MULTIDIMENSIONAL SVC BITSTREAM ADAPTATION AND EXTRACTION FOR RATE-DISTORTION OPTIMIZED HETEROGENEOUS MULTICASTING AND PLAYBACK*

[1]*Wen-Hsiao Peng,* [1]*John K. Zao,* [1]*Hsueh-Ting Huang,* [1]*Tse-Wei Wang, and* [2]*Lun-Chia Kuo*

[1]Department of Computer Science, National Chiao Tung University, Taiwan
[2]Industrial Technology Research Institute, HsinChu, Taiwan

ABSTRACT

In this paper, we propose an *optimal SVC bitstream extraction scheme* that can produce scalable layer representations for different viewing devices scattered over a multicasting network with diverse link bandwidth. Our scheme can determine *optimal extraction orders/paths* that are implementable at different multicasting nodes by performing successive steps of *multiple adaptation*. In addition to this basic scheme, we also develop an unambiguous denotation for the optimal extraction paths, and an algorithm for deducing the optimal extraction paths for less capable devices through path truncation. Extensive SVC encoding and adaptation experiments have been performed using JSVM 9 and both objective quality metrics (PSNR and MSE) as well as subjective metric (VQM). The experiment results showed that our scheme works well if *convexity* of R-D performance is maintained throughout the SVC bitstream.

Index Terms— Scalable Video Coding, Bitstream Adaptation, Rate-Distortion Optimization

1. INTRODUCTION

Production of scalable bitstreams that can be played back by a garden variety of viewing devices is a long pursued goal of video compression technology. The emerging scalable video coding (SVC) standard [12] achieved that goal by employing *multilayer coding* along with *adaptive inter-layer prediction* and *hierarchical temporal reference*. By encoding a video sequence into an interdependent set of *network abstraction layer (NAL)* units, an SVC bitstream allows different viewing devices to extract and decode subsets of its NAL units according to playback capability and network connectivity. How to offer different devices with appropriate NAL sets through *bitsream adaptation and extraction* thus becomes an intriguing problem.

Several approaches have been proposed to develop optimal adaptation schemes which ensure the best playback quality for viewing devices while making the best use of available transport bandwidth. All these approaches tackled the problem as a form of rate-distortion (R-D) optimization. Amonou *et al.* [3] proposed to arrange SVC Quality Increments according to their R-D contribution. The idea is similar to the creation of Quality Layers in JPEG 2000 [2] and is used also in multidimensional adaptation [6][7]. In search for the best compromise between spatial and temporal quality, many studies [4][5][8][10] showed that the peak-signal-to-noise-ratio (PSNR) or mean squared error (MSE) measurements of adapted bitstreams do not correlate well with the perceptual quality of their playback. The discrepancy becomes most obvious when both spatial and temporal quality are varied over a wide range of bit rates. To amend this discrepancy, quality metrics correlatd with perceptual viewing quality were used in recent studies [5][7]. Kim *et al.* [6], for example, resorted to use the *mean opinion scores* (MOS) of subjective viewing tests to infer the best extraction paths for different video contents. In addition to the search for optimal adaptation of a SVC bitstream, signaling mechanisms have been devised to transport the adaptation settings along with the bitstream. Three different signaling mechanisms have been proposed: (1) Priority Identifiers [12], which is a 6-bit field in the NAL unit header that indicates the extraction priority of that unit; (2) Extraction Tables [6], which specify all spatiotemporal switching points in the bitstream suitable for discretionary adaptations; (3) Scalability Information Supplemental Enhancement Information (SSEI) Messages [11][12], which describe the dependence relations existing among different SVC layers, based upon which bitstream adaptation is performed.

All these previous studies, however, address the problem from a unicasting perspective, i.e. they devise adaptation schemes only to suit the playback capability and transport bandwidth of a single device. In this paper, we consider the problem of adapting a SVC bitstream to suit various types of viewing devices that are scattered all over a multicasting network with different transport bandwidth at each link. In this situation, the adaptation scheme not only has to find an *optimal extraction order/path* that guarantees R-D optimized decoding on different viewing devices. It also has to ensure that the extraction order can be carried out at different multicasting nodes by performing *multiple adaptation steps*. Furthermore, the extraction path can also be "contracted" by deleting spatial/temporal/SNR layers that are not needed for particular device types.

In order to devise an adaptation scheme that can generate optimal extraction paths for the *heterogeneous multicasting* scenarios mentioned above, we carried out a series of experiments to examine how the optimal extraction paths of viewing devices may differ from one another owing to their differences in display formats, processing power or transport data rates. In particular, we evaluated the R-D costs of different scalable layers by interpolating their decoded videos to the spatiotemporal resolution of target devices. Our experiments showed that optimal extraction paths for different viewing devices exhibits regular, predictable patterns (especially when objective quality metrics such as PSNR or MSE are used) if different layers of a SVC bitstream were encoded to ensure *convexity* of their R-D performance curves. This result enables us to deduce the optimal extraction paths for less capable devices from those for the more capable ones and develop a succinct denotation of these extraction paths. Specifically, the contributions of this paper includes (1) a R-D optimized, multidimensional bitstream adaptation scheme based on successive scalable layer extraction, (2) an explicit, unambiguous path denotation for the optimal extraction paths, and (3) an algorithm for deducing the optimal extraction paths for less capable

*THIS WORK WAS SUPPORTED IN PART BY THE NSC UNDER GRANT 96-2628-E-009-015-MY3

Table 1. Encoder Configurations

JSVM 9_10	
Spatial Scalability	QCIF (176x144), CIF (352x288)
Temporal Scalability	Hierarchical B + GOP Size = 8
SNR Scalability	QCIF: 3 Layers, CIF: 2 Layers
Device Types	QCIF15/QCIF30, CIF15/CIF30
Settings of Qp and Inter-layer Dependency	
Akiyo	QCIF(50←43←37), CIF(44←40)
Foreman	QCIF(46←40←34), CIF(41←34)
Football	QCIF(41←35←30), CIF(36←30)
Mobile	QCIF(41←35←30), CIF(41←34)

viewing devices through truncation of the path denotation.

The rest of this paper is organized as follows: Section 2 introduces the concepts of successive layer extraction. Section 3 analyzes the R-D optimized extraction paths in different conditions. Based on the analysis results, Section 4 presents a succinct denotation and an inference algorithm for determining the optimal extraction paths for different viewing devices. We then compare our scheme with the Quality Layer extraction implemented in JSVM 9 in Section 5 before concluding the paper with a summary of our work.

2. SUCCESSIVE SCALABLE LAYER EXTRACTION

In order to support spatial, temporal and fidelity (SNR) scalability, a SVC bitstream is decomposed into an array of interdependent layer representations. The spatial, temporal and SNR dependence relations impose a *partial order*[1] among these layered representations. Decoding of a SVC representation must conform to this partial order in the sense that the decoder must obtain all the NAL units on which the *target representation* depends before it can decode that representation. In mathematical terms, all these NAL units are included in a *transitive closure* of dependence relations that originate from the target representation. If these dependence relations are depicted by a graph then all these NAL units must reside in a *convex set* on the graph with the target representation being a vertex. A convex set of NAL units are always decodable because, by definition, these units always satisfy all the dependence relations among them. In the SVC community, these decodable sets of NAL units are known as *scalable layer representations* or simply *scalable layers*.

If a viewing device needs to extract a decodable set of NAL units at every refinement step during its playback process (such is the case if the device needs to adapt itself to the fluctuation of transport data rates) or different viewing devices need to obtain decodable NAL units from a heterogeneous multicasting network then the *extraction path* must traverse *a subset-ordered sequence of scalable layer representations*[2] starting from the one containing the base layer and ending with the one containing its target representation. Furthermore, a particular extraction path is regarded *optimal* for a viewing device if and only if the corresponding sequence of scalable layers provides the best rate-distortion (R-D) performance for that device. For a multi-layer video code with incremental refinement support such as SVC, the R-D performance of a viewing device is measured by the magnitude of rate-distortion ratios at each refinement step during the decoding process. Our experiments with the setting of

[1]In an SVC bitstream, all decodable sets of NAL units actually form a *semi-lattice*, which has the base layer as its bottom element and the highest enhancement layers as its top elements.

[2]A *subset-ordered sequence of scalable layer representations* is a sequence of these representations with each representation (except the last one) being a subset of the next one.

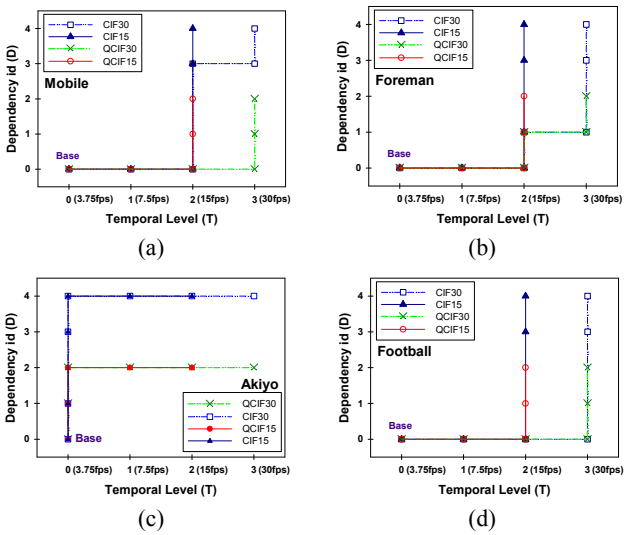

Fig. 1. Comparison of optimized extraction paths on different viewing devices: (a) Mobile, (b) Foreman, (c) Akiyo, and (d) Football. Frame replication and perceptual quality model are used in search for optimal paths.

quantization parameters Qp and dependence relations during SVC encoding [9] suggested that a SVC bitstream must satisfy the following two rules regarding its R-D performance in order to produce optimal extraction paths suitable for heterogeneous multicasting with multiple adaptation.

1. *Monotonic Reduction of Distortion in Successive Refinement*. A *proper dependency setting* for a SVC bitstream must guarantee that every subset-ordered sequence of its scalable layers representations exhibit a monotonic decrease of distortion values (in mean squared error) when these representations are decoded in order.

2. *Convexity of Rate-Distortion Curves*. A *well-adapted dependency setting* for a SVC bitstream must guarantee that every subset-ordered sequence of its scalable layers representations exhibit a monotonic decrease of the rates of distortion reduction as well as a monotonic decrease of distortion values.

The first rule implies that every plausible extraction path through an SVC bitstream should enable the decoder to improve its video quality at every refinement step. Any scalable layer representation that violates this rule contains wasted data bits because some of its NAL units do not contribute to the improvement of video quality. The second rule is even stronger. It requires that the *slope of R-D curve* decreases monotonically through each refinement step. This rule ensures the *convexity* of R-D performance curve along every extraction path of the bitstream.

On the other hand, our experiments with scalable layer extraction [§3] discovered a localized search algorithm that can construct an optimal extraction path for heterogeneous multicasting based on steepest decent search strategy if and only if the SVC bitstream adopts a well-adapted dependency setting. A localized search is made possible because the *convexity* of the R-D performance of a well-adapted bitstream eliminates the existence of any local minimum.

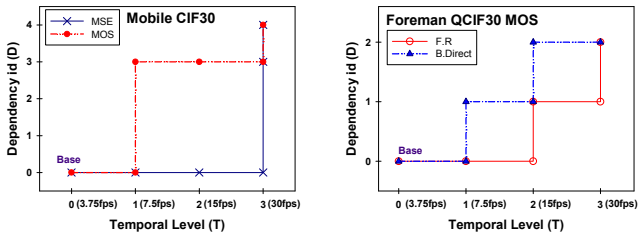

Fig. 2. Influeces of distortion measure and temporal interpolation on optimal extraction paths. (a) Comparison of MOS- and MSE-based extractions. (b) Comparison of temporal interpolation using frame replication (F.R.) and B_Direct_16x16 (B.Direct).

Table 2. Different Types of SVC Bitstream Extraction Steps

Extraction Steps			*ID Changes*		
Types	*Symbols*	*Indices*	*D*	*Q*	*T*
Temporal	T	0	+0	+0	+1
Spatial	S	1	+1	+0	+0
SNR (CGS)	C	1	+1	+0	+0
SNR (MGS)	M	$\widetilde{1}$	+1	$+1^*$	+0

3. ANALYSIS OF R-D OPTIMIZED EXTRACTION PATHS

This section provides a detailed, informative study on the R-D optimized extraction paths with respect to (1) device types, (2) video content, (3) distortion measure, and (4) spatiotemporal interpolation. In addition to MSE, the perceptual quality model [1] is also adopted in our experiments to better characterize the visual effects of time-varying spatiotemporal quality. Moreover, to simulate the actual use of SVC, extracted videos are interpolated to the spatiotemporal resolutions of target devices using standard-compliant spatial filtering followed by frame replication or motion field estimation likes B_Direct_16x16. Table 1 details the encoder configurations, where the arrows specify the inter-layer dependencies among CGS layers and the underlines mark the QCIF reference layers for spatial scalability.

In Figure 1, the R-D optimized extraction paths of different video sequences and device types are compared. It is obvious that *the optimal extraction paths are content dependent.* For instance, the SNR/spatial layers are extracted with higher priority in Akiyo sequence, while the temporal layers are more preferable in high-motion sequences such as Football and Mobile. The differences are caused by the very different visual characteristics in these sequences. In addition, it is interesting to note that *the optimal extraction paths of different viewing devices reveal regular, predictable patterns.* The paths of less capable devices can be deduced by contracting the ones of more capable devices in both spatial and temporal dimensions, thereby eliminating the need to specify a separate extraction path for each device type. The result is exploited in our path inference rule to provide a succinct denotation of extraction paths.

In Figure 2 (a), we compare the optimal extraction paths based on MSE and MOS, which is produced with the VQM software [1]. From the figure, the MSE-based extraction tends to have a bias in favour of temporal quality; that is, the distortion due to frame skipping is considered more significant. A similar tendency is also observed in other test sequences. The temporal effect, however, may not achieve the same degree of impairment in perceptual quality, as indicated by the extraction path with MOS. The result once again

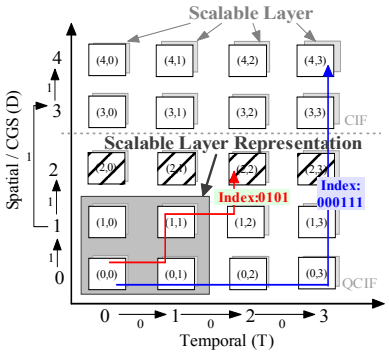

Fig. 3. An example of extraction paths and their index notations.

Table 3. Extraction Steps along the Path 0101

Extraction Steps	*Target Layers*	*Extracted Layers*
<Base>	(0,0)	–
T(0)	(0,1)	(0,1)
C(1)	(1,1)	(1,0),(1,1)
T(0)	(1,2)	(0,2),(1,2)
M($\widetilde{1}$)	$(2,2)^*$	$(2,0)^*$,$(2,1)^*$,$(2,2)^*$

emphasizes the limitation of MSE in reflecting the true perceptual quality, especially when multidimensional adaptation is in use.

In Figure 2 (b), we further illustrate the impacts of temporal interpolation on the optimal extraction paths. In comparison to frame replication, the better quality of B_Direct_16x16 allows the extraction to preferentially improve spatial quality. The results suggest that the temporal interpolation of target devices will also crucially affect the choices of optimal extraction paths. It is, however, reasonable to assume the use of frame replication as most of viewing devices simply display the last decoded picture when frame skipping occurs. Of equal importance is the spatial interpolation. However, as discovered in our experiments, the influence of temporal interpolation is more apparent.

4. DENOTATION AND INFERENCE OF OPTIMIZED EXTRACTION PATHS

Our multiple adaptation scheme [§2] implies the presence of *four* different types of bitstream extraction. At every refinement step, the decoder may take in additional scalable layers in one of *temporal*, *spatial* or *SNR (CGS/MGS) dimensions* to enhance the quality of playback video. We denoted these four different steps as T/S/C/M respectively. Their notations and effects in terms on the (D, Q, T) *identifiers* are listed in Table 2. Index values of 0, 1 and $\widetilde{1}$ can also be affixed to these steps so that we can assign distinct values to different extraction paths by simply concatenate the indices of consecutive refinement steps.

As an example, Figure 3 shows three different extraction paths for traversing a CIF30 bitstream and their index notations. Individual steps taken along the path (0101) are listed in Table 3. This example clearly shows that the decoder may acquire multiple layer representations in a single extraction step in order to obtain a decode NAL set.

It is obvious that the *optimal extraction paths* for various viewing devices may differ from one another. Hence, we may need to embed the denotation of multiple extraction paths into a single SVC

bitstream. Fortunately, our experiments [Fig.1] showed that we can simply *contract* the optimal extraction path for a more capable device to suit a less capable one by simply deleting the steps that acquire layer representations lying beyond the playback capability of that device. This *contraction rule* only works if the bitstream is produced using a *well-adapted encoding setting* [§2] and an objective quality metric (e.g. PSNR or MSE) is used for its R-D optimization. In the cases that the perception-based quality metrics are used, the contraction rule may yield a reasonable but non-optimal extraction path. In those case, we recommend the use of multiple extraction path denotations with one for each device type of a different spatial resolution as the extraction paths tend to vary most significantly along the spatial dimension.

5. COMPARISONS AND CONCLUDING REMARKS

Before drawing our conclusions, we would like to compare our adaptation scheme with the Quality Layer (QL) and Basic extractions implemented in JSVM 9. In our experiments, we examine two types of scalability: (1) QCIF SNR and (2) QCIF/CIF combined scalability. Two quality enhancements from the base quality are encoded for QCIF SNR scalability, while each spatial resolution in QCIF/CIF combined scalability is encoded with a base quality and one quality enhancement. Both experiments use the MGS vector mode {3, 3, 4, 6} without key pictures. In addition, each layer is simply predicted from the previous layer and the Quality Layers are assigned independently across spatial layers, i.e., the QCIF substreams must be entirely extracted prior to the extraction of the CIF layers.

From Figure 4, the proposed scheme is far superior to the other two approaches in Akiyo sequence while showing comparable performance in Foreman sequence. The reasons are twofold. Firstly, our scheme allows optimal extraction paths to preferentially improve spatial quality without extracting the entire base layer. However, both the QL and Basic extractions must initially extract the base layer at full frame rate. Secondly, our extraction paths are derived based on the *real* R-D costs of scalable layers. Contrarily, the Quality Layers are computed by estimating the R-D information in open- and/or closed-loop decoding.

Summarizing, in this paper, we propose an *optimal SVC bitstream extraction scheme* that can produce scalable layer representations for different viewing devices scattered over a multicasting network with diverse link bandwidth. Our scheme can determine *optimal extraction orders/paths* that are implementable at different multicasting nodes by performing successive steps of *multiple adaptation*. In addition to this basic scheme, we also develop an unambiguous denotation for the optimal extraction paths, and an algorithm for deducing the optimal extraction paths for less capable devices through path truncation. Extensive SVC encoding and adaptation experiments have been performed using JSVM 9 and both objective quality metrics (PSNR and MSE) as well as subjective metric (VQM). The experiment results showed that our scheme works well if *convexity* of R-D performance is maintained throughout the SVC bitstream.

6. REFERENCES

[1] "Institute for Telecommunication Sciences (ITS) Video Quality Research," *http://www.its.bldrdoc.gov/n3/video/index.php*.

[2] "Part I Final Draft International Standard (ISO/IEC FDIS 15444-1)," *ISO/IEC JTC1/SC29/WG1 N1855*, 2000.

[3] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, "Optimized Rate-Distortion Extraction With Quality Layers in the Scalable Extension of H.264/AVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, pp. 1186 – 1193, September 2007.

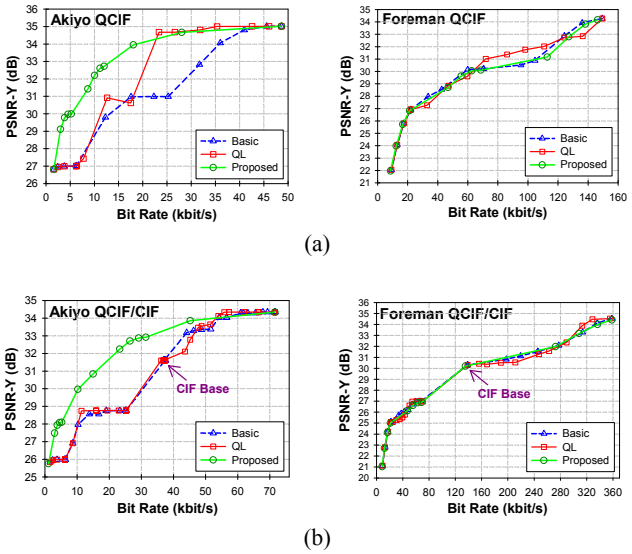

Fig. 4. R-D preformance comparison of the proposed scheme with the Quality Layer and Basic extractions in JSVM 9: (a) QCIF SNR Scalability, (b) QCIF/CIF Combined Scalability.

[4] M. A. J. Barzilay, J. R. Taal, and R. L. Lagendijk, "Subjective Quality Analysis of Bit Rate Exchange Between Temporal and SNR Scalability in the MPEG4 SVC Extension," *Proceedings, IEEE International Conference on Image Processing (ICIP)*, September 2007.

[5] R. Feghali, D. Wang, F. Speranza, and A. Vincent, "Quality Metric for Video Sequences With Temporal Scalability," *Proceedings, IEEE International Conference on Image Processing (ICIP)*, vol. 3, September 2005.

[6] Y. S. Kim, Y. J. Jung, T. C. Thang, and Y. M. Ro, "Bit-stream Extraction to Maximize Perceptual Quality Using Quality Information Table in SVC," *Proceedings, SPIE Conference on Visual Communications and Image Processing*, vol. 6077, January 2006.

[7] J. Lim, M. Kim, S. Hahm, K. Lee, and K. Park, "An Optimization-theoretic Approach to Optimal Extraction of SVC Bitstreams," *ISO/IEC JTCI/SC29/WG11 and ITU-T SG16 Q.6, JVT-U081*, October 2006.

[8] Z. Lua, W. Lina, B. C. Sengb, S. Katob, S. Yaoa, E. Onga, and X. K. Yanga, "Measuring the Negative Impact of Frame Dropping on Perceptual Visual Quality," *SPIE-IS&T Electronic Imaging*, vol. 5666, pp. 16 – 20, January 2005.

[9] W. H. Peng, L.-S. Huang, J. K. Zao, J.-S. Lu, T.-W. Wang, H.-T. Huang, and L.-C. Kuo, "Rate-Distortion Optimized SVC Bitstream Extraction for Heterogeneous Devices : A Preliminary Investigation," *IEEE Workshop on Scalable Video Coding and Transport*, 2007.

[10] D. Wang, F. Speranza, A. Vincent, T. Martin, and P. Blanchfield, "Towards Optimal Rate Control: A Study of the Impact of Spatial Resolution, Frame Rate, and Quantization on Subjective Video Quality and Bit Rate," *SPIE Proceedings, Visual Communications and Image Processing (VCIP)*, vol. 5150, pp. 198 – 209, July 2003.

[11] Y.-K. Wang, M. Hannuksela, S. Pateux, A. Eleftheriadis, and S. Wenger, "System and Transport Interface of SVC," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1149 – 1163, September 2007.

[12] T. Wiegand, G. Sllivan, J. Reichel, H. Schwarz, and M. Wien, "Joint Draft ITU-T Rec. H.264 — ISO/IEC 14496-10/Amd.3 Scalable Video Coding," *ISO/IEC JTCI/SC29/WG11 and ITU-T SG16 Q.6, JVT-X201*, July 2007.