be decoded at any bit rate. However, when the bit rate of the spatial base layer is very low, if the high-resolution video from the current frame is directly up-sampled from the low quality and low-resolution video, there are serious blocking artifacts in this reference. But if several low-resolution enhancement layers are first used to enhance the video quality at the low-resolution level, and then up-sampled to the high-resolution, a better high-resolution visual quality can be obtained. Therefore, the second key point is that the enhancement layers may include the enhancement to both the low-resolution video and the high-resolution video. The lower enhancement layers are used to improve the quality of the low-resolution video, and the higher enhancement layers are used to improve the quality of the high-resolution video.

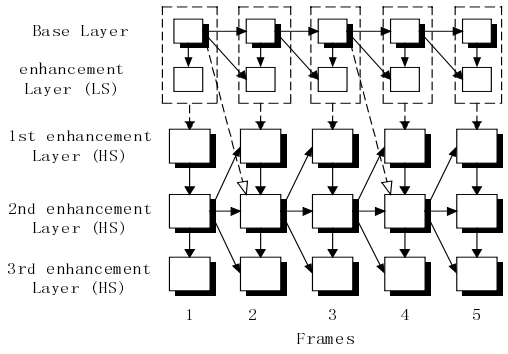


Figure 2 The proposed framework of FGSS coding

Figure 2 conceptually illustrates the architecture of the FGSS coding scheme. In this architecture, input video sequence is first down-sampled and compressed to a given bit rate with any existing non-scalable coding techniques. In the traditional spatially scalable coding, spatial resolution is switched immediately to high-resolution for the enhancement layer coding. However, the FGSS scheme provides a flexible method scaling from low-resolution to high-resolution. If the bit rate of the base layer is very low (e.g. 32kbit/s), several FGS[4] lower enhancement layers are first used to improve the video quality at the low-resolution level. If the quality of the low-resolution video is good enough in the base layer, the spatial resolution can also be immediately switched to high-resolution in the enhancement layer. Therefore, the enhancement layers at low-resolution level are optional. It depends on the bit rate of base layer, sequence contents, and application requirements. How to choose the most suitable point to switch from low-resolution to high-resolution is an encoding optimization issue.

3. THE ENHANCEMENT LAYER CODING OF FGSS SCHEME

The most different part between the FGSS scheme and the

traditional spatially scalable coding scheme is the enhancement layer coding. In the tradition spatially scalable coding scheme, the enhancement bit-stream should only be completely transmitted and decoded or completely dropped, otherwise, partially decoded enhancement layer will cause serious error drifting. However, since the bit-plane coding technique is used in the FGSS scheme, the enhancement layer bit-stream can be decoded at any bit rate that fits the available channel bandwidth.

In general, there are many bit planes at the high-resolution level. Therefore, the first problem is how to choose the high-resolution reference in the enhancement coding, i.e., which enhancement layer the high-resolution reference should locate. The first enhancement layer after up-sampling to the high-resolution level cannot be selected as the high-resolution reference, because bits used in this layer are relative less. The reconstructed quality of this layer is only slightly better than that of the up-sampled low-resolution reference. Therefore, using this layer as the high-resolution reference cannot achieve high coding efficiency. The higher enhancement layers are not good references either, because the bit rate of higher enhancement layers is too high for most applications and it is generally not always available to the receiver. Therefore, the second or third enhancement layer in high-resolution level is often used to reconstruct the high-resolution reference.

The high-resolution video encoded in the enhancement layer has three different prediction modes similarly to B frames. In the first mode, the current macroblock is predicted from the previous high-resolution reference, and the data encoded at the low-resolution level are subtracted from the predicted residues as well. In the second mode, the current macroblock is still predicted from the previous high-resolution reference, but those data encoded at the low-resolution level are not subtracted from the predicted residues. In the third mode, the current macroblock is predicted from the reconstructed low-resolution image in the current frame.

When the bit plane coding technique is applied to the spatial enhancement layer coding, another problem is how to terminate or reduce the drifting error, which could possibly appear due to network bandwidth fluctuations. When the enhancement bit-stream of the FGSS scheme is transmitted over the Internet, the network bandwidth fluctuates in a wide range from time to time. If the high-resolution references in some frames are partially or completely dropped during transmission, the decoded high-resolution reference will differ from that in the encoder. This will inevitably cause the error drifting.

A method proposed in [6] can effectively reduce the drifting error. The basic idea is that the high-resolution reference should be alternatively reconstructed from the previous low-resolution level and the previous high-resolution level to reduce the drifting error. When the high-resolution reference is reconstructed from the previous low-resolution level image, the encoder and decoder can always obtain the same temporal prediction. Therefore, the drifting errors propagated from the previous frames are eliminated. An example is shown in Figure 2. The high-resolution references in even frames are reconstructed from the low-resolution level in previous frames (indicated by the dash line and hollow arrow).

4. EXPERIMENTAL RESULTS

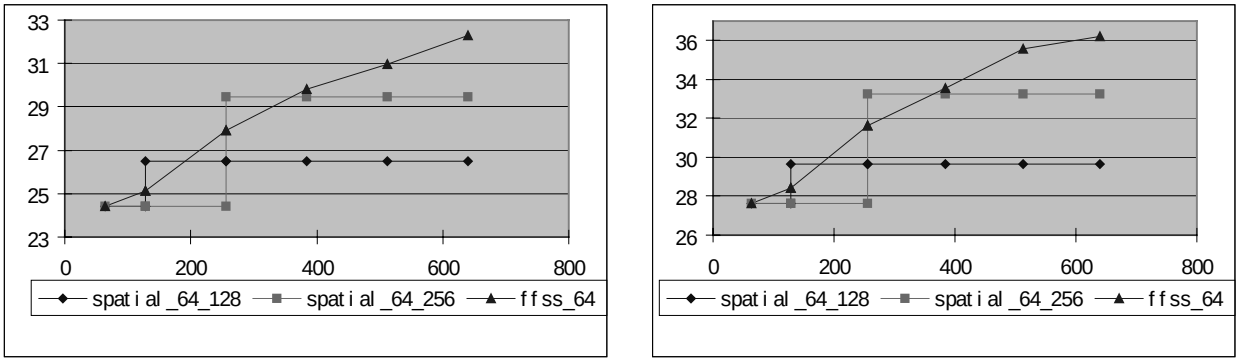
Two experiments have been performed to verify the performance of the proposed FGSS scheme. The FGSS architecture used in the two experiments is shown in Figure 2. In the first one, video sequence is directly up-sampled to high-resolution in the enhancement layers. In the second one, the video quality of low-resolution is first improved by several low enhancement layers, and then scaled from low-resolution to high-resolution.

The MPEG-4 test sequences Foreman and Coastguard (CIF format) are used in the two experiments. Only the first frame of each sequence is encoded as an I frame, and other frames are encoded as P frames. The base layer is encoded with the MPEG-4 natural visual coding scheme. A simple half-pixel motion estimator independently gets the motion vectors between two adjacent images in low and high resolution. The ranges of motion vectors are set at ± 15.5 or ± 31.5 pixels in low and high resolution, respectively.

In the first experiment, the bit rate of base layer is 64kbits/s with TM5 rate control, and the encoding frame rate is 10Hz. The bit rate of enhancement layers is 128kbits/s or 256kbits/s in the traditional spatially scalable coding. In the FGSS scheme, the bit rate of enhancement layer is not constrained. From the experimental results shown in Figure 3, the traditional spatially scalable scheme only provides two different video qualities. Only within a small bit rate range after video sequence is switched to high-resolution, the coding efficiency of traditional spatially scalable is higher than that of the FGSS scheme. On the other hand, the bit rate of the FGSS scheme can be arbitrarily selected within a wide range. Meanwhile, video quality of the FGSS scheme can be consistently improved while the bit rate increases. This means that the proposed scheme has better bandwidth adaptation capability.

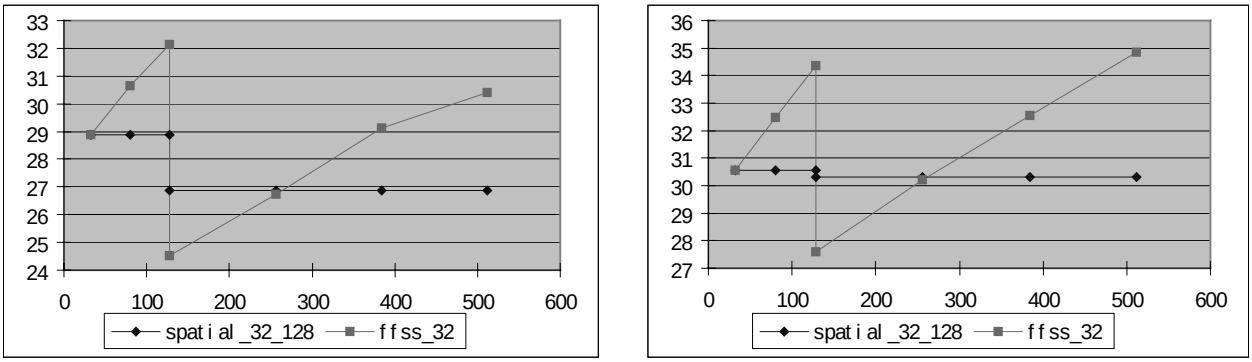
In the second experiment, the bit rate of base layer is 32kbits/s. Since the bit rate of base layer is too low, several enhancement layers are used to first improve the low-resolution video. When the enhancement bit rate is higher than 96kbits/s, the decoded video sequence is switched to high-resolution. The bit rate of the enhancement layer in the traditional spatially scalable coding is 128kbits/s. The experimental results are given in Figure 4. The decoded video is in QCIF format when bit

rate varies from 32kbits/s to 128kbits/s. When the bit rate is higher than 128kbits/s, the decoded video is in CIF format. Although the video quality of traditional spatially scalable coding is higher than that of the FGSS scheme up to 2.0dB at the 128kbits/s, the visual quality of the FGSS scheme is not bad because there are serious blocking artifacts in the tradition spatially scalable coding. However, with the bit rate increasing, the quality of the FGSS scheme is consistently improving while the traditional spatially scalable coding is limited by the quality at 128kbps/s bit rate.



(a) Coastguard (b) Foreman

Figure 3 The PSNR versus bit rate comparison between FGSS and the traditional spatially scalable coding for Y component.



(a) Coastguard (b) Foreman

Figure 4 The PSNR versus bit rate comparison between FGSS and the traditional spatially scalable coding for Y component.

5. CONCLUSIONS

This paper proposed a novel framework for fine-granularity spatially scalable video coding (FGSS). Compared with traditional spatially scalable scheme, new functionalities are extended to the spatial scalability, such as fine granular scalability, good bandwidth adaptation and error recovery, these new functionalities enable flexible video streaming over the internet to different users with a wide range network bandwidth and screen resolutions.

REFERENCES

[1] ITU-T International Telecommunication Union, “Draft ITU-T Recommendation H.263 (Video Coding for Low Bit Rate Communication)”, KPN Research, The Netherlands, Jan., 1995.

[2] ISO/IEC International Standard 13818-2, “Generic Coding of Moving Pictures and Associated Audio Information: Video”, Nov., 1994.

[3] ISO/IEC International Standard 14496-2, “Information Technology – Generic Coding of Audio-Visual Objects, Part 2: Visual”, MPEG98/N2502a, Oct., 1998.

[4] W. Li, “Streaming video profile in MPEG-4”, IEEE trans. Circuit and systems for video technology, special issue on streaming video (accepted).

[5] ISO/IEC, “JPEG 2000 Verification Model 8.5 (Technical Description)”, Sept., 2000.

[6] F. Wu, S. Li and Y.-Q. Zhang, “A framework for efficient progressive fine granularity scalable video coding”, IEEE trans. Circuit and systems for video technology, special issue on streaming video (accepted).