Rate-Distortion Optimized SVC Bitstream Extraction for Heterogeneous Devices : A Preliminary Investigation

Wen-Hsiao Peng†, Lin-Shung Huang†, John K. Zao†, Jiun-Shien Lu†, Tse-Wei Wang†,
Hsueh-Ting Huang†, and Lun-Chia Kuo‡
†Department of Computer Science, National Chiao-Tung University, HsinChu, Taiwan
‡Industrial Technology Research Institute, HsinChu, Taiwan

Abstract— **The emerging SVC standard enables partial extraction and decoding of video bitstreams, and thus allows various viewing devices to adapt their video reception and playback according to devices capability and network performance. This desirable feature, however, comes with a caveat: the parts of a bitstream needed for providing good quality playback at different devices may differ significantly depending on the visual characteristics of video programs, the quantization and dependency settings of SVC encoders as well as the display formats of viewing devices. In this paper, we present the results of our preliminary investigation on the intricate relations among these factors. We discovered a set of constraints on the *setting of quantization parameters* and the *choices of reference layers* for inter-layer dependencies that ensures *good rate-distortion trade-offs* and *regular bitstream extraction paths*. We called these constraints, the *adaptation rules* for inter-layer encoding, and their resulting outputs, the *well-adapted layers/bitstreams*. We further discovered that bitstream extraction by different viewing devices follow predictable paths if the SVC bitstream is *well-adapted* and the playback process extracts a complete set of interdependent network abstraction layer (NAL) units in every refinement step. These regular extraction paths enabled us to develop an efficient bitstream extraction algorithm based on local optimization of rate-distortion ratios. The experimental results using standard testing sequences confirmed our findings.**

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

I. Introduction

PRODUCTION of *scalable bitstreams* that can be played back by a garden variety of viewing devices such as cellular telephones, laptop/desktop computers and high-definition televisions has been a long pursued goal of video compression technology. The emerging scalable video coding extension of H.264/AVC standard (referred to as SVC) *[1][2][3]* promises to achieve that goal by employing *adaptive inter-layer prediction* along with *hierarchical multi-reference temporal prediction*. By encoding a video sequence into an interdependent set of *network abstraction layer (NAL) units*, SVC allows different viewing devices to extract and decode subsets of NAL units according to their display formats, processing power and/or data throughput. SVC also enables video multicasting sessions to deliver only the necessary NAL units to different viewing devices by performing discretionary data transport and ensure graceful degradation of viewing quality over lossy communication by applying unequal error/erasure protection onto different subsets of NAL units. However, in order to implement these scalable features, the SVC encoders, the transport network nodes and the SVC decoders must all have the full knowledge of dependence relations and playback importance of the NAL units. As we discovered in our investigation, obtaining such knowledge is never a trivial task because the interdependency among the NAL units, the setting of *quantization parameter (Qp)* during their encoding process and the order of extraction during their decoding process all affect the *coding efficiency* and the *rate-distortion (R-D) performance* of an SVC bitstream. Consequently, all these factors must be fine tuned during the encoding/decoding of individual video program and even different parts of a single program. With this realization, we decided to investigate the relationship among video content characteristics, encoder parameter settings, viewing device capability and decoder extraction paths. This paper presents the preliminary findings of our investigation.

Our investigation aimed at answering two questions: (1) how does the tuning of *Qp* coupled with the changing of inter-layer dependencies affect the R-D performance at the decoding of spatial and SNR layers? (2) how does the optimal extraction paths of viewing devices differ from one another owing to their differences in display formats and processing power? Our attempt to answer each question led to a separate study. Our studies distinguished from similar previous studies *[4][5][6]* in two significant ways: (1) we examined the effects of different encoder settings and inter-layer dependencies with respect to a heterogeneous community with different viewing devices (rather than a single device or a homogeneous community) and (2) we compared the R-D performances of intermediate decoded videos by interpolating them all to the highest spatiotemporal resolution. By doing so, we created a unified framework for comparing the R-D performance of different intermediate representations and the extraction paths of various devices.

The remaining of this paper is divided into five parts. Section II contains a review of SVC inter-layer dependency structure and spatial-temporal-fidelity prediction mechanisms. This review tries to introduce the concepts and terms we used in our studies. Sections III and IV offer our answers to the two questions we raised. In these sections, we prescribe several rules that govern the choices of quantization parameters, dependency relations and extraction paths. We then present the results of our encoding/decoding experiments in Section V to support our prescription. This paper ends with a summary of our observations and a list of future works in the conclusion.

II. SVC Dependency and Prediction Structure

In order to support spatial, temporal and fidelity (SNR) scalability, an SVC bitstream is decomposed into an array of interdependent layer representations. In the spatial/SNR dimension, layer representations are assembled into *access units (AU)*. Within an AU, every layer representation may depend on another layer representation as its *base layer* through *inter-layer motion, residual, and textural predictions*. In principle, the value of *Qp* and the choice of base layer of each layer representation may differ among the access units. In practice, as in our studies, they remain constant in a coded video sequence. In the temporal dimension, the AUs are aligned in *groups of pictures (GOPs)*. Within each GOP, decoded pictures are related through a hierarchy of temporal dependence relations. Each picture can choose multiple decoded pictures as its *temporal references,* and the dependence relations may be dyadic or irregular. One such mesh of dependencies is depicted in Fig. 1.

The spatial/SNR and temporal dependencies in an SVC bitstream impose a *partial order* among its layer representations. Decoding of

This work was supported in part by a grant (NSC-96-2628-E-009-015-MY3) from National Science Council of Taiwan.

0-7695-3084-2/07 $25.00 © 2007 IEEE
DOI 10.1109/ISM.Workshops.2007.74

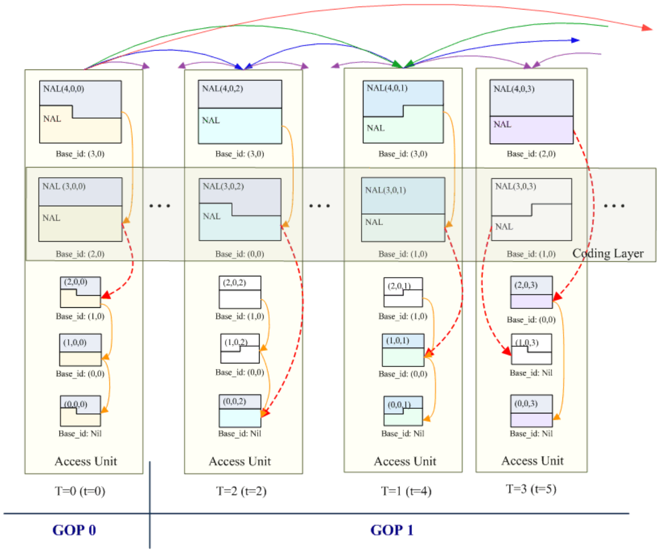


Fig. 1. SVC Dependency Structure

an 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 dependencies originate from the target representation. If these dependencies are depicted by a graph then all these NAL units should 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. 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 lacks error concealment capability) then its *extraction path* must traverse *an ordered sequence of convex subsets*[1] starting from the one containing the base layer and ending with the one containing its target representation. Among these extraction paths, one is regarded *optimal* for a target device if and only if the corresponding series of convex sets provides the best R-D performance when they are decoded on that device. The setting of proper inter-layer dependence relations and the search for optimal extraction paths for different viewing devices are the subjects of the next two sections.

III. Proper Setting of Inter-Layer Dependencies

Our first study aimed at investigating the combined effects of quantization parameter (*Qp*) settings and inter-layer (spatial/SNR) dependencies in determining the coding efficiency and the R-D performance[2] of an SVC bitstream. To simplify our preliminary study, we assigned the same *Qp* value to all NAL units within a coding layer and specified the same dependency setting for all the access units within a GOP. From our experiment results in Section V-A, we deduced the following two rules for setting *Qp* values and choosing reference layers.

Proposition 1: Monotonic Reduction of Distortion in Successive Refinement. For a set of SVC NAL units, a *proper dependency setting* for the set guarantees that every ordered sequence of its convex subsets exhibit a monotonic decrease of distortion values (in mean squared error) when these subsets are decoded in order.

Proposition 2: Convexity of Rate-Distortion Curves. For a set of SVC NAL units, a *well-adapted dependency setting* for the set guarantees that every ordered sequence of its convex subsets exhibit a monotonic decrease of distortion values as well as a monotonic decrease of the rates of distortion reduction.

The first rule implies that each plausible extraction paths embedded in an SVC bitstream should enable the decoder to improve picture quality at every refinement step. Consequently, a viewing device can produce better pictures by extracting more dependent NAL units from the bitstream and running them through the decoder. Any convex set that violates this rule contains wasted data bits because some of its NAL units do not contribute to the improvement of picture quality (or the reduction of picture distortion).

The second rule is even stronger. It requires that the *rate of distortion reduction* or the *slope of R-D curve* decreases monotonically as the decoder takes the convex subsets in successive refinement steps. The subtle difference between these two rules can be explained using Fig. 2. In the figure, each layer (from B to E) in Dependency Setting #1 simply depends on its previous layer; hence, the reconstruction of layer E requires the decoding of all its dependent layers from A to D. However, because the rate of R-D improvement produced by D is not as good as the one produced by E, Setting #1 is merely a proper setting but not a well-adapted setting. In contrast, Dependency Setting

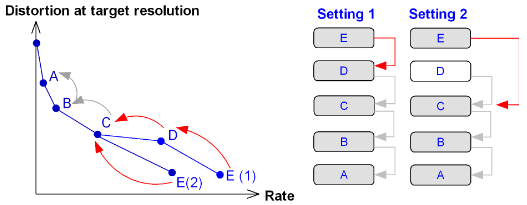


Fig. 2. Comparisons of R-D performances using different dependency settings. A, B, C, D, and E stand for the five coding layers of different quality levels with E being the target layer to be reconstructed.

[1] An *ordered sequence of convex subsets* is an ordered sequence of convex sets with each element (except the last one) being a subset of the next element.

[2] For a multi-layer video code with refinement support, its rate-distortion (R-D) performance is measured by the magnitude of rate-distortion ratios at each refinement step.

#2, which links C directly to E by skipping D, is a well-adapted setting. As explained in the next section, a well-adapted dependency setting enables the decoder to identify an optimal extraction path by merely examining the local R-D improvement provided by each layer.

IV. Optimization of Bitstream Extraction

Starting from a scalable bitstream with maximum spatial, temporal, and SNR resolutions, SVC allows selected layers to be extracted and decoded according to viewing device capability and network performance. Thus, an optimal bitstream extraction algorithm should be devised for a target device to search for an extraction path that minimizes the distortions of decoded pictures over a range of bit rates while maximizing the efficiency of the extracted layers, which is equivalent to enlarging the convexity of R-D curve.

As mentioned in Section II, an SVC decoder should take an *ordered sequence of convex sets* that contain dependent NAL units during its successive refinement steps. Each such sequence corresponds to an *extraction path* within the SVC bitstream. In search for an optimal extraction path for a target device, we compared two (*local* vs. *global*) search strategies in our study. Starting from the convex subset with only the base layer, the local search strategy builds the extraction path by adopting a steepest-descent algorithm, which always includes the convex subset that offers the highest incremental R-D improvement as the next point on the path. In contrast, the global search strategy performs an exhaustive search on all possible extraction paths. The optimal path is the one that maximizes the convexity of R-D curve.

Section V-B describes the experiments we performed on optimized bitstream extraction. Our experiment results suggested the following rule.

Proposition 3: Localized Search for Optimal Extraction Path. The optimal extraction path for a target viewing device can be found using *local steepest-descent search strategy* if the SVC bitstream adopts a well-adapted dependency setting.

V. Experiments and Analyses

In this section, we present the experiments showing the foundations of the rules prescribed in Sections III and IV. In the first part, the combined effects of *Qp* and dependency settings in determining the R-D performance and coding efficiency of an SVC bitstream are studied. In the second part, the bitstream extractions based on local and global optimizations are explored for different video contents, device types, optimization schemes, and temporal interpolation algorithms.

A. Experiments with Inter-layer Dependency

1) Proper Settings of Quantization Parameter: The first experiment is conducted to investigate the proper settings of *Qp* for well-formed R-D performance. Without loss of generality, two SVC layers are encoded by firstly fixing the enhancement-layer *Qp* and then sweeping a range of the base-layer *Qp* from 0 to 51. Both CGS and spatial scalability are tested with their R-D performances being shown in Fig. 3 (a) and (b), respectively.

From Fig. 3, the R-D performance of the enhancement layer is dependent on the quality of the base layer. Normally, the enhancement layer is expected to provide better quality than the base layer. However, in Fig. 3 (a), it is interesting to note that improper *Qp* settings in CGS may cause the base layer to have better quality. Such ill-formed R-D performance is resulted by encoding the base layer with higher fidelity (using smaller *Qp* values). Similar effects are also found in spatial scalability as the interpolated base layer reveals superior quality than the enhancement layer. Unlike CGS, the spatial enhancement layer shows degraded R-D performance rather than an ill-formed R-D curve.

In view of the experimental results, we see that a coding layer *shall* be predicted from a dependent layer with *worse* quality so as to achieve well-formed R-D performance. This implies that proper *Qp* settings must ensure a monotonic decrease of distortion values when the SVC layers are decoded in order.

2) Proper Settings of Inter-layer Dependency: Having devised a criterion for assigning *Qp*, the dependency settings[3] for well-formed R-D performance follow almost immediately. In this experiment, we shall see that well-adapted dependency settings are determined by both video contents and *Qp* assignments. For the experiments, the service requirements of SVC bitstreams are firstly defined in Fig. 4, where the combined scalability is required to provide two spatial resolutions (including QCIF and CIF) with each having three CGS layers (QCIF: A0, A1, A2; CIF: B0, B1, B2)[4] and four temporal levels (3.75Hz∼30Hz). To show how the QCIF videos are perceived on CIF devices, the corresponding PSNR values with spatial interpolation are shown as A0′, A1′, and A2′.

Fig. 5 compares the R-D performances of the following dependency settings with layer B2 being the target representation. In the first three settings, we simply have the CIF layer B0 be predicted from different QCIF layers while maintaining a fixed and straightforward dependency relation between CGS layers. Setting #4 is specifically designed for the Foreman sequence, where layer B0 is skipped for coding.

- Setting #1: (A0←A1←A2), (A2←B0←B1←B2).
- Setting #2: (A0←A1←A2), (A1←B0←B1←B2).
- Setting #3: (A0←A1←A2), (A0←B0←B1←B2).
- Setting #4: (A0←A1←A2), (A1←B1←B2).

As expected, the R-D performance for reconstructing the target layer changes significantly according to dependency settings and video contents. For the Mobile sequence, in which the spatial interpolation works inefficiently, the QCIF layers A1, A2 reveal extremely poor R-D performance when interpolated and displayed on CIF devices. Thus, the CIF layer B0 should be prevented from being predicted by the QCIF layers A1 and A2, which explains the superior R-D performance of Setting #3 in Fig. 5 (a).

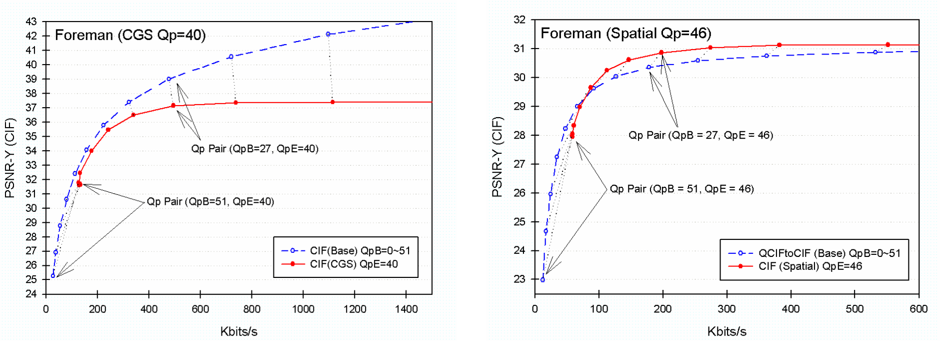
Contrarily, inspection of Fig. 4 (b) shows that the interpolated QCIF layers A1, A2 of the Foreman sequence offer better quality than the CIF layer B0. The fact justifies the poor R-D performances of Settings #1 and #2 in Fig. 5 (b), where the layer B0 is improperly predicted from the QCIF layers of better quality. Moreover, in contrast to Setting #3, the better R-D performance of Setting #4 proves the improper *Qp* assignment for the layer B0 since interpolating the QCIF layer A1 achieves the same or even better quality than the CIF layer B0.

In conclusion, the principle motivation for using well-adapted dependency settings is to provide better R-D performance. This can be achieved by ensuring the rate of distortion reduction decreases monotonically when the SVC layers are decoded incrementally. As illustrated by the examples above, such well-adapted settings vary with video contents and *Qp* assignments, and a feasible mechanism for reference layer selections is to follow the concave contour formed by the PSNR curves of different spatial layers.

3) Comparisons of Coding Efficiency: In the discussions above we have concentrated on the R-D performance of an SVC bitstream. However, in many applications of SVC, it is equally important to consider the coding efficiency. Fig. 6 compares the coding efficiency

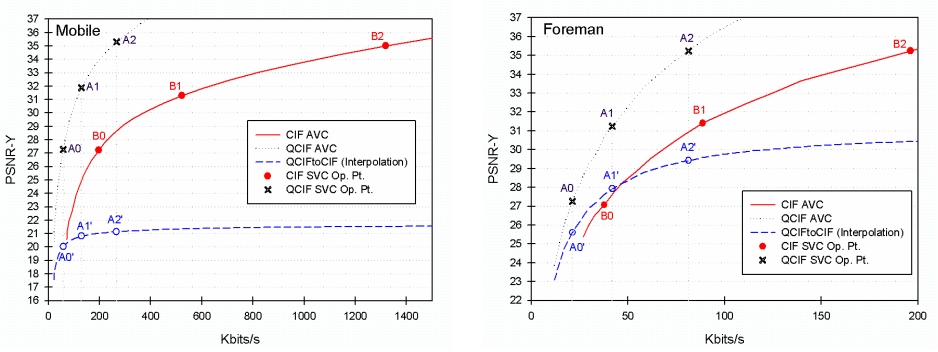
[3] By our definitions, proper dependency settings should produce concave rate-distortion functions when the distortion is expressed in PSNR.

[4] The PSNR values of the 3 CGS layers are uniformly distributed from 27dB to 35dB. We use the encoding results of H.264/AVC to obtain the exact *Qp* value for each layer.



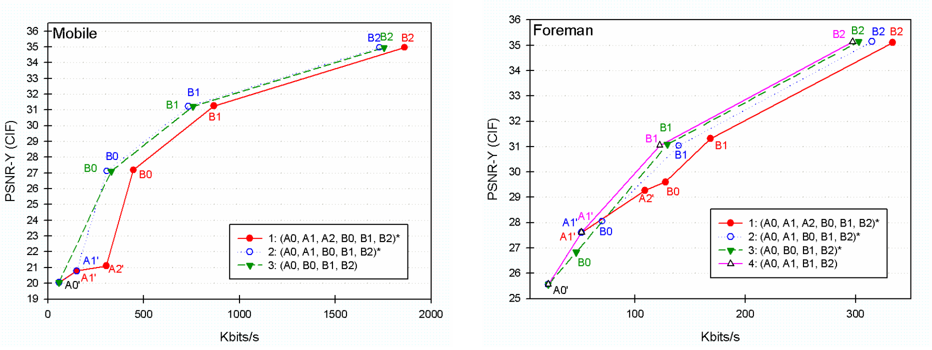
(a) (b)

Fig. 3. Comparisons of R-D performances using different Qp combinations with the configurations of (a) coarse grain scalability (CGS) and (b) spatial scalability.



(a) (b)

Fig. 4. Examples of service requirements of the SVC bitstream for the combined scalability: (a) Mobile and (b) Foreman.



(a) (b)

Fig. 5. Comparisons of R-D performances of different dependency settings with the layer B2 being the target layer: (a) Mobile and (b) Foreman.

of different dependency settings in terms of total bit rates. In contrast to Setting #1, in which a layer representation is always predicted from a dependent layer with the best quality, it is of no surprise that well-adapted settings introduces 3∼11% extra bit rates, which is a consequence of minimizing the number of dependent layers for a target representation. However, it should be noted that the loss in coding efficiency is traded for better R-D performance.

B. Experiments with Bitstream Extraction

Based upon well-adapted layers and dependency settings, in this section we further investigate optimal bitstream extraction paths for different video contents, device types, optimization schemes, and temporal interpolation algorithms. Video sequences of different visual characteristics are tested using similar configurations in Section V-A.2 and optimal extraction paths are explored for the three target devices having display formats of CIF@30Hz, CIF@15Hz, and QCIF@30Hz, respectively. In order to compare the R-D performances along different extraction paths, we interpolated the results to the highest spatiotemporal resolution available on all the target devices. By doing so, we created a unified framework for comparing the R-D performance of different intermediate representations and the extraction paths of various devices. Particularly, the spatiotemporal interpolation

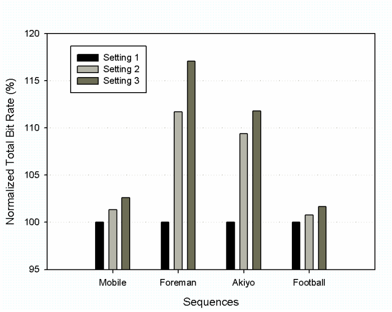


Fig. 6. Comparisons of coding efficiency with different dependency settings. (Sequence, Proper Setting): (Mobile, Setting #3), (Foreman, Setting #2), (Akiyo, Setting #2), and (Football, Setting #3).

is performed in two sequential steps. Firstly, the standard-compliant filter provided in JSVM [7] is used for spatial interpolation. Secondly, based on the interpolated pictures, the temporal interpolation is conducted by predicting the motion vectors of missing pictures in a way similar to the temporal B_Direct_16x16 mode.

1) Influence of Video Content: In Fig. 7, the globally optimized extraction paths of different video sequences and device types are compared. Starting from the same base layer, it can be seen that optimal extraction paths proceed differently towards the top-left corner according to the spatiotemporal characteristics of video contents. The layers with improved SNR and/or spatial quality are always extracted with higher priority in the static Akiyo sequence, while on the other hand, the higher temporal layers are more preferable in the fast-motion or highly-textured sequences such as Football and Mobile. The phenomena can be explained by the better efficiency of temporal interpolation in static sequences. Interestingly, the Foreman sequence represents an in-between scenario, in which the optimal path for CIF@30Hz firstly improves the temporal resolution and then turns to enhance the SNR quality before reaching the full frame rate.

2) Influence of Device Type: In Fig. 7, it is interesting to note that optimal extraction paths of different viewing devices actually reveal regular predictable patterns for the same video content. The optimal paths for devices having lower spatiotemporal resolutions are mostly found to be the projections and/or the truncated versions of the ones for higher resolutions. For instance, one can project the optimal paths for CIF@30Hz onto a three-dimensional space with a reduced temporal resolution in order to predict the ones for CIF@15Hz. Similarly, the paths for QCIF@30Hz can be deduced from the projected ones of CIF@30Hz on the SNR-temporal plane. However, an exception is observed in Fig. 7 (b), in which the optimal path for QCIF@30Hz is different from the projected one of CIF@30Hz. The discrepancy is caused by the less efficient temporal interpolation that uses a fixed block size regardless of the spatial resolution. Consistent results are expected by incorporating a temporal interpolation algorithm with adaptive block size.

3) Influence of Optimization Scheme: As in most optimization problems, local optimization generally results in suboptimal solutions. However, from the experimental results, the same extraction paths are found by both global and local optimizations with well-adapted dependency settings. Exceptions occur only when improper dependency settings are employed. An example is presented in Fig. 8, where the local optimization turns to enhance the temporal quality at the branch point owing to the ill-formed R-D performance of the CGS layer A2. As it can be seen from the example, without the knowledge of global R-D performance, local optimization could be easily and mistakenly led by ill-formed R-D performance. On the other hand, by ensuring proper dependency settings, the consistent extraction paths enable us to develop an efficient bitstream extraction algorithm based on local optimization of rate-distortion ratios.

4) Influence of Interpolation Algorithm: The bitstream extraction paths are also highly dependent on the spatiotemporal interpolation algorithms adopted by target devices. Due to the complicated relations, including both translation and dilation, between successive pictures, the changes on extraction paths are more apparent with different temporal interpolation algorithms. As we discovered in our experiments, using straightforward frame replication normally produces only two types of extraction paths. One always extracts the SNR/spatial layers with higher priority, and the other one always preferentially extracts the temporal layers. By implementing better interpolation schemes, more varieties can be observed. Such results suggest that the spatiotemporal interpolation schemes on target devices should also be considered in search for an optimized extraction path.

VI. Conclusion

In this paper, we proposed two adaptation rules for setting the quantization parameters and dependence relations among SVC spatial and SNR layers. The first rules prohibit a coding layer from choosing a reference layer with higher PSNR value[5]. The second rule requires that the layer representations fed into SVC decoders provide diminishing distortion improvement in successive refinement steps. We also introduced a localized steepest-descent algorithm to search for the optimal extraction path that ensures best playback performance for different viewing devices if the underlying dependence relations are well adapted.

We plan to extend our preliminary investigation in several directions: (1) to devise assignment schemes/rules for priority identifiers in SVC NAL headers so as to signal optimal extraction paths, (2) to include with SVC bitstreams supporting medium-grain scalability (MGS), and (3) to include experiments with error concealment techniques.

References

[1] T. Wiegand, G. Sllivan, J. Reichel, H. Schwarz, and M. Wien, "Joint Draft 9 of SVC Amendment," *ISO/IEC JTCI/SC29/WG11 and ITU-T SG16 Q.6, JVT-V201*, January 2007.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *Proceedings, IEEE International Conference on Image Processing (ICIP)*, October 2006.

[3] H. C. Huang, W. H. Peng, T. Chiang, and H. M. Hang, "Advances in the Scalable Amendment of H.264/SVC," *IEEE Communications Magazine*, vol. 45, pp. 68 – 76, 2007.

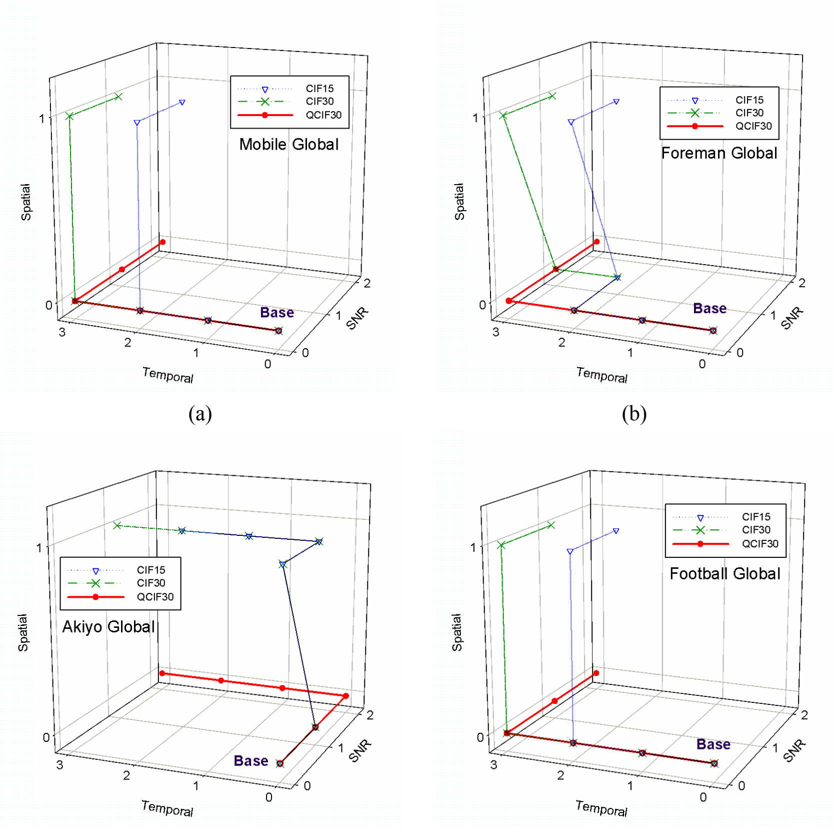
[4] J. Lim, M. Kim, S. Hahm, K. Lee, and K. Park, "An Optimization-theoretic Approach to Optimal Extraction of SVC Bitstreams," *ISO/IEC JTCI/SC29/WG11 and ITU-T SG16 Q.6, JVT-U081*, October 2006.

[5] 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.

[6] I. Amonou, N. Cammas, and S. Kervadec, "Optimized Rate-Distortion Extraction with Quality Layers in the H.264/SVC Scalable Video Compression Standard," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, JVT-U144*, October 2006.

[7] J. Reichel, H. Schwarz, and M. Wien, "Joint Scalable Video Model JSVM-8," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, JVT-U202*, October 2006.

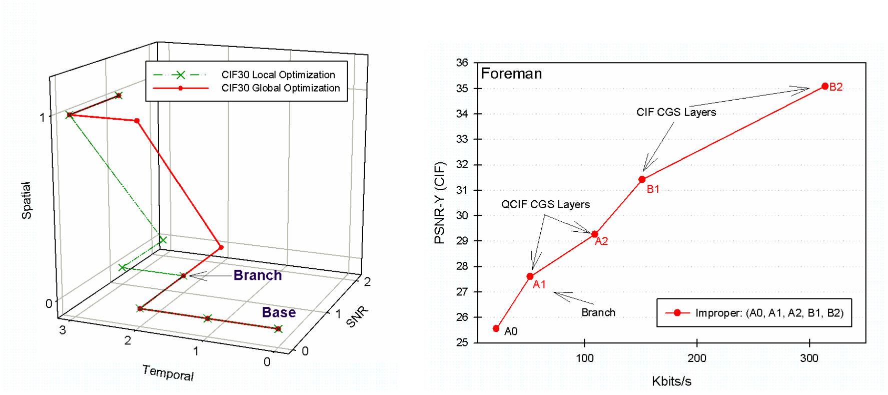
[5] The PSNR values of the coding layer and its reference layer should be measured after both layers are interpolated to the same spatiotemporal resolution.



(a) (b)

(c) (d)

Fig. 7. Comparisons of globally optimized bitstream extraction paths with proper dependency settings: (a) (Mobile, Setting #3), (b) (Foreman, Setting #4), (c) (Akiyo, Setting #2), and (d) (Football, Setting #3).



(a) (b)

Fig. 8. Bitstream extraction paths of Foreman sequence with improper dependency setting: (a) Comparisons of the extraction paths with local and global optimizations, (b) R-D performance of the improper dependency setting, (A0←A1←A2), (A2←B1←B2).