A Highly Rate-Scalable System Based on MPEG-4 Spatial-Temporal-FGS Video Coding

Yi-Shin Tung, Ja-Ling Wu, Hsi-Kang Tsao, Chi-Hui Huang
Communication and Multimedia Laboratory, Department of Computer Science and Information Engineering
National Taiwan University
#1 Roosevelt Rd. Sec. 4, Taipei
Taiwan 106, R.O.C.

Abstract: Video scalability is the most important functionality provided in MPEG-4. Among all the available scalabilities, rate scalability is the most fundamental and practical one. It enables a single video stream applicable to different bandwidth environments. In this paper, we address a four-layer coding system based on MPEG-4 syntax, and extend its semantic by proper combination of the spatial, temporal and SNR scalabilities. The accomplished system provides video compression ratios from 5 to 1000, and can be fine-grain adjusted. The technology may benefit many applications such as network streaming, editing preview and content analysis.

Key-Words: Scalable video, MPEG-4, FGS, Hybrid scalability, Dependency graph

1 Video Scalabilities
To make the video resources more flexible, the coding efficiency is no longer the only consideration while designing the codec. In the real world, same videos may need to be universally accessed and delivered. A movie film may be stored in the high-quality beta-cam tape played in the theater, or in DVD or VCD discs used personally at home; a conferencing video may be transmitted through LAN, Internet, or wireless networks at different bitrates. In order to meet these requirements, ISO included spatial, temporal, and SNR scalabilities in the MPEG-2 standard. It gifts a single compressed video stream the ability of being used in all kinds of aforementioned environments.

The latest international coding standard, MPEG-4 [1], includes a general scalability framework. It supports temporal and spatial scalabilities in its first version, and treats the SNR scalability as a special case of spatial scalability (when vertical and horizontal resampling factors are set to 1). Besides, it augments the fine granularity scalability (FGS) in the MPEG-4 amendment 4 [2], which is a special kind of SNR scalability using bitplane DCT coding.

The spatial scalable encoder can compress one video clip to two-layer streams, which is originally designed for the interworking of TV and HDTV to deliver TV and HDTV programs simultaneously. The first stream (base stream) carries the low spatial resolution content, while the other (enhancement stream) takes the part to increase the spatial resolution of the content. The total occupation of the two streams is much less than the sum of separately encoded streams. All research about the scalabilities tries to orthogonally decompose the streams such that they will not lose the coding efficiency too much. However, due to the property of using the same fixed block (8x8) DCT in different spatial resolutions and Laplacian Pyramid in downsampling process, it is impossible to produce the layered streams without introducing some rate overhead as compared to the wavelet-based codec.

Temporal scalability is proposed later. It provides the intermediate stage for migration 30 fps HDTV to 60 fps higher temporal resolution systems. Temporal consecutive frames are usually similar with each other and provide an intuitive way for data compression. Thus, in MPEG standard series, motion estimation and compensation are the major techniques for temporal redundancy reduction. In order to take full advantage of similarity between the neighboring frames, bi-directional prediction (B frame) is frequently used. Consequently, the temporal scalability will not seriously affect the coding efficiency if the coding patterns of pictures in each layer are carefully designed. For example, if the non-scalable stream adopts the pattern “IBPBP…”, the bitrate will equate to the two-layer streams using patterns “IPPP…” and “BBB…”. Therefore, it seems a good strategy to adopt B frame coding to achieve temporal scalability.

In addition to the basic scalabilities mentioned above, MPEG-2 and MPEG-4 also support the so-called hybrid scalability, the combination of these basic scalabilities. It can benefit more applications in different fields. However, the integration of different scalabilities will result in a high-cost design. Fortunately, the advanced silicon manufacturing has made the high-performance and low-priced
processors available today. Thus, it is the time to bring the hybrid scalable system into effect. The latest FGS is a special kind of SNR scalability. The concept of SNR scalability comes from the well-known image-coding standard, JPEG. In the JPEG progressive mode, different quantizers are used to generate images with different qualities. In the general scheme, coarse quantizer is adopted first in the base stream, and the quantization error is further coded under another finer quantizer. Based on this scheme, FGS takes a particular set of quantizers for each stage where every element in the set is an integral power of 2. The accumulation of the bitplane coding results forms the original DCT residues. By transmitting bitplanes according to their significance, the decoding process can gradually refines the DCT coefficients and a fully scalable bitstream is obtained. In this paper, we address a highly rate scalable system based on the MPEG-4 syntax, which integrates spatial, temporal, and FGS scalabilities. The general architecture of our design consists of arbitrary layers of streams. To reduce the implementation complexity, and to make a demonstrative example, we implemented a four-layered system. The compression ratio of the exemplar system ranges from 5 to 1000, compared with non-scalable MPEG-4 system that ranges from 5 to 50. Besides, we augment an amendment FGS stream such that the whole compressed streams can be partially transmitted at byte precision. The organization of the paper is as follows. In Section 2, we first present the bitrate controlling strategy, which reduces the source data to be compressed and adjusts the picture quality simultaneously. After making some simple perceptual assessment, we also decided the properties of each layer, for example spatial, temporal, or SNR enhancement. In Section 3, the combined spatial-temporal-FGS encoding/decoding system and the coding patterns designed for fast error recovery are described. We also address an efficient pipeline example. By using an augmenting stream, the delivering strategy proposed here can make full use of all available bandwidth. In Section 4, some experimental results about coding efficiency and its performance are given. Finally, the future work toward the perceptually scalable system is described in section 5.

2 Hybrid Source and Compression Data Rate Control

Compression techniques enable video contents to be stored in a reasonable space. To reduce the large quantity of video data, the general compression scheme will inevitably introduce some distortions. The reconstructed video is then not fully identical to the original video. Fitting the video data rate to some desired target, the quantizer adjustment is the basic controlling tool commonly used in the encoding process.

In the case of high bitrate video coding (such as MPEG-2 in DVD), the produced artifacts are usually hidden within several insignificant bits, and is almost insensible to the observers. The frequency-domain contrast sensitivity of human perception gives the theoretical foundation of this lossy video coding [3]. However, if we continue decreasing the bitrate of output stream, it will inevitably produce some visible distortions. The most celebrated distortions, which occurred in the MC-DCT based codec, are block effects, blurring effects and grain effects. These effects are vexatious, and often appear in the high-motion clips of VCD, which applied the constant bitrate strategy. If the desired bitrate needs to be reduced further, the video quality may become worse and worse, and even become unacceptable to the observers. Putting away the quality issues, there still exists a low bound of bitrate adjustment in most of the existing MPEG encoders. This is due to the limitation of quantization scale in the MPEG syntax. The quantization scale cannot be set larger than the value of 31. Thus, it cannot lower the bitrate any further, if there is no special handling such as changing the quantization matrix, forcing the skipped MB, or even frame dropping. In summary, quantizer adjustment is the basic tool for bitrate control in MPEG codecs. Coarser quantization may lead to lower bitrate, poor quality, and subjectively lower PSNR.

In addition to quantizer adjustment, source data reduction is another way used for low bitrate coding. For example, video conferencing systems usually adopt QCIF sized video at 10 fps. Multi-resolution representation offers us an idea for such processing. In general, decreasing spatial or temporal resolutions is a shortcut to reduce the data source so as to lower the compressed bitrate. It is intuitive that halving the width and height of video quarters the amount of video data, while halving the frame rate reduces half of the video data. The process of spatial downsampling produces smaller viewing size. To browse the video under the original resolution, upsampling of output video is required. Spatial downsampling results in a smaller spatial resolution or a blurring video in original resolution. A lot of research about perceptual upsampling has been done [4], and may result in a better visual quality. These algorithms can be adopted arbitrarily for display purpose only, because the upsampled frames will not
be referred in the prediction loop. Temporal downsampling interleaves the video into different categories. The Base-Layer encoding process only compresses the frames in the first set. Due to low temporal resolution, the motion jitter may be perceived under high motion clips. Unlike the spatial upsampling, there is no good temporal interpolation method publicly. Therefore, to eliminate this undesired jitter effect, representative frame selection should be performed instead of equal-spaced sampling to eliminate possible perceptual quality loss. However, in this paper, we still use simple spatial upsampling and equal-spaced temporal subsampling processes for simplifying the problem, and leave these possible issues for later research. Therefore, we know that the compressed bitrate can be adjusted by changing quantization scale, spatial resolution or temporal resolution. We can generate bitstreams of various bitrates by applying different values to the three-parameter scalability set. It can be treated as a multi-resolution representation along the three axes: spatial size, temporal frame rate, and the precision of coefficients (SNR). The three orthogonal axes construct a 3-dimensional space. If we map the full-quality video into the space, a cube that represents the quality of source video can be obtained. The reconstructed video will undoubtedly be part of the full volume, since either the value of spatial, temporal, or SNR resolutions may be shrunk. In the layered coding, each layer of video can be represented by a smaller cube, whose volume is decided by its corresponding resolution parameter vector. The union of all non-overlapped cubes can restore the volume corresponding to the full resolution video. It is obvious that the reconstruction quality of layered decoding process can be equivalent to any non-scalable stream generated by traditional non-scalable encoder, if their corresponding volumes are the same. The goal of scalable coding is to increase video quality critically. That is to say, any surplus bitrate should be spent in such a way that the quality can be improved the most. Following this idea, to design a multi-layered coding system, we have to decide the three-dimensional vectors for all layers based on human perception. These vectors inherently constrain the quantity of source video and quantization scale of encoding process.

In the existed literature, most research about human perception is based mainly on single axis analysis. It is hard to assess or find out a good multi-axes measurement. Although there is no apparent theory helps us to set the vectors, some evidences can still be found. 1) Most people cannot observe the distortion when the least 3 bits of pixel intensity are rounded off. 2) When viewing medium-motion videos, some examinees do not disgust the 15 fps videos, or even cannot distinguish them from 30 fps videos if no supervised training given in advanced. 3) The upsampled video can be acceptable only if the quality of the small-sized video source is originally good. All these inspections give us the hints that can help us to decide the refining path among the video layers.

As a result, in our four-layer coding system, we would like to enhance the spatial resolution first, then the temporal resolution, and finally the SNR improvement. Besides, in order to achieve the widest range of rate scalability, the data rate of base layer stream should be set as low as possible. In our implementation, all parameters for the base stream such as spatial resolution, frame rate and quant. scale, \( q_b \), are firstly set to their minimum values. Then, the bitrate is spent on the first enhancement layer, where the spatial resolution enlarges four times and the picture quality is also enhanced by using a smaller quant. scale, \( q_s \). In the third layer, the video frame rate is doubled while the quant. scale for the temporal resolution stream, \( q_t \), is slightly finer than \( q_s \), such that the quality of the consecutive frames will not vary too much. Up to this layer, the video has been brought to the highest spatial and temporal resolution. The rest of bitrates are used for increasing the video SNR. Fig. 1 depicts the representative cubes corresponding to layers of the bitstreams generated by our proposed codec structure.

![Fig. 1: The Representative Cubes of the Proposed Four-Layer Video Coding System.](image)

### 3 The Hybrid Spatial-Temporal-FGS Coding

According to the preferred layer definitions described in the last section, the properties of the four layers are set to be the coarsest quality, spatial enhancement, temporal enhancement, and SNR.
enhancement, respectively. We will now address the corresponding pairs of layered architectures for encoding and decoding processes. Figs. 2 and 3 show the corresponding block diagrams of our design. In the encoding process, the video source is first processed by the temporal splitter and the spatial downsampler to reduce the amount of source data. The resulting videos are then treated as the source for each layer accordingly, and are operated independently. The spatial-temporal reduced video flows into the base layer, while the temporally interleaved videos are sent to the second and the third layers as the inputs of the corresponding spatial and temporal enhancement modules. The quantization errors of DCT coefficients in these layers are fed into the last layer for further quality enhancement. For the object of adjusting bitrate arbitrarily during transmission, FGS is used to enable random truncation of bitstream. Besides, the reconstructed frames of the base layer are upscaled in compliance with the MPEG-4 specification, such that they can be treated as possible referencing candidates for the spatial enhancement layer to remove data correlation between the first two layers. The decoding process inverses the operation of the encoding process. However, if only the base-layer stream is delivered, the output video can be upscaled by a proprietary algorithm when the full-resolution display is desired. Moreover, when temporal enhancement is applied, the temporal combiner can reorganize all output frames to the full frame-rate video. Further quality improvement is realized at last stage after decoding the available partial FGS streams.

In the figured scenario, the encoder is processed offline, finding optimal coding parameters (motion vector, MB type, and etc.), generating the full set of bitstreams from lowest video quality (lowest bitrate) to the highest which may be almost identical to the source video, and producing the indexes of all access units (AUs). The stored archive can then be accessed by the provider, such as streaming server, and transmitted any possible amount of data to the client for best presentation. The decoding process consumes the available data under the time constraint, and required to be designed efficiently. We can observe that the layered architecture of decoding process can be realized efficiently using pipeline and parallel implementation. Fig. 4 shows the possible timeslot of the decoding pipeline. By adding the time shift to each stage, the circuits for each layer can be independently operated and scheduled. Thus, the full process is decomposed, and the difficulties of implementing the multi-layer decoder are overcome. In the decoding process, there are two framestores to memorize the referable frames belonging to different spatial resolutions. In the simple streaming applications, the uni-directional prediction is adopted for less memory requirement and low delay intrinsically. Fig. 5 depicts the dependency graph of all AUs and the reference-select-code defined in the MPEG-4 under this case. It is obvious that only one frame needs to be stored in each resolution. In the case of applications whose coding efficiency is important, such as home video, the bi-directional prediction must be used. This is a tradeoff between coding efficiency and implementation complexity. Fig. 6 represents the dependency graph of all AUs in this case. To quickly recover to higher layer quality when the bitrate variation occurs, the P and B coding modes are used one after another. It is helpful to quickly build a recovery path when some referable AUs are lost, as the idea delivered in the PFGS [5]. The four-layer system can make a stream possess high scalable range, but still cannot be matched to arbitrary desired degree (e.g. in byte). To make the stream meet any desired budget within the allowable range, an amendment FGS stream is augmented and a delivery strategy is applied to the streaming process. The augmented stream refines the base-layer video. The amount of augmented stream is small in size and only required to be equal to that of the second layer (spatial enhancement layer in our design), which can replenish the discontinuity between the bitrates of the first layer and the second layer.
The delivery strategy properly selects partial bitstreams from different layers to meet the desired bitrate constraint at byte precision. The server side applies this strategy among all AUs within the allowable lookahead window. In our basic assumption, the minimum bandwidth should be supplied and exceeds the rate of the base layer (Path 1 in Fig.7a), which can be very low in our design. When the bandwidth is between the first and second layer, the augmented FGS is selected. Because the FGS AUs can be inherently truncated at any location, the bitrate can be precisely adjusted (Path 2 in Fig.7a). The augmented FGS AUs are discarded when the available bandwidth exceeds the sum of the first two layers. The AUs of the second layer are delivered instead under this situation (Path 3 in Fig.7a). The same rule is applied when the bitrate is less than the third temporal layer. We adopt the corresponding FGS AUs (Path 4 in Fig.7b) until the first three layers can be fully delivered (Path 5 in Fig.7b). Utilizing the leftover bandwidth, the FGS AUs of all frames are delivered last (Path 6 in Fig.7c). By this strategy, both the frame rate and quality of each frame will vary with the available bandwidth at that moment. The overall perceptual quality increases when network bandwidth becomes higher.

Under the proposed strategy, the selected AUs at the same instance will be multiplexed for transmission. The multiplexed units are then sent in accordance with their decoding order. Fig. 8 represents the transmitted order when only P frame coding is used. If B frame coding is used in the enhancement layers, the referable AUs are sent first to guarantee the casality.
4 Experimental Results
In our experiments, we compress several source videos by our proposed system and non-scalable MPEG-4 encoder. In our system, we set the base-layer video to the quarter spatial resolution, half frame rate, and parameters of quantization scales, $q_b$, $q_s$, $q_t$ to 31, 16, 16, respectively. For comparison purpose, the non-scalable encoder adopts constant quantization mode with the scale of 16. In general case, the base-layer compression ratio in our system is around 1000, and it is superior to the maximum compression ratio, 50, of the standard MPEG-4 encoder. Compared with the MPEG-4 non-scalable stream, the PSNR of output video that decodes the first three layers is almost the same, and the total size is, in average, 20% larger. The 20% overhead is reasonable when layered coding is used. Besides, by integrating the augmented stream, the bitrate can then be adjusted at byte precision within the range of compression ratio from 5 to 1000. For example, a 720x480-sized video can be scale from 100Kbps to 10Mbps.

5 Conclusions and Future Works
In this paper, we proposed a hybrid scalable system, which integrates the spatial, temporal and latest FGS scalabilities. We adopt the syntax of the newest International Standard, MPEG-4, such that the high coding efficiency and possible future extension may easily be attained. The output bitstream of proposed system is a fully scalable bitstream. Its compression ratio is ranging from 5 to 1000 and the average PSNR is around 20 to 44. Such a highly scalable codec makes a single bitstream applicable anywhere. Besides, by introducing an amendment stream and the devised delivery strategy, the bitrate can then be adjusted arbitrarily within the full range, while only 10% extra storage is required. These attractive features make our proposed scalable codec suitable for bandwidth varying environments or bandwidth sharing networks. In our current study, we do not include the research about human perception, such as perceptual quantizer selection for regions of different properties, representative frame categorization, or even inter-axes criteria measurements for the optimal layer definition. We will put emphasis on these topics in the near future, which enable the generated streams to meet perceptual assessment.

References:

Fig. 7: The AU Delivery Strategy to Optionally Transmit Bitstreams of Different Layers: (a) when the bitrate is below the sum of the first two layers, (b) within the range from the sum of the first two layers to the sum of the first three layers, (c) above the sum of first three layers.


