A Novel Scalable Video Codec Based-on MPEG-4 Visual Texture Coding

Chi-Hui Huang, Yi-Shin Tung, and Ja-Ling Wu, IEEE Senior Member

Communication and Multimedia Laboratory
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan China

{cindy, tung, wjl}@cmlab.csie.ntu.edu.tw

Abstract-**In this paper, a novel wavelet-based scalable video codec is proposed and implemented. Due to the superb coding efficiency and also the scalability that MPEG-4 visual texture coding (VTC) possessed, it is adopted as the intra-coding kernel. In addition, combined with the simplest and most effective way for dealing with temporal redundancy, motion estimation/compensation techniques, inter-coding can be integrated into our codec easily and efficiently.**

In our design, multiple scalabilities including spatial scalability, temporal scalability, rate scalability, SNR scalability, and their combinations are achieved, and they can be adjusted dynamically at the time of operation. Moreover, in order to make the video codec able to synchronize with user's resolution changing request as fast as possible, a modified hierarchical motion estimation scheme is also presented. The resultant coding efficiency of the proposed codec is comparable to that of the non-scalable MPEG-1 and outperforms most of the existed wavelet-based scalable codecs, such as SAMCoW and Taubman and Zakhor's approach.

I. Introduction

In the last decade, the popularization of Internet, the extensive use of multimedia, the development of wireless and video communication accelerate the transmission and the use of digital video data. However, due to the huge volumes of video and the limitation of available network bandwidth, it would be impractical to directly store or deliver video data without compression. Therefore, data compression has become an inevitable part of video communication. Practically, a user may request a video sequence with his specific quality, according to his display device, network environments, computational power, and preference. It is hard and inefficient for a video server to transcode the content to serve various incoming requests, simultaneously. The most intuitive way to overcome this difficulty is to store various versions of one video sequence at the server side. Nonetheless, this straightforward approach requires huge resources in terms of disk space and management overhead, which makes it especially infeasible to real-time applications. Another approach is to exploit the scalabilities of video codec, which provides a set of adjustable parameters to decide the properties and qualities of decoded video at the time of decoding. Through this means, the server only needs to compress a video sequence in detail once, and the generated bitstream can fit in with various client requirements.

Generally speaking, scalable coding techniques can be classified into three categories, including MC-DCT based, 3D-Wavelet based, and MC-Wavelet based ones. Most of the well-known and widely used video coding standards are MC-DCT based. However, when spatial scalability is also taken into account, MC-DCT coders are less efficient then wavelet-based coders [1]. This may due to the low-resolution DCT coefficients obtained by the downsampling process are not exactly the low frequency part of the original signal. That is to say, the low-pass and high-pass filters used by most of the well-known MPEG series are not critical sampled filtering pairs [2]. (Here, the low-pass filter refers to the downsampling process, and the high-pass filter is the zero-motion compensated of the upsampled low band reconstruction.)

Three-dimensional (3D)-Wavelet video c oder is another approach to realize scalable codecs. The idea of 3D wavelet-based video coding comes from the extension of 2D zerotree coding schemes. The wavelet transform is applied not only to the spatial domain but also the temporal domain. Many 2D zerotree coding schemes have been extended to design 3D wavelet-based video codecs, including 3D-IEZW[3], 3D-SPIHT[4], and so on. However, all the 3D wavelet video codecs suffer from the defect that the temporal redundancy of the video sequence cannot be exploited efficiently and the poor temporal low-resolution representation, which does not fit the human perception. Also, performing 3D wavelet decomposition requires multiple video frames to be processed at the same time; therefore, more memory is needed for both the encoder and the decoder, which results in a longer processing delay. Due to these defects, 3D-Wavelet is not suitable for video coding and is not addressed in this work.

Since spatial scalability is a must of the fully scalable codecs, the third category, MC-Wavelet, seems to be a good approach, which takes wavelet transform in spatial domain, and motion estimation/compensation techniques to deal with temporal correlations between successive frames. In this paper, we present a MPEG-4 visual texture coding (VTC) based scalable video coder belonging to the MC-Wavelet category. In our design, the hierarchical motion estimation for wavelet subbands is devised, and some modifications and strategic decisions of VTC are made to fit the requirement of the full scalability.

This paper is organized as follows. In Section II, a brief description of MPEG-4 visual texture coding [5] which is adopted as our spatial domain coding scheme is given, and then the hierarchical motion estimation scheme used to remove the temporal redundancy is presented. The detail of multiresolution encoding and decoding of the proposed scalable video coder is described in Section III. Section IV provides the modified hierarchical motion estimation scheme that helps the scalable codec to synchronize with user's resolution raising request as fast as possible while decoding. The experimental results of the proposed system are given in section V. Finally, we conclude our paper in Section VI.

II. The Proposed Scalable Video Codec

Video data is usually highly correlated both in spatial and temporal domains. In order to achieve good coding efficiency, it is necessary to remove redundant information existed within the video data. In the proposed video codec, we utilized MPEG-4 VTC to deal with the spatial redundancy of each frame (and

0-7803-7488-6/02/$17.00 © 2002 IEEE.

predicted error frame), and motion estimation/compensation techniques to remove the correlations between successive frames.

A. Spatial Redundancy Removal: The MPEG-4 Visual Texture Coding

MPEG-4 VTC is the state-of-art technique for encoding the texture of objects and still images, and it is based on discrete wavelet transform and embedded zerotree algorithm[6]. Compared with DCT-based approaches, zerotree wavelet coding provides superior coding efficiency as well as spatial and SNR (quality) scalabilities. The MPEG-4 VTC consists of the following procedures: (1) taking discrete wavelet transform, (2) generating wavelet zerotree, (3) quantizing wavelet coefficients, and (4) entropy coding the quanitzed wavelet coefficients. Moreover, due to the different properties of the lowest frequency subband (DC band) and other high frequency subbands (AC bands), the coding schemes of them are accordingly different.

There are three possible quantization modes for the AC coefficients in the MPEG-4 VTC: single quantization mode (SQ), multiple quantization mode (MQ), and bi-level quantization mode (BQ). Each quantization mode provides different degree of scalability, and takes distinct computing complexity. In order to meet the real-time constraint of video application and maintain basic frame rate for supporting acceptable video quality, SQ mode is adopted in this work. However, the other two quantization modes can easily be integrated into the proposed codec to offer better quality adjustment, as well. Besides, two different scanning orders of quantized wavelet coefficients (tree-depth and band-by-band) have been defined to serve for difference purposes. In order to support progressive transmission without causing too much overhead, band-by-band scanning order is preferred in the streaming environments.

B. Temporal Redundancy Removal: The Hierarchical Motion Estimation and Refinement

Motion estimation and compensation technique is known to be one of the simplest and most effective way to remove temporal redundancy existed within usual video data. The proposed system adopts motion estimation-based approach combined with some modifications that utilize the correlations between wavelet subbands. Since the properties of the DC band and AC bands are different, distinct strategies are applied to deal with them.

1) DC Motion Estimation

Since the DC band, which is known as the lowest frequency subband, of an image obtained by taking wavelet transform can be viewed as the spatially low resolution representation of the original image, we simply apply traditional block-based motion estimation to the DC band of each frame. On the encoder side, motion compensation is also included to form the prediction of current DC used for referencing, and the prediction is constructed by using quantized reference frame instead of the original frame to prevent error accumulation and possible drifting problems. The motion vectors obtained during the DC band motion estimation are kept as references for the further motion refinement of AC bands.

2) *AC Motion Refinement*

Observing that the motion activities between different subbands are highly correlated, we do not directly perform traditional motion estimation on AC bands. Instead, we imitate the idea of hierarchical motion estimation and compensation given in [7]. The motion vector of the previous level at the corresponding position, with proper scale, is used as the initial motion candidate for the next level. Then the motion refinement is done by searching within some predefined search range, e.g. +/-2 pixels, and merely the refined motion vectors are needed to be entropy coded. The overall hierarchical motion refinement process for AC bands is illustrated in Fig. 1. Through the proposed hierarchical motion estimation and compensation technique, not only the data rate of motion vector is decreased, the computational time can also be reduced, dramatically.

However, in our experiments the performance of applying motion estimation/compensation on higher wavelet subbands is not always good if inter-coding mode is adopted all the time. This may due to the constantly changing of high-frequency signals or fast object translations. Therefore, we provide the selectivity of intra and/or inter coding block when encoding P frames to compensate this contradictory situation. In B frame, since it is less possible that there exists no good reference in both neighboring frames, the inter-coding is always adopted.

III. MULTIRESOLUTION ENCODING AND DECODING

Scalability makes one single bitstream applicable to various kinds of environments and able be progressively decoded and transmitted. In the proposed video codec, spatial, temporal, bitrate, and SNR scalabilities are all included.

A. Spatial Scalability

The spatial scalability is intrinsic to the proposed codec, since wavelet is the transform kernel. One of the most desirable properties of wavelet transform is that the resulting DC band of wavelet-transformed image is visually equivalent to the spatially low-resolution representation of the original image. Therefore, the spatial scalability can be achieved by simply selecting suitable bitplanes and discarding unnecessary levels of high frequency subbands.

B. Temporal Scalability

Our temporal scalability comes from strategically placing predicted frames during the encoding process and selectively transmitting encoded frames while delivering. The temporally low resolution of the original video sequence is obtained by direct temporal subsampling, and the decoded frames are exactly the same as those that would be decoded at full frame rate. The resultant visual quality is better as compared with the blurred video produced by the 3D-Wavelet coders.

In order to achieve better compression ratio and higher flexible temporal scalability at the same time, the technique of B-frame partitioning is included in the proposed codec. Through this way, the realization of temporal scalability will not affect the coding efficiency. For example, if the GOP pattern "IPBPBPB..." is adopted, the bitstream can be separated into two layers with patterns "IPPP..." and "BBB...," respectively. The base-layer stream with pattern "IPPP..." obviously can be decoded independently since the enhancement-layer stream consists of merely B-frames, which would not be used as reference frames. This technique also serves a good error resilient property, because the enhancement frames only refer to those frames in the base-layer stream. The adopted structure of the temporal layers is shown in Fig. 2.

C. Bitrate and SNR Scalabilities

The bitrate and the SNR scalabilities are achieved by discarding bits from high frequency subbands when the bitrate budget as-

signment of the frame is exhausted. Since the Intra-frames and predictive error frames are coded using the MPEG-4 VTC, which utilizes the bit-plane coding technique [5][8] to code the magnitudes of coefficients from the most significant to the least significant bit-planes. Those bits that carry the most important information (to visual quality) are placed in the front part of the bitstream, and it is permissible to cut off the bitstream at any point within the frame when the available bitrate is not enough to decode the full resolution signal. Consequentially, the fine-grained bitrate scalability is achieved.

IV. Modified Hierarchical Motion Estimation

In addition to those basic scalabilities mentioned above, the proposed codec also supports the so-called hybrid scalability, which is the combination of the prescribed basic scalabilities. The codecs that provide hybrid scalability are more practical and useful. However, when integrating those basic scalabilities together we observed that the motion estimation and compensation techniques used to support the mutiple-layer-coding would always cause some problems. This situation also happens in most of the existing motion-based scalable codecs [5][9]. Considering the situation that when the available bandwidth increases during the decoding process and if the user decides to raise the spatial resolution one level higher. We notice that the codec cannot respond to the user's request immediately. This is because the decoder lacks for the one-level higher frequency subband of reference frame. Since we apply motion estimation and compensation technique, the decoding of each AC level demands for the presence of that level's reference frame(s) (except for I frames which can be decoded independently). Fig. 3 illustrates the original prediction structure of a GOP, which can eliminate most correlations between the same subbands of neighboring frames. However, following such structure, if the resolution of reference frames is at a certain level, any incoming requests higher than that cannot be achieved until the appearance of the next GOP. The same situation happens when users attempt to increase SNR scalability, but the SNR scale of reference frames is not high enough. Therefore, any requests that attempt to raise the resolution scale cannot be immediately accomplished. Actually, it can never be achieved until the appearance of next intra-frame. In order to overcome this undesirable shortage, we slightly modify the motion refinement algorithm applied to the AC bands of P frames. As depicted in Fig. 4, the motion estimation performed on DC bands of P frames remains the same, but some of the motion refinements of AC bands are disabled. For those odd numbered P frames, the motion refinements only take place on the odd numbered enhancement layers. As to those even numbered P frames, the motion refinements only apply to the even numbered enhancement layers.

With this slight modification of motion refinement process, the request for increasing spatial resolution or SNR scale can be achieved within a very short time[8]. Since the motion correlations of high frequency subbands between successive P frames are kept or removed alternatively, every one of two AC levels in a frame can be decoded independently. That is to say, the spatial layer decoding can be moved one level higher whenever encountering a P frame. There exists a quick and automatic recovery path that can increase the decoding resolution/scale from the coarsest base layer to the finest enhancement layer, as that indicated by the thickest arrow-line, in Fig. 4. The encoding flowchart of the proposed codec can be found in Fig. 5.

V. Experimental Results and Performance Analysis

In this section, we provide some experimental results of the proposed scalable coding system. Comparisons are also made with some existing scalable video coders and the international coding standard, MPEG-1. The PSNR of luminance and color components (Y, U, and V) will be calculated separately to objectively measure the quality. The coding pattern of the proposed system is set to be "IBBBPBBB...," with GOP lengths 150. The evaluation of compression performance of the proposed codec is made on several CIF format (352x288) video sequences with 30 frames per second (such as, Football, Miss America and Salesman).

The performance comparison between non-scalable MPEG-1 [9], Taubman and Zakhor's algorithm [10], Ke Shen's SAMCoW [11], and our proposed video codec is also included. Taubman and Zakhor's algorithm is based on the 3D-Wavelet coding. The SAMCoW is also a MC-Wavelet approach, but applies block-based motion estimation on time domain and utilizes wavelet-based zerotree algorithm to code the intra-frames and prediction error frames.

The resulting PSNR values of football sequence with GOP size 150 are given in TABLE I and our proposed codec is named as MCVTC (Motion-Compensated Visual Texture Coding) for short. The proposed codec outperforms SAMCoW, Taubman and Zakhor's algorithm, but is inferior to MPEG-1 for 1 to 1.5db. Experiments are also made with the Miss America sequence between the proposed codec and MPEG-1, and the result is shown in TABLE II. Although the purposed codec does not perform as well as MPEG-1 in terms of PSNR, the subjective experiments have shown that our algorithm produces comparable visual quality.

VI. Conclusions

In this paper, a novel wavelet-based scalable video codec based on MPEG-4 VTC is presented. We applied motion estimation/compensation techniques to the lowest frequency subband and motion refinement to all the high frequency subbands to effectively remove the temporal redundancy existed between successive frames of the video sequence. And the MPEG-4 VTC is adopted to code the intra-frames and predicted error frames. The resulting coding efficiency outperforms most of the wavelet-based scalable video codecs and is inferior to non-scalable MPEG-1 for at most 1.5db in average.

Moreover, our proposed codec generates embedded bitstreams and provides multiple scalabilities. The adopting of wavelet transform, which inherently possesses better spatial scalability as compared with the DCT based counterpart, gifts our video codec the spatial scalability. This gives us an idea that wavelet-based approach can be a good candidate if spatial scalability is of great concern. The temporal scalability comes out from the careful design of temporal coding pattern and selective dropping of the inter-frames. Through this means, temporal scalability is realized without introducing any overhead. By utilizing bit-plane coding scheme, precise bitrate control and bitrate scalability are achievable. Also, the scalable nature of our video coding scheme allows the decoding data rate to be dynamically changed. The ability of automatically adjusting data rate to meet network load is very appealing to network oriented applications.

In conclusion, the proposed wavelet-based video codec based on the MPEG-4 VTC achieves acceptable coding efficiency together with flexible scalabilities. It also demonstrates the potential superiority of MC-wavelet coders.

References

[1] M. D., A. L., and S. M., "Spatial-Temporal Scalability for MPEG Video Coding," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 10, no. 7, Oct. 2000.

[2] Y. S. Tung, J. L. Wu, and C. H. Huang, "Toward Better Coding Efficiency of Spatial Scalability via Separate DCT Subband Prediction," *IEEE Intl. Conf. Fundamentals of Electronics, Communications and Computer Sciences, 2002.*

[3] Y. Chen and W. A. Pearlman. "Three Dimensional Subband Coding of Video Using the Zerotree Method." *Proceeding of the SPIE Conference on Visual Communications and Image Processing, March 28-30 1996, San Jose, CA.*

[4] B. J. Kim and P. A. William. "An Embedded Wavelet Video Coder Using Three-Dimensional Set Partitioning in Hierarchical Trees (SPIHT)," *Data Compression Conference*, pp. 251-260, 1997.

[5] International Organization for Standardization (ISO), "ISO/IEC IS 14496-2, Information Technology — Coding of Audio-Visual Objects: Part 2: Visual," 1998.

[6] J. M. Shapiro. "Embedded Image Coding Using Zerotrees of wavelets coefficients," *IEEE trans Signal Processing*, vol. 41, no. 12: pp. 3445 - 3462, Dec. 1993.

[7] P. C. Chang and T. T. Lu, "A Scalable Video Compression Technique Based on Wavelet Transform and MPEG Coding," *IEEE Trans. Consumer Elec.*, vol. 45, no. 3, pp. 788-793, Aug. 1999.

[8] Microsoft Research China, "Study of a New Approach to Improve FGS Coding Efficiency," ISO/IEC JTC1/SC29/WG11 M5583, Dec. 1999.

[9] International Organization for Standardization (ISO), "ISO/IEC 11172-2, Information Technology — Coding of Moving Pictures and Associated Audio for Digital Storage Media at Up to About 1.5Mbit/s." May 1993.

[10] D. Taubman and Z. Zakhor, "Multirate 3-D Subband Coding of Video," *IEEE Trans IP*, vol. 3, pp. 572-588, Sept. 1994.

[11] K. Shem and E. J. Delp. "Wavelet Based Rate Scalable Video Compression," *IEEE Trans. CSVT*, vol. 9, no. 1, Feb 1999.

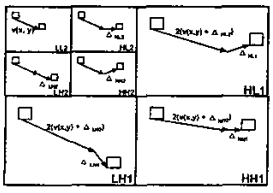

Fig. 1. AC bands hierarchical motion refinement.

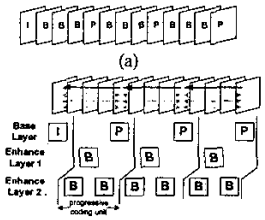

Fig. 2. The coding pattern used by the proposed system: (a) original sequence, (b) the three temporal layers.

TABLE II: The PSNR comparison of non-scalable mpeg-1 and MCVTC with CIF format Miss America sequence (30frames/sec) in various bitrates.

Components		Average	Y	U	V
1 Mbps	MCVTC	39.4	38.8	39.6	39.8
	MPEG-1	40.4	39.8	40.7	40.7
1.5 Mbps	MCVTC	40.0	39.3	40.2	40.6
	MPEG-1	40.9	40.4	40.8	41.4
2 Mbps	MCVTC	40.5	39.7	40.7	41.0
	MPEG-1	41.7	41.2	41.8	42.2

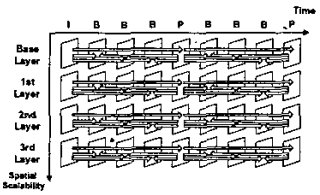

Fig. 3. The correlations between each frame and each spatial layer.

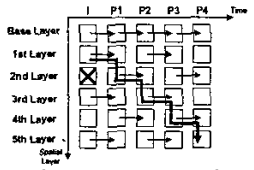

Fig. 4. The modified reference correlations of between P frames. The cross denotes that the second enhancement layer of I frame is not received by the decoder side. And the thickest arrow line denotes the automatic recovery path of the decoder.

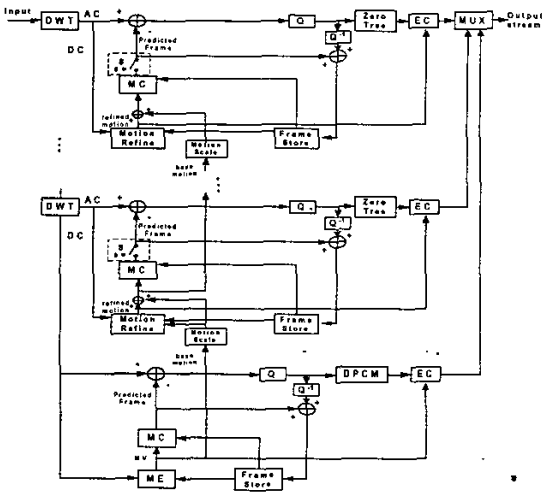

Fig. 5: The encoding flowchart of the proposed codec

TABLE I: The average PSNRs of the CIF format Football sequence (30frames/sec) with GOP size 150 in various bitrates.

Components		Average	Y	U	V
1 Mbps	MCVTC	30.5	26.4	31.1	33.9
	SAMCoW	27.1	25.8	26.6	29.9
	Taubman	25.2	21.5	28.8	32.2
	MPEG-1	31.6	27.8	32.0	35.0
1.5Mbps	MCVTC	31.7	28.1	32.2	34.7
	SAMCoW	28.1	27.1	27.7	30.1
	Taubman	26.3	22.6	29.6	32.8
	MPEG-1	33.1	29.5	33.9	35.9
2 Mbps	MCVTC	32.7	29.5	33.4	35.3
	SAMCoW	28.8	28.2	28.4	30.4
	Taubman	26.7	23.1	30.1	33.0
	MPEG-1	34.2	31.0	34.9	36.6
4 Mbps	MCVTC	36.1	33.8	36.8	37.7
	SAMCoW	30.9	30.9	30.4	31.8
	Taubman	28.9	25.0	31.5	33.9
	MPEG-1	37.1	35.0	37.6	38.8