J. Vis. Commun. Image R. 19 (2008) 543–557

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

A rate-distortion optimization model for SVC inter-layer encoding and bitstream extraction

Wen-Hsiao Peng *, John K. Zao *, Hsueh-Ting Huang, Tse-Wei Wang, Lin-Shung Huang

Department of Computer Science, National Chiao Tung University 1001 Ta-Hsueh Road, Hsinchu 30010, Taiwan

ARTICLE INFO

Article history:
Received 6 December 2007
Accepted 1 August 2008
Available online 15 August 2008

Keywords:
Scalable Video Coding
Bitstream extraction
Inter-layer dependency
H.264/AVC
Rate-distortion optimization
Steepest-descent search
Multiple adaptation
Video streaming
Subjective quality
Quality metric

ABSTRACT

The Scalable Video Coding (SVC) standard enables viewing devices to adapt their video reception using bitstream extraction. Since SVC offers spatial, temporal, and quality combined scalability, extracting proper bitstreams for different viewing devices can be a non-trivial task, and naive choices usually produce poor playback quality. In this paper, we propose a two-prong approach to achieve rate-distortion (R-D) optimal extraction of SVC bitstreams. For SVC encoding, we developed a set of adaptation rules for setting the quantization parameters and the inter-layer dependencies among the SVC coding layers. A well-adapted SVC bitstream thus produced manifests good R-D trade-offs when its scalable layers are extracted along extraction paths consisting of successive refinement steps. For extracting R-D optimized bitstreams for different viewing devices, we formalized the notion of optimal and near-optimal extraction paths and devised computationally efficient strategies to search for the extraction paths. Experiment results demonstrated that our R-D optimized adaptation schemes and extraction strategies offer significant improvement in playback picture quality among heterogeneous viewing devices. Particularly, our adaptation rules promise R-D convexity along optimal extraction paths and permit the use of steepest-descent strategy to discover the optimal/near-optimal paths. This simple search strategy performs only half of the computation necessary for an exhaustive search.

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 scalable extension of H.264/AVC standard (referred hereafter as SVC) [11,15] achieved that goal by employing *multilayer coding* along with *adaptive inter-layer prediction* and *hierarchical temporal reference*. By encoding a video sequence into an inter-dependent set of scalable layers, SVC allows different viewing devices to extract and decode parts of the scalable layers according to their playback capability, processing power, and/or network connection quality. How to offer different devices with appropriate scalable layers through *bitsream adaptation and extraction* thus becomes an intriguing problem.

Currently, the Joint Scalable Video Model (JSVM) [10,14] recommends three different ways to perform bitstream extraction: (1) given a target bit rate, an extractor shall produce a scalable layer representation with a bit rate that is closest to but not greater than the target bit rate; (2) given a specific scalable layer, an extractor shall extract not only the target layer but all the reference layers needed for decoding the target layer; (3) given a specific frame rate, frame size, and bit rate of the extracted bitstream, the extractor shall extract a decodable representation that satisfies those constraints. While these extraction mechanisms provide flexibility for discretionary bitstream extraction, the current standard does not specify what the resultant bitstream should be if there exist several possible extraction sequences.

Several approaches have been proposed for finding optimal bitstream adaptation and extraction schemes that ensure the best playback quality on a viewing device while making the best use of available transport bandwidth. Amonou et al. [2] formulated the problem as a rate-distortion (R-D) optimization process and shuffled the quality increments in an R-D sense. The idea is similar to the creation of Quality Layers in JPEG 2000 [13] and has since been expanded and applied to the combined scalability for multi-dimensional adaptation. For instance, in [4] Kim et al. evaluated the perceptual preference for spatial and temporal quality over a range of bit rates, and recorded the spatiotemporal switching points using quality information tables. Lim et al. [6], on the other hand, defined a quality index to measure the perceptual quality and performed bitstream extraction by maximizing the quality index of the extracted substream subject to a bit rate constraint. All these approaches tackled the problem as a form of rate-distortion (R-D) optimization.

Although various R-D optimization techniques have been applied to bitstream extraction, the playback quality of extracted

* Corresponding authors.
E-mail addresses: pawn@mail.si2lab.org (W.-H. Peng), jkzao@cs.nctu.edu.tw (J.K. Zao).

1047-3203/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2008.08.002

bitstreams is still far from being satisfactory. This is because successful R-D optimization of bitstream extraction requires the cooperation of SVC encoding, extraction and decoding processes. Poor R-D performance can be caused by improper setting of quantization parameters and inter-layer dependencies as likely as the performance of bitstream extraction along a suboptimal path. In this paper, we propose a novel R-D optimization scheme that tackles the problem by mending both encoding and extraction processes. In the course, this paper also offers answers to the following questions:

(1) How can the tuning of quantization parameters coupled with the changing of inter-layer dependencies affect the R-D performance of SVC bitstreams?
(2) What are the criteria on SVC encoder/decoder settings that ensure the existence of optimal or near-optimal extraction paths for different viewing devices?
(3) How can the optimal extraction paths of different viewing devices be found using computationally efficient strategies when an SVC bitstream is to be extracted through successive refinements?

The main contributions of our works include

- Specification of adaptation rules for setting the quantization parameters and the inter-layer dependencies during SVC encoding,
- Formalization of the notion of optimal/near-optimal extraction paths according to the playback quality of extracted bitstreams on different viewing devices.
- Depiction of R-D optimized bitstream extraction using the novel graphical tool of R-D Trellis diagram.
- Development of efficient search strategies for optimal/near-optimal extraction paths.

Experiment results showed that our optimization scheme offer significant improvement in playback quality of extracted bitstreams. Our adaptation rules promise R-D convexity along the optimal extraction paths and enable the steepest-descent method to achieve comparable performance as the exhaustive search while reducing the computation complexity in half.

The remaining of this paper is organized as follow: Section 2 contains a review of SVC dependency structure and presents our R-D optimization model for bitstream extraction. Section 3 introduces our strategies for finding an optimal or near-optimal extraction path. Section 4 describes the necessary criteria of well-adapted SVC encoding that guarantee the existence of optimal/near-optimal extraction paths. Section 5 addresses the issues of establishing well-adapted inter-layer dependencies among the SVC layers. Section 6 provides a detailed analysis on optimal extraction paths and evaluates the performance of the steepest-descent method in search for optimal paths. The differences between our extraction scheme and previous works are also compared before the paper is concluded with a summary of our observations and a list of future works.

2. Models of SVC dependencies and bitstream extraction

Before discussing our optimal bitstream extraction process and well-adapted encoding scheme, we introduce in this section a *partial order model* for the *dependence relations* among the SVC layers and the notion of *optimal* and *near-optimal extraction* from an SVC bitstream. This section does not intend to serve as an introduction to SVC. The reader is referred to the specification and overview papers [3,11,15] for details of the SVC standard.

2.1. Partial order of SVC dependence relations

As shown in Fig. 1, an SVC bitstream is composed of an array of picture *slices*, each of which is encapsulated in a Network Abstraction Layer (NAL) *unit* ς. For the purpose of supporting *spatial, quality (SNR)* and *temporal* scalability, the NAL units are grouped into *scalable layers* $\aleph$, each of which is identified by three indices: *dependency identifier*[1] D, *quality identifier*[2] Q and *temporal identifier* T. In order to establish necessary references for motion and textural predictions, the scalable layers $\aleph(D,Q,T)$ or $\aleph(L,T)$ with $L = D\|Q$ are related to one another through two types of *dependence relations* Ψ_L and Ψ_T that span across two sets of scalable layers: the *spatial or quality layers* $\mathscr{L}(L) = \{\aleph(D,Q,T)|L = D\|Q; T = [0,T_{\max}]\}$ and the *temporal layers* $\mathfrak{F}(T) = \{\aleph(D,Q,T) \mid D = [0,D_{\max}]; Q = [0,Q_{\max}]\}$.

In the temporal dimension, the *temporal dependencies* $\Psi_T : \mathfrak{F}(T) \to \{\mathfrak{F}(T')|T' < T\}$ establish a hierarchical structure among the temporal layers. A temporal layer $\mathfrak{F}(T)$ may depend on *any* temporal layer $\mathfrak{F}(T')$ with $T' < T$ within the same *group of pictures* (GOP). If the temporal dependence relations are consisted of *dyadic inter-frame mappings* then a temporal layer shall depend on *all* the temporal layers that have lower temporal identifier values.

In the spatial or quality dimension, every coding layer may depend on exactly one other layer according to the *inter-layer dependence relations* $\Psi_L : \mathscr{L}(L) \to \mathscr{L}(L')$. The inter-layer dependencies may vary among different *access units*, which contain coded pictures to be played back at different time instances. In this paper, however, we assume Ψ_L remain invariant within each GOP. This assumption enabled us to use the SVC codec in the current release of JSVM software.

The dependence relations Ψ_L and Ψ_T supply every scalable layer $\aleph(L,T)$ with two sets of reference layers $\Psi_L(\aleph)$ and $\Psi_T(\aleph)$. These dependence relations impose a *strict partial order*[3] $\succ$ onto the scalable layers. According to the dependence relations, every NAL set that can be played back by a viewing device (referred hereafter as a *scalable layer representation*) must form a *partial-order lattice*. This is because in order for a scalable layer representation $S(\hat{L},\hat{T})$ to be decodable, it must include all the scalable layers $\aleph(L,T)$ on which the target layer $\aleph(\hat{L},\hat{T})$ are directly or indirectly dependent. Thus, $S(\hat{L},\hat{T})$ must be a *transitive closure* (*) of all dependence relations $\Psi = \Psi_L \cup \Psi_T$ originating from $\aleph(\hat{L},\hat{T})$:

$$S(\hat{L},\hat{T}) = \Psi^*(\aleph(\hat{L},\hat{T})) \subseteq \bigcup_{L,T=\bar{L},\bar{T}}^{\hat{L},\hat{T}} \aleph(L,T) \tag{1}$$

The transitive closure implies that every scalable layer $\aleph(\hat{L},\hat{T})$ in an SVC bitstream must depend ultimately on an independent scalable layer, known as the *base unit* $S(\bar{L},\bar{T}) = \aleph(\bar{L},\bar{T})$. This implies the partial-order lattice of $S(\hat{L},\hat{T})$ must have $\aleph(\bar{L},\bar{T})$ as the *bottom element* $\perp$ and $\aleph(\hat{L},\hat{T})$ as the *top element* $\top$. Furthermore, the set of scalable layer representations $\{S(L,T)\}_{L,T=\bar{L},\bar{T}}^{\hat{L},\hat{T}}$ that are related by the dependence relations Ψ_L and Ψ_T also forms a partial-order lattice with the bottom element $\perp = S(\bar{L},\bar{T})$ and the top element $\top = S(\hat{L},\hat{T})$ simply because the transitive closure of a *directed acyclic graph (DAG)* such as the one formed by the dependence relations is also a strict partial order and a DAG.

[1] The SVC dependency identifier D serves as an *index* of the *spatial* or *coarse-grain quality scalable (CGS)* layers.
[2] The SVC quality identifier Q serves as an index of the *medium-grain quality (MGS)* layers. In this paper, we do not include MGS layers into our SVC bitstreams. Hence, $Q = 0$ and $L = D$.
[3] Every *directed acyclic graph (DAG)* such as the one formed by the dependence relations is a *strict partial order*.

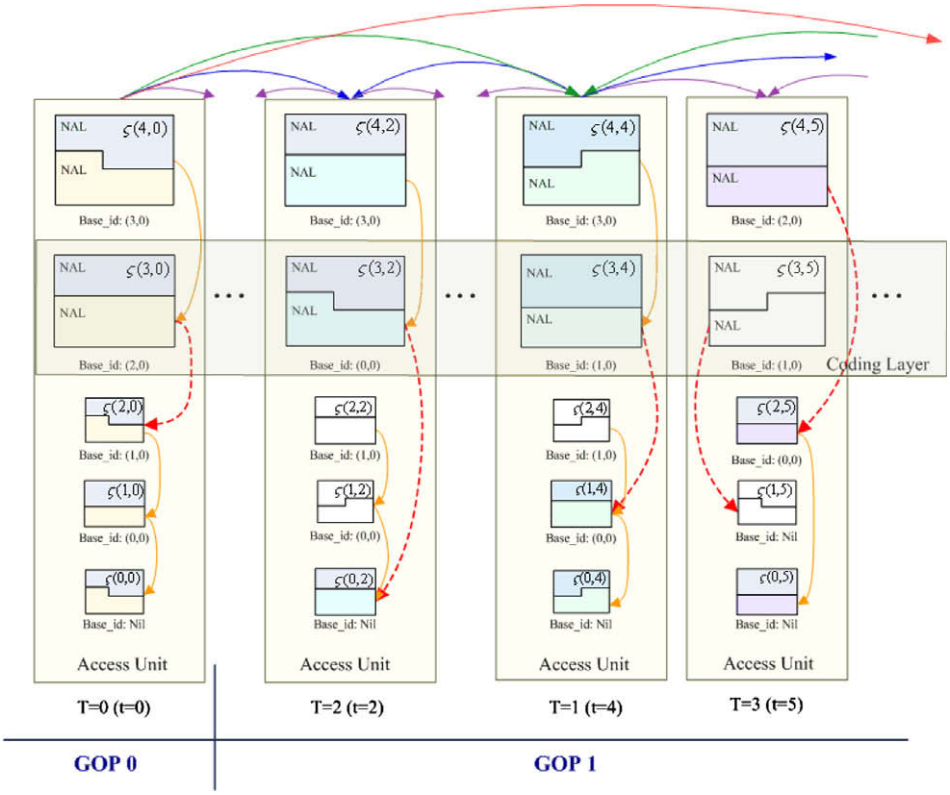

Fig. 1. SVC dependency structure.

2.2. Extraction paths through SVC bitstream

While playing back an SVC bitstream, a viewing device may choose to extract and decode various scalable layers (with possible use of error concealment) based on its display format, decoding capability and network throughput. A decoding sequence of these scalable layers arranged according to their dependence relations is known as an *extraction path* $\Pi_{\wp}$ for the viewing device. The subscript $\wp$ indicates a denotation of the extraction path.

2.2.1. Successive refinement

Beside of satisfying the dependence relations, one may want to fulfill some additional criteria while choosing the extraction paths for one or more viewing devices:

(1) One may want to feed a viewing device with scalable layer representations of lower bit rates when the network throughput deteriorates. Such an act of bit-rate adaptation enables a viewing device to support graceful degradation of playback quality.
(2) One may want to perform *successive extraction* en-route a multicasting tree. Significant reduction of transport bandwidth can be achieved by having an up-stream provider extract only the scalable layers needed by its down-stream subscribers. Careful selection of extraction paths for different down-stream subscribers may minimize the bandwidth consumption of a multicasting session [8].

People may also want to optimize the playback quality for multiple viewing devices connected to a service network while minimizing network bandwidth utilized to transport the SVC bitstream. This problem of rate-distortion optimized bitstream extraction for heterogeneous viewing devices, however, is not addressed in this paper.

The first criterion mentioned above requires that an extraction path $\Pi_{\wp}(\overline{L},\overline{T};\hat{L},\hat{T})$ must be a totally ordered sequence of the scalable layer representations $S(L,T)$ running from $S(\overline{L},\overline{T})$ to $S(\hat{L},\hat{T})$:

$$\Pi_{\wp}(\overline{L},\overline{T};\hat{L},\hat{T}) = [S(L_i,T_i)]_{i=0}^{N} \text{ with } (L_0,T_0) = (\overline{L},\overline{T}) \text{ and } (L_N,T_N) = (\hat{L},\hat{T}) \quad (2)$$

The denotation $\wp$ of the extraction path can be a sequence of (L,T) indices as shown in Section 3.3.3 or a sequence of dependency identifiers such as the ones specified in [8].

The criterion of *successive extraction* implies that every element along the extraction path must be the *supremum* $\wedge$ of all previous elements:

$$S(L_{i+1},T_{i+1}) = S(L_{i+1},T_{i+1}) \wedge S(L_i,T_i), \quad \forall i = [0,N-1] \quad (3)$$

Such an extraction path $\Pi_{\wp}$ is a *locus of suprema* traversing the lattice of $S(\hat{L},\hat{T})$ from its bottom element $\perp = S(\overline{L},\overline{T})$ to its top element $\top = S(\hat{L},\hat{T})$. Since every element along the path must have the previous element being its proper subset $S(L_{i+1},T_{i+1}) \supset S(L_i,T_i)$ (Fig. 16(a)), the extraction path can be produced by *successive refinement* of the SVC bitstream.

2.2.2. Incremental and cumulative rate-distortion performance

Since multiple loci of suprema may exist in a lattice, several extraction paths are available for traversing an SVC bitstream between the base unit and a target layer representation. These extraction paths are differentiated by their *rate-distortion (R-D) performance*, which measures the *effectiveness* that an extracted bitstream uses its data bits to enhance the playback picture quality. The R-D performance of an SVC bitstream can be quantified in two ways: (1) using a ratio between the increase in bit rate and the decrease in playback distortion at every refinement step

and (2) calculating the area underneath the R-D curve that spans the refinement steps. The two measures are defined below and used in Section 3.

The first (incremental) measure of R-D performance evaluates the *R-D improvement*[4] Γ incurred through successive refinement:[5]

$$\Gamma(L,T;L'',T'') \triangleq -\frac{d(L'',T'')-d(L,T)}{r(L'',T'')-r(L,T)} \tag{4}$$

where $d(L,T)$ is the distortion value and $r(L,T)$ is the total bit rate of $S(L,T)$. Note that the R-D improvement is path independent because each $S(L,T)$ has unique r,d values. The R-D improvements Γ_L,Γ_T along the L or T dimension deserve special attention. They are used in Section 3.2.1 to specify the *global R-D conditions* that ensure the feasibility of using a *steepest-descent* strategy to search for the optimal extraction path.

$$\Gamma_L(L,T;L'',T) \triangleq -\frac{d(L'',T)-d(L,T)}{r(L'',T)-r(L,T)} \tag{5a}$$

$$\Gamma_T(L,T;L,T'') \triangleq -\frac{d(L,T'')-d(L,T)}{r(L,T'')-r(L,T)} \tag{5b}$$

We further define the *local R-D improvements* γ of a single refinement step in either L or T dimensions as

$$\gamma_L(L,T) \triangleq -\frac{d(L',T)-d(L,T)}{r(L',T)-r(L,T)} \tag{6a}$$

$$\gamma_T(L,T) \triangleq -\frac{d(L,T')-d(L,T)}{r(L,T')-r(L,T)} \tag{6b}$$

where L' and T' denote the subsequent spatial/quality and temporal layers reached through a single refinement step. Note that these local R-D improvements are uniquely identified by their *reference identifiers* (L,T). We also define the R-D improvement Γ' of two successive refinement steps (one in each of the L and T dimensions):

$$\Gamma'(L,T) \triangleq \Gamma(L,T;L'',T'') = -\frac{d(L'',T'')-d(L,T)}{r(L'',T'')-r(L,T)} \tag{7}$$

This is the R-D improvement incurred during the traversal of a four-node trellis in the grid of $S(L,T)$ (Section 3.1). These traversals play a pivotal role in our proposed strategies to search for an optimal extraction path.

The second (cumulative) measure of R-D performance is the *underlying area* $\Omega_\wp$ of an R-D curve corresponding to an extraction path $\Pi_\wp(\bar{L},\bar{T};\hat{L},\hat{T})$:

$$\Omega_\wp(\bar{L},\bar{T};\hat{L},\hat{T}) \triangleq \frac{1}{2} \times \sum_{i=0}^{N} (d(L_{i+1},T_{i+1})+d(L_i,T_i))(r(L_{i+1},T_{i+1})-r(L_i,T_i)) \tag{8}$$

where $(L_0,T_0)=(\bar{L},\bar{T})$ denotes the base representation and $(L_N,T_N)=(\hat{L},\hat{T})$ denotes the target representation. Unlike the local R-D improvement, $\Omega_\wp$ depends on the chosen extraction path $\wp$. Also, rather than measuring the rate of R-D improvement in a single refinement step, $\Omega_\wp$ measures the *efficiency* of an SVC bitstream in using its data bits to enhance its playback quality through a series of refinement steps along the path $\wp$. Furthermore, $\Omega_\wp(\bar{L},\bar{T};\hat{L},\hat{T})/(r(\hat{L},\hat{T})-r(\bar{L},\bar{T}))$ can be interpreted as the *average playback quality* along the extraction path.

[4] The *negation* of the slope is used to ensure that a *positive* value reflects an improvement in playback picture quality.

[5] In the definitions of *R-D improvements* and the equations hereafter, we use indices L',T' to denote the scalable layer representations $S(L',T')$, that can be reached through a *single refinement step* from the reference representation $S(L,T)$, and use L'',T'' to denote $S(L'',T'')$ reachable through *multiple refinement steps* from $S(L,T)$.

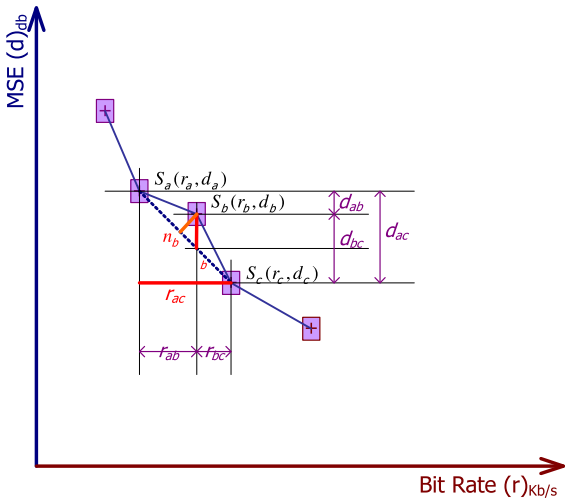

Fig. 2. Measuring components of the *deviation from convexity* of a *NAL cluster* along an SVC R-D curve.

2.2.3. Optimal extraction paths

When we select an extraction path through an SVC bitstream for a specific viewing device, we intend to choose the *optimal extraction path* that offers the viewing device with the best R-D performance as prescribed by the following criteria.

Criterion 1. *Minimum underlying area for corresponding R-D (MSE) curve*[6]. The optimal extraction path $\Pi_\wp$ produced by successive refinement should be the one that has *minimum total underlying area* $\Omega_\wp$ for the corresponding R-D curve if *mean square errors*[7] *(MSE)* are used to measure the playback distortion of the extracted bitstream.

Criterion 2. *Convexity of corresponding R-D (MSE) curve.* The optimal extraction path $\Pi_\wp$ produced by successive refinement should have the corresponding R-D curve maintain its convexity[8] at *every refinement step.* More precisely, the R-D curve should have *monotonically decreasing MSE* values and R-D improvement γ at every step.

Note that among the two criteria, the first one is used as the optimization objective while the second one serves as a constraint.

2.2.4. Near-optimal Extraction Paths

In our experiments, we discovered in some rare cases (especially when subjective measures such as *mean opinion scores* are used to quantify the playback picture quality), some extraction paths with slightly non-convex R-D curves may have better performance than the ones with convex R-D curves. In those cases, we should choose a *near-optimal extraction path* that has the smallest area underneath its R-D curve while the deviation from the convexity of the R-D curve falls below a tolerance limit.

Criterion 3. *Tolerance limit for deviation from convexity.* An SVC extraction path can be considered as *near optimal* if and only if the *deviation from convexity* ζ of its R-D curve at any refinement step (as defined by the following formula) lies within a specified

[6] In the cases that the *peak signal-to-noise ratios (PSNR)* are used as the measure of playback distortion, the *minimum/maximum* conditions of the criteria must be reversed.

[7] The uncompressed videos with a spatiotemporal resolution that matches the display format of target devices are used as the references for MSE computations. Also, intermediate representations are interpolated to the same target format before measuring their MSE values.

[8] A R-D curve with distortion measured in terms of mean square errors (MSE) is *convex* or *concave upward* if and only if its *epigraph* (the sets of points lying on or above the curve) is a *convex set*.

tolerance limit and the *total underlying area of its R-D curve* is *minimum* among all the satisfying paths.

$$\zeta(S_b) \triangleq \frac{\epsilon_b}{r_{ac}} = \frac{r_{ab}d_{ac} - r_{ac}d_{ab}}{r_{ac}^2} \tag{9}$$

Fig. 2 illustrates the quantities appearing in Eq. 9 and offers a physical interpretation of the measure ζ. As shown in the figure, $\zeta(S_b)$ is a ratio between the increment in MSE distortion ϵ_b and the increment in bit rate r_{ac} within a non-convex segment $[S_a, S_b, S_c]$ of an R-D curve. This ratio must be *small* in order for the deviation from convexity to be deemed *acceptable*. This is particularly true at the early refinement steps, in which the increases in bit rates are moderate while the decreases in distortion are steep. Only minute deviation of convexity can be tolerated in those early steps.

3. Optimization of SVC bitstream extraction

Our investigation began with an attempt to devise strategies for finding an *optimal or near-optimal extraction path* of an SVC bitstream for a viewing device. The extraction path should be amenable to successive refinement of the SVC bitstream. The R-D performance along the extraction path is expected to satisfy Criterion 1, 2 or 3 stated in Section 2.

The approach we proposed for the search of an optimal SVC extraction path $\widehat{\Pi}_{\wp}(\overline{L}, \overline{T}; \widehat{L}, \widehat{T})$ consists of two steps: (1) the discovery of the extraction paths $\Pi_{\wp}(\overline{L}, \overline{T}; \widehat{L}, \widehat{T})$ with convex R-D curves and (2) the selection among these convex paths, the *optimal extraction path* $\widehat{\Pi}_{\wp}(\overline{L}, \overline{T}; \widehat{L}, \widehat{T})$ with the smallest total area $\Omega_{\wp}$ underneath its R-D curve.

In the cases that the selected path $\widehat{\Pi}_{\wp}$ has a sub-optimal R-D performance—i.e., its convex R-D curve has a total underlying area larger than that of some non-convex R-D curves with the same base and target representations—then we shall search for a *near-optimal extraction path* $\widetilde{\Pi}_{\wp}(\overline{L}, \overline{T}; \widehat{L}, \widehat{T})$ as a substitute. The search process is a mere repetition of the previous steps except that the extraction paths with non-convex R-D curves that contain tolerable deviation from convexity [Criterion 3] are considered in the selection process.

3.1. Graphical tools

To aid our search for the optimal/near-optimal extraction path of an successively refined SVC bitstream, we developed two graphical tools and named them, the *R-D mesh* and the *trellis diagram* of the bitstream. Following paragraphs explain the essence and the uses of these tools.

For the sake of examining the *R-D improvement* contributed by different refinement steps, we displayed in a single diagram all the piecewise-linear *R-D curves* of the extraction paths produced by successive refinement of an SVC bitstream. The R-D curves form a mesh, which we call the *R-D mesh* of the SVC bitstream. Every node in the R-D mesh represents a scalable layer representation $S(L, T)$ in the bitstream and is labeled explicitly by its layer L and temporal T identifiers. The coordinates (r, d) of the node represent the bit rate and the distortion of $S(L, T)$. Every line segment in the mesh, on the other hand, corresponds to a refinement step π in either L or T dimension:

$$\pi_L(L,T) : S(L,T) \to S(L',T) \tag{10a}$$
$$\pi_T(L,T) : S(L,T) \to S(L,T') \tag{10b}$$

where L' and T' denote the subsequent spatial/quality and temporal layers. The slope of each segment equals to the negation of the R-D improvement contributed by the corresponding refinement step.

Similarly, for the sake of exhibiting all possible extraction paths of an SVC bitstream, we superimpose them onto a grid of all scalable layer representations $S(L, T)$ embedded in the bitstream, and call the composite diagram the *trellis diagram* of the SVC bitstream. Again, every node and edge in the trellis diagram represents a scalable layer representation and a refinement step, respectively. In the trellis diagram, however, the coordinates of the nodes are their identifier values (L, T) while the edges are explicitly labeled with the R-D improvement $\gamma_L(L, T)$ and $\gamma_T(L, T)$ offered by the corresponding refinement steps. Plausible extraction paths and their segments are also drawn on top of the trellis diagram to illustrate the iterative process of searching for the optimal/near-optimal path.

Fig. 3 displays the R-D mesh and the trellis diagram of the Aikyo test video clip. Box (a) shows the *R-D mesh*; box (b) shows the *trellis diagram*, and box (c) gives a conceptual rendering of a simple trellis diagram. These tools are used in the rest of this paper both to expound the search strategies and to interpret the experiment results.

3.2. Discovery of R-D convex extraction paths

The discovery of the extraction paths with convex R-D curves (referred hereafter as *R-D convex paths*) is carried out by concatenating overlapping R-D convex segments that share at least one refinement step. We define an *R-D convex segment* as a section of an extraction path that has a convex R-D curve. Due to transitivity of inequality $(A > B \wedge B > C \Rightarrow A > B > C)$, the concatenation of two piecewise-linear functions that has monotonically decreasing slopes and share at least one overlapping segment is yet another function of the same sort. Hence, a convex extraction path can be constructed by concatenating two or more overlapping R-D convex segments.

3.2.1. Intra-trellis and inter-trellis convex segments

All R-D convex extraction paths can be constructed from two *elementary* types of *R-D convex segments* as shown in Fig. 3(c):

(1) *Intra-trellis (local) convex segments*, which consist of two refinement steps, one of which in L and T dimensions:

$$\Pi'_L(L,T) = \pi_L(L,T) \| \pi_T(L',T) \tag{11a}$$
$$\Pi'_T(L,T) = \pi_T(L,T) \| \pi_L(L,T') \tag{11b}$$

Each of these convex segments traverses a single four-node trellis.

(2) *Inter-trellis (global) convex segments*, which also consist of two refinement steps, both of them in either L or T dimensions:

$$\Pi_L(L,T;L'',T) : \pi_L(L,T) \| \pi_L(L',T) \tag{12a}$$
$$\Pi_T(L,T;L,T'') : \pi_T(L,T) \| \pi_T(L,T') \tag{12b}$$

Each of these inter-trellis convex segments traverses two connected trellises in L or T dimension.

The existence of intra-trellis segments Π'_L and Π'_T cannot be controlled directly by the setting of SVC encoding process. However, they can be verified by comparing the R-D improvement γ_L or γ_T of their first refinement steps $\{\pi_L, \pi_T\}$ against the R-D improvement Γ' of the intra-trellis segments $\{\Pi'_L, \Pi'_T\}$:

$$\Pi'_L(L,T) \text{ exists iff } \gamma_L(L,T) \geqslant \Gamma'(L,T) \tag{13a}$$
$$\Pi'_T(L,T) \text{ exists iff } \gamma_T(L,T) \geqslant \Gamma'(L,T) \tag{13b}$$

The existence of inter-trellis segments Π_L and Π_T, nonetheless, can be manipulated indirectly by the setting of *quantization parameter* QP, *inter-layer dependencies* Ψ_L and *temporal dependencies* Ψ_T among the SVC coding layers. In fact, as mentioned in Section 4, R-D convex paths in L and T dimensions, $\Pi_L(\overline{L}, T; \widehat{L}, T)$ or

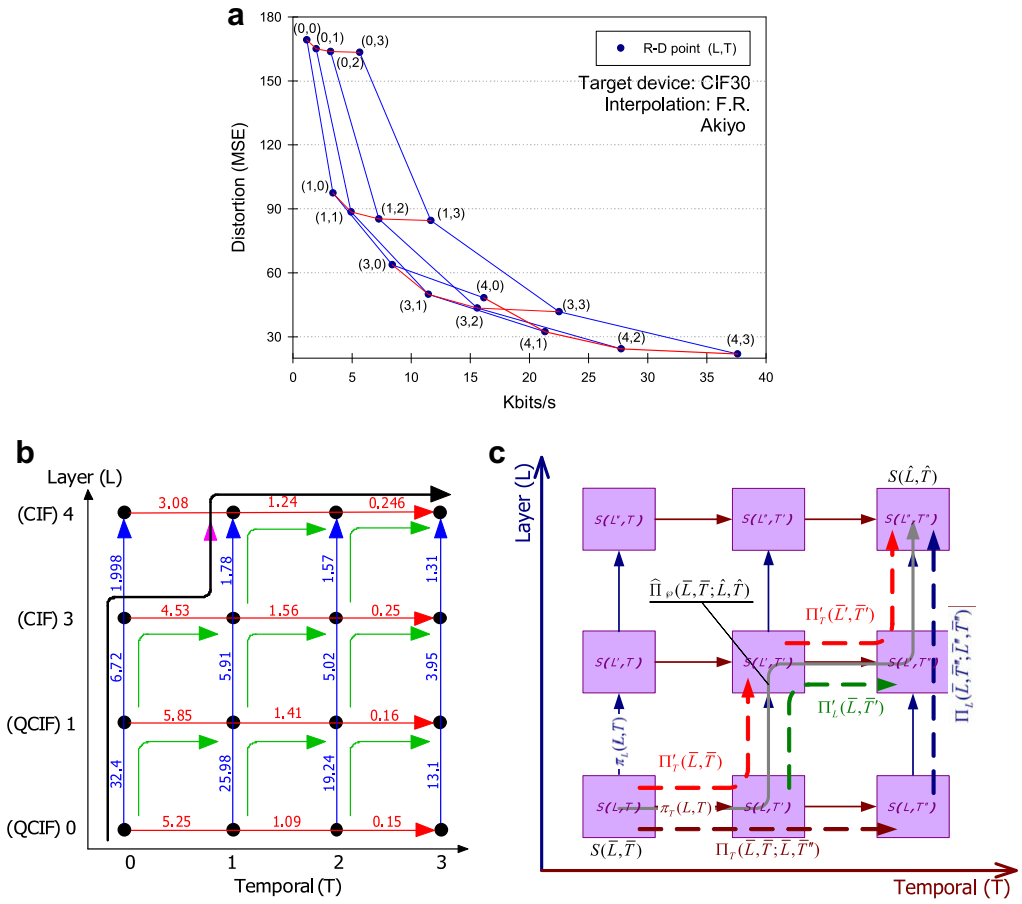

Fig. 3. R-D mesh and trellis diagram of an SVC bitstream, Akiyo (CIF30).

$\Pi_T(L,\overline{T};L,\widehat{T})$, may exist at every L and T values if QP, Ψ_L and Ψ_T satisfy certain constraints for *well-adapted SVC encoding*. The discovery of this correlation between SVC encoder setting and decoder (extraction) operation is a major contribution of this paper.

3.2.2. Connected trellises and extended convex segments

The intra-trellis (local) and inter-trellis (global) convex segments are the basic building blocks of R-D convex extraction paths. A convex extraction path can be constructed by concatenating at least two *overlapping R-D convex segments* that share a common refinement step. Fig. 4 illustrates the construction of convex extraction paths that traverse two connected trellises in L dimension (a) and T dimension (b), respectively. In these connected trellises, there exists four possible combinations. Three of them produce unique R-D convex paths between the diagonal end points $S(L,T)$ and $S(L'',T'')$ as the overlapping convex segments form a single connected path in each of these cases. In the forth case, however, *two* R-D convex paths (each of which consists of one intra-trellis and one inter-trellis convex segment) are formed between the end points, and the uniqueness of convex paths is violated. It is also possible for a trellis to contain two intra-trellis convex segments. These cases can simply be treated as the combination of the cases with single intra-trellis convex segments. They are further studied in Section 3.3.1 and 3.3.2.

Once we find the convex segments that traverse the connected trellises, we can discover all R-D convex extraction paths $\Pi_\wp(\overline{L},\overline{T};\widehat{L},\widehat{T})$ by concatenating the overlapping convex segments that exist in all the *consecutive pairs of connected trellises* lying between $S(\overline{L},\overline{T})$ and $S(\hat{L},\hat{T})$. In our approach, we regard two trellises as being *connected* if they share a common edge, and two pairs of connected trellises as being *consecutive* if they share a common trellis. All the R-D convex extraction paths $\Pi_\wp(\overline{L},\overline{T};\widehat{L},\widehat{T})$ form a directed graph between $S(\overline{L},\overline{T})$ and $S(\hat{L},\hat{T})$.

3.3. Search for optimal/near-optimal extraction paths

After we discover all the R-D convex extraction paths $\Pi_\wp(\overline{L},\overline{T};\widehat{L},\widehat{T})$, we can find the optimal extraction path $\widehat{\Pi}_\wp(\overline{L},\overline{T};\widehat{L},\widehat{T})$ by comparing the underlying area $\Omega_\wp$ of these convex paths. However, since the calculation of $\Omega_\wp$ involves the decoding of all $S(L,T)$ along the convex path $\Pi_\wp$, the operation can be computationally intensive. In the following paragraphs, we propose several strategies that can reduce the computation involved in the search for the optimal/near-optimal extraction path under different R-D conditions.

3.3.1. Strong local conditions

The search strategy becomes the simplest when *only one intra-trellis convex segment*, either $\Pi'_L(L,T)$ or $\Pi'_T(L,T)$, exists in every trellis. This situation arises when there is a *clear domination of R-D improvements* in either L or T dimension:

$$\text{Only } \Pi'_L(L,T) \text{ exists iff } \min\left(\gamma_L(L,T),\gamma_L(L,T')\right) > \max\left(\gamma_T(L,T),\gamma_T(L',T)\right) \tag{14a}$$

$$\text{Only } \Pi'_T(L,T) \text{ exists iff } \min\left(\gamma_T(L,T),\gamma_T(L',T)\right) > \max\left(\gamma_L(L,T),\gamma_L(L,T')\right) \tag{14b}$$

With this *strong intra-trellis (local) condition* and the existence of convex R-D curves in every spatial/quality and temporal layer (referred to as the *global condition*), the search for the optimal

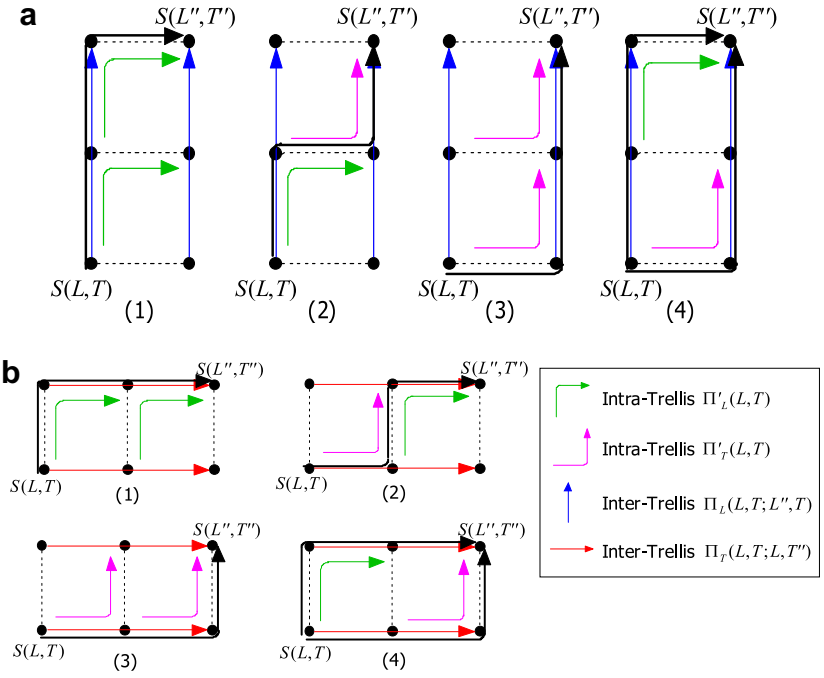

Fig. 4. Plausible concatenation of R-D convex segments among connected trellises: (a) in L dimension and (b) in T dimension.

extraction path can be reduced to a traversal of trellises based on the *steepest descent* of R-D improvement at every refinement step.

Proposition 1. *Search by steepest descent for the optimal extraction path under strong intra-trellis (local) R-D condition. If either Π_L or Π_T exists in every trellis of an SVC bitstream, then an optimal extraction path $\widehat{\Pi}_{\wp}(\overline{L},\overline{T};\hat{L},\hat{T})$ between $S(\overline{L},\overline{T})$ and $S(\hat{L},\hat{T})$ can be found by choosing the refinement step, $\pi_L(L_i,T_i)$ or $\pi_T(L_i,T_i)$, that offer the largest R-D improvement $\gamma(L_i,T_i)$ at every $S(L_i,T_i)$ between $S(\overline{L},\overline{T})$ and $S(\hat{L},\hat{T})$.*

$$\widehat{\Pi}_{\wp}(\overline{L},\overline{T};\hat{L},\hat{T}) = \|_{i=0}^{N-1}\{\pi_L(L_i,T_i),\pi_T(L_i,T_i)\}_{\max} \tag{15}$$

where

$$(L_0,T_0) = (\overline{L},\overline{T}),$$
$$(L_N,T_N) = (\hat{L},\hat{T}),$$
$$S(L_{i+1},T_{i+1}) = \begin{cases} \pi_L(S(L_i,T_i)) & \text{if} \quad \gamma_L(L_i,T_i) > \gamma_T(L_i,T_i) \\ \pi_T(S(L_i,T_i)) & \text{if} \quad \gamma_T(L_i,T_i) \geqslant \gamma_L(L_i,T_i) \end{cases} \tag{16}$$

This simple search strategy is feasible because there exists a *unique convex extraction path* between the base representation $S(\overline{L},\overline{T})$ and any target representation $S(\hat{L},\hat{T})$ if *both local and global conditions of R-D performance* are satisfied in an SVC bitstream. Fig. 5 illustrates a typical example. Note that the intra-trellis convex segments Π'_L and Π'_T tend to concentrate in two separate regions of the trellis diagram: Π'_L (drawn as magenta arrows) gathers in the upper-left corner while Π'_T (drawn as green arrows) gathers in the lower-right corner. Both types of convex segments bend their paths towards the boundary that separates the two regions. This distinct distribution of intra-trellis convex segments is a direct consequence of the dominant R-D improvement in either L or T dimension. The inequalities in Eqs. 14a and 14b prohibit the existence of two R-D convex paths in a pair of connected trellises and thus eliminate the chance for Π'_L (a magenta arrow) to appear underneath or to the right of Π'_T (a green arrow). The boundary between the two regions defines a convex and optimal extraction

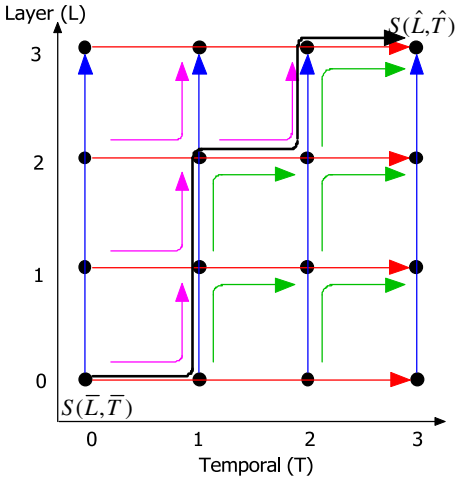

Fig. 5. A trellis diagram with convex segments satisfying *strong* intra-trellis (local) and inter-trellis (global) R-D conditions.

path (with maximum convexity and minimum underlying area) of the SVC bitstream because any other extraction path between the same end points would inevitably traverse at least one non-convex segment and thus yield a sub-optimal R-D performance. In the cases that the target representation $S(\hat{L},\hat{T})$ does not lie on the boundary, the optimal extraction path coincides with the boundary until it reaches either $S(\hat{L},T)$ or $S(L,\hat{T})$. Then, the path simply follows the spatial/quality layer $\mathscr{L}(\hat{L})$ or the temporal layer $\mathfrak{I}(\hat{T})$ until it reaches $S(\hat{L},\hat{T})$.

Before ending this discussion, we want to make two further comments on the steepest-descent search strategy:

(1) The steepest-descent strategy is the most computationally efficient method to search for optimal extraction paths. Our experiments showed that it merely needs to decode *approximately half* of the scalable layer representations lying

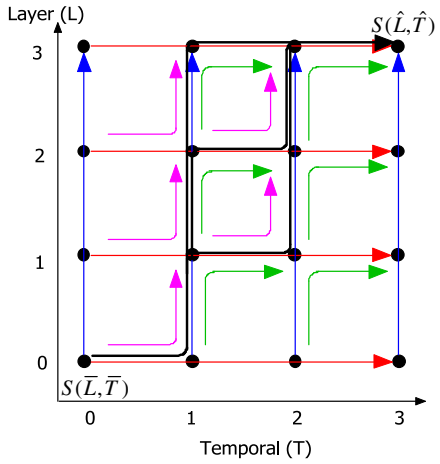

Fig. 6. A trellis diagram with convex segments satisfying *weak* intra-trellis (local) and inter-trellis (global) R-D conditions.

between the two end points. The improvement can be even more significant when an SVC bitstream contains a large number of scalable layers.

(2) Our experiments also showed that the steepest-descent strategy often found an extraction path very close to the optimal/near-optimal path (especially if an objective distortion measure such as MSE was used) even though the SVC bitstream did not satisfy the strong local condition. The differences in R-D performance among these paths were insignificant in most cases.

3.3.2. Weak local conditions

Among all the R-D trellises of an SVC bitstream, some of them contain R-D convex segments but lack a clear domination of R-D performance in either L or T dimension. In these cases of *weak local conditions*,[9] both Π'_L and Π'_T exist in each of these trellises. The existence of multiple convex segments in one or more trellises revokes the unique existence of convex extraction path and thus refutes the use of steepest-descent as the proper strategy to search for the optimal extraction path. In every cluster of connected trellises that satisfy the weak local conditions, *exhaustive search* must be used to find all the R-D convex extraction paths in the cluster. These paths can be concatenated with the convex extraction paths in the other parts of the trellis diagram to form the plausible candidates for optimal extraction. Fig. 6 shows an example of this non-ideal situation. All the marked extraction paths between $S(\overline{L},\overline{T})$ and $S(\hat{L},\hat{T})$ must be examined and have the area underneath their R-D curves be compared before the optimal path can be found.

Although the search strategy under weak local conditions is more elaborate than steepest-descent, it is less computation intensive than an exhaustive search over the entire trellis diagram. Note that the exhaustive search for R-D convex paths and the comparison of areas underneath corresponding R-D curves only take place within the clusters of connected trellises.

3.3.3. Fractional violation of local or global conditions

In some rare cases (when a subjective distortion measure such as the mean opinion scores is used to quantify playback picture quality), the local R-D condition (i.e. the existence of intra-trellis R-D convex segments) and/or the global R-D condition (i.e. the convexity of R-D curves along spatial/quality and temporal layers) may fail to be upheld. As a result, no convex extraction path exists between some $S(\overline{L},\overline{T})$ and $S(\hat{L},\hat{T})$ pairs. A *near-optimal extraction path* $\widetilde{\Pi}_{\wp}(\overline{L},\overline{T};\hat{L},\hat{T})$ with a slightly non-convex R-D curve (Section 2.2.4) may have to be accepted as a substitute instead. In the search for the near-optimal extraction path, extraction path segments with R-D curves that contain slight deviation from convexity [Criterion 3] are included into consideration. Steepest descent or localized exhaustive search strategies may be used to find the near-optimal path depending whether the trellises contain one or more extraction path segments. Fig. 7 provides an example that contains a violation of local R-D condition in the lower-left trellis. A slightly non-convex extraction path segment $\Pi'_T(0,0)$ shown as a dashed magenta arrow is included in the concatenation of path segments. Two plausible extraction paths were found. Among them, the path $\widetilde{\Pi}_{\wp}$ that traverses the sequence of nodes $[(0,0),(0,1),(1,1),(1,2),(2,2)]$ is chosen because it has the smallest area underlying its R-D curve.

[9] The *weak intra-trellis (local) condition* of R-D performance satisfies Eqs. 13a and 13b but not Eqs. 14a and 14b.

4. Production of well-adapted SVC bitstreams

The second part of our investigation aimed at establishing the necessary criteria that must be satisfied during SVC encoding in order to guarantee the existence of optimal/near-optimal extraction paths. Specifically, we examined the combined effects of *QP settings* and *inter-layer dependencies* Ψ_L on the *R-D performance* of an SVC bitstream.

4.1. Settings of quantization parameters

One important issue in SVC encoding is to determine the QP value for every spatial/quality layer so that the resultant bitstream can meet the predefined quality or bit rate constraints. Although the application requirements seem arbitrary, it should be noted, however, that improper QP settings may produce redundant representations and thus ill-formed R-D performance. To this end, we proposed two criteria for evaluating the properness of QP settings when combined scalability is in use.

Criterion 4. *Monotonic decrease in QP value for successive refinement.* In a given spatial resolution, the QP value of quality layers should *decrease* monotonically from one layer to the next in order to *successively* refine the *textural information.*

Criterion 5. *Elimination of redundant representations.* For different spatial resolutions, the high-resolution layers should have a reconstruction quality (fidelity) that is higher than that of the *spatially interpolated* low-resolution ones in order to eliminate *redundant representations.*

Criterion 4 requires the picture quality to be successively refined as the size of the bitstream increases by extracting more quality layers. Criterion 5 further prevents redundant layers from being encoded. We say that a high-resolution layer is redundant if there exists another low-resolution layer that can provide the same or even higher fidelity by simply performing spatial interpolation. Clearly, such redundancy should be detected and removed during SVC encoding.

In particular, the two criteria specify only the relative QP levels among the spatial and quality layers, i.e., the exact values still need to be decided by the intended applications. For instance, by focusing our attention on mobile streaming applications, in our experiments the PSNR of spatial/quality layers is set to fall between 27 dB and 35 dB. Exhaustive encoding was carried out offline to obtain proper QP settings for different test sequences.

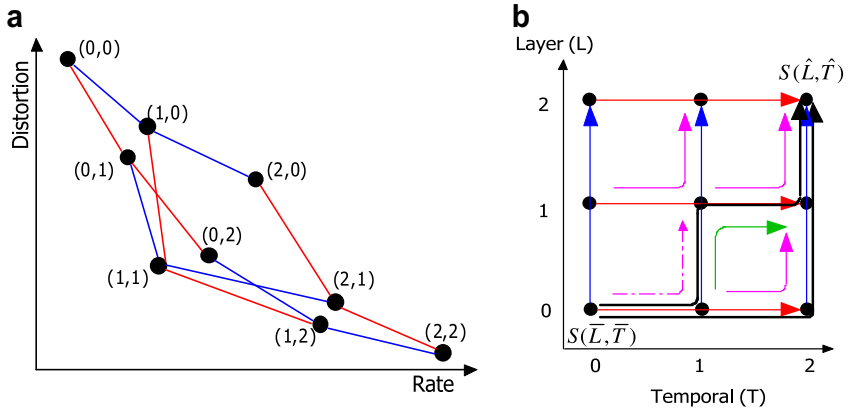

Fig. 7. R-D mesh and trellis diagram of an SVC bitstream with fractional violation of intra-trellis (local) R-D conditions.

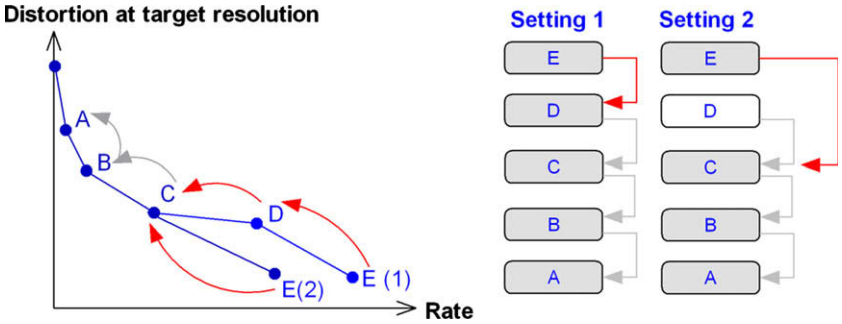

Fig. 8. R-D performance of SVC bitstreams with different inter-layer dependency settings. Labels A, B, C, D, and E denote five coding layers of different SNR levels with E being the target layer for reconstruction.

4.2. Settings of inter-layer dependencies

In our efforts to devise efficient search strategies for optimal/near-optimal extraction paths, we discovered that the global condition can be satisfied by maintaining the R-D convexity across spatial/quality and temporal layers during SVC encoding.

With hierarchical and dyadic temporal dependencies, the cascading QP assignment in current JSVM [10] can already make the R-D curves across temporal layers convex in most cases, especially when MSE is used for distortion measure. This is because higher temporal layers are coded with larger QP values, which inherently leads to diminishing R-D improvement with increasing temporal level.

On the other hand, the R-D convexity along the spatial/quality dimension relies on a well-adapted inter-layer dependency, which must satisfy the following criterion.

Criterion 6. *Convexity of rate-distortion curve across spatial and quality layers.* An SVC encoder should produce an SVC bitstream according to a *well-adapted inter-layer (spatial and quality) dependence relation* $\hat{\Psi}_L$ that ensures every successive refinement of scalable layer representations $[S(L_i,\hat{T})]_{i=0}^{N} = \hat{\Psi}_L^*(S(\hat{L},\hat{T}))$ from $S(\bar{L},\hat{T})$ to $S(\hat{L},\hat{T})$ with $L_0 = \bar{L}, L_N = \hat{L}$ and $\hat{T} = T_{\max}$ exhibits a *monotonic decrease in MSE distortion* $d(L_i,\hat{T}) > d(L_{i+1},\hat{T})$ as well as a *monotonic decrease of R-D improvement* $\gamma_L(L_i,\hat{T}) > \gamma_L(L_{i+1},\hat{T}) > 0$.

This criterion forbids the *slope* of the R-D curve to steepen (or equivalently the *R-D improvement* to rise) as a viewing device takes in a sequence of coding layers in successive refinement steps. Its practical implication can be explained using the example shown in Fig. 8. In the example, each layer (from B to E) in Setting #1 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 R-D improvement produced by D is not as good as the one produced by E, Setting #1 cannot maintain the R-D convexity. In contrast, Setting #2, which links C directly to E by skipping D, is a well-adapted dependency setting.

We must advise readers to exercise caution when they try to set up a well-adapted inter-layer dependency because the adaptation can easily be overdone. In Fig. 8, although Setting #2 (which ensures R-D convexity along the spatial/quality dimension) produces a better R-D performance for a single viewing device even if it takes layer E in one moment and layer D in another, Setting #1 (which fails to maintain R-D convexity) consumes less bandwidth when it comes to serving two viewing devices existing in the same network. This observation confirms a well-known fact that the SVC coding gain over simulcasting is at the cost of the R-D performance of individual layers. Our advice of caution can be summarized in the following proposition.

Proposition 2. *Minimal adaptation of successive inter-layer dependencies. An SVC encoder should choose a successive inter-layer dependence relation*[10] $\hat{\Psi}_L(L,\hat{T}) : S(L,\hat{T}) \to S(L-1,\hat{T}), \forall L > \bar{L}$ *to be the default dependency setting. The dependence relation should only be modified at the refinement steps that produce non-convex R-D improvements. At those refinement steps, the reference layers* $\hat{\Psi}_L(S(L,\hat{T}))$ *should be chosen to be the nearest spatial/quality layers that can produce convex R-D improvements.*

Again using the example in Fig. 8, a proper adjustment of inter-layer dependencies is to make E depend on C rather than B. This minimal adjustment of inter-layer dependence relations shall only cause a small increase in the total bit rate of the SVC bitstream.

[10] *The default successive inter-layer dependence relation usually produces a lowest total bit rate.*

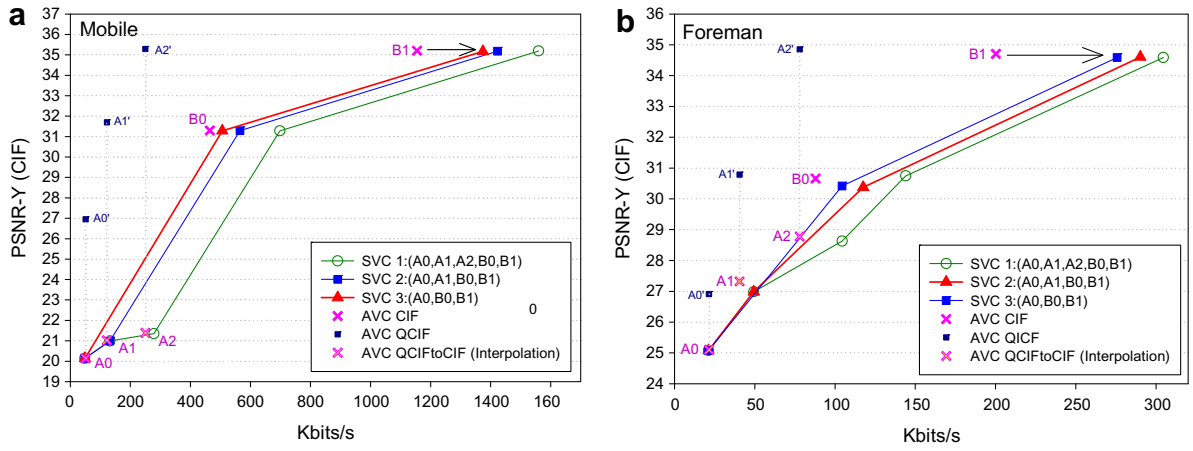

Fig. 9. Comparison of R-D performance between SVC and H.264/AVC for the sequences "Mobile" and "Foreman". The R-D curves of SVC with various dependency settings were produced using the bottom-up encoding process and the fixed-quality configurations.

Moreover, it should be noted that such strategy is to promise the R-D convenxity along the spatial/quality dimension rather than to optimize the R-D performance for individual coding layers. For the latter, the reader is referred to the paper published by Yao et al. [16] for more complete discussion.

5. Implementation of well-adapted inter-layer dependencies

Having described the criteria for well-adapted bitstreams, we further presents in this section a practical approach to establishing well-adapted inter-layer dependencies among the coding layers of an SVC bitstream.

5.1. Prediction of R-D convexity

To predict the R-D convexity of an SVC bitstream along the spatial/quality dimension, one effective approach is to evenly add 10% or more redundancies[11] to the R-D points of H.264/AVC [12]. The results generally hold when the *multi-loop encoder control* and the *fixed-quality* configurations are used [5,12]. Moreover, the predictability remains valid with the *bottom–up encoding process* [10] if the fact that enhancement layers usually suffer more coding efficiency losses than their reference layers is additionally taken into account. These observations enable us to predict the R-D convexity of various dependency settings by inspecting the R-D points of H.264/AVC.

For validation several SVC bitstreams, each corresponds to one of the following dependency settings, were firstly encoded using the bottom–up encoder control and the fixed-quality configuration. In particular, Setting #1 denotes the default dependency setting (in which each layer simply depends on its previous layer), whereas Settings #2 and #3 adapt the default setting by merely changing the reference layer of layer B0. The R-D performance of these dependency settings is compared with that of H.264/AVC in Fig. 9.

- *Setting #1:* (QCIFA0 ← A1 ← A2), (CIF A2←B0←B1).
- *Setting #2:* (QCIF A0←A1←A2), (CIF A1←B0←B1).
- *Setting #3:* (QCIF A0←A1←A2), (CIF A0←B0←B1).

In Fig. 9, one can deduce from the R-D points of H.264/AVC that Setting #3 would be a well-adapted setting for the sequence "Mobile". The result can be readily shown correct by comparing the R-D curves of SVC. Llikewise, in the sequence "Foreman", both Settings #2 and #3 are likely to ensure the R-D convexity. Although Setting #3 has maximal R-D convexity, Setting #2 is considered a better choice because, as will be seen in the next section, it incurs a minimal increase in the total bit rate of the bitstream.

In Fig. 10 we further present the results with the *fixed-rate* configurations, in which the quality (and the QP) of each layer is not fixed; rather, the cumulative rate to each layer is kept constant regardless of the dependency settings. Comparing with the R-D performance of H.264/AVC, the coding efficiency loss of SVC can be seen from the drop of its R-D curves. Similar to the fixed-quality configuration the increasing PSNR drop along the spatial/quality dimension helps to predict the R-D convexity of various dependency settings when an SCV bitstream is encoded with a fixed-rate. From Fig. 10, we can obtain exactly the same prediction results as with fixed-quality configurations. Interestingly, in the sequence "Foreman", there is a "bump" in the R-D curve with Setting #1. This is because the QP value of layer B0 is improperly chosen to meet the bit rate constraint. The result stresses the importance of having proper QP settings.

Our discussions so far have assumed that the R-D points of H.264/AVC are available prior to the convexity prediction. The assumption does not generally hold unless each layer is pre-encoded with H.264/AVC. Although collecting the R-D data of H.264/AVC seems time consuming, performing exhaustive SVC encoding with various dependency settings incurs much higher complexity. In addition the proposed approach merely ensures the R-D convexity at *full frame rate*. Nevertheless, the global condition requires the R-D convexity to be upheld at *all possible frame rates* supported by an SVC bitstream. We have found empirically from the extensive experiments that maintaining the R-D convexity at full frame rate would also likely to ensure the convexity at all the lower frame rates. After all, the R-D behavior at full frame rate represents the average performance over all video frames.

5.2. Degradation in coding efficiency

Having analyzed the R-D convexity under various dependency settings, we now turn our attention to the coding efficiency of an SVC bitstream, which is characterized by its total bit rate. As described previously, long-term inter-layer reference may be needed for the sake of R-D convexity. It is natural then to question whether and to what extent the total bit rate will increase. The answers can be found by the comparison shown in Fig. 11, where it can be seen that the well-adapted dependency settings (Setting #2 for the Foreman; Setting #3 for the Mobile) incur, on average, 15—20% bit rate

[11] Comparing with the single layer coding, the coding efficiency loss of SVC is generally proportional to the number of coding layers. In some cases, the R-D gap between H.264/AVC and SVC can be much greater than 10%.

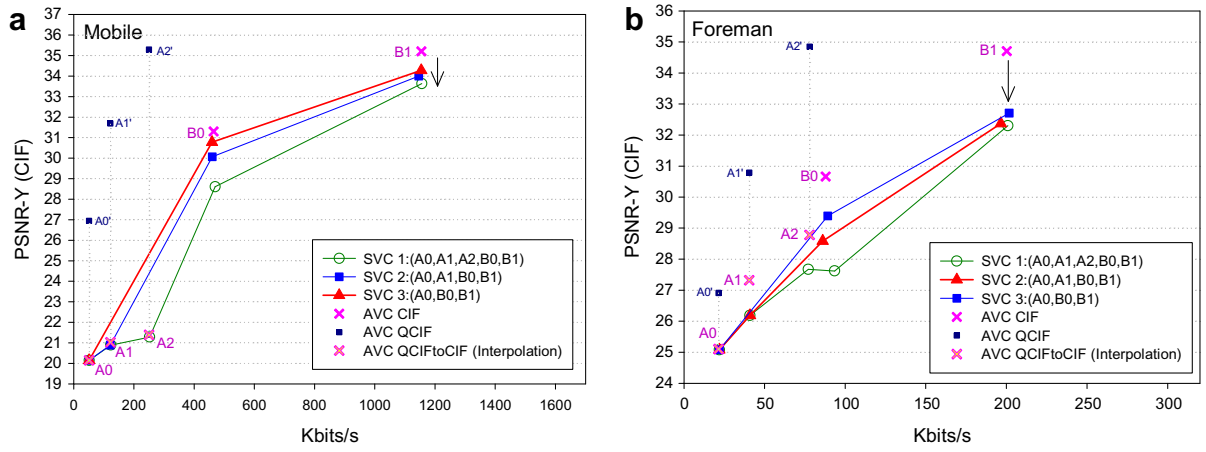

Fig. 10. Comparison of R-D performance between SVC and H.264/AVC for the sequences "Mobile" and "Foreman". The R-D curves of SVC with various dependency settings were produced using the bottom-up encoding process and the fixed-rate configurations.

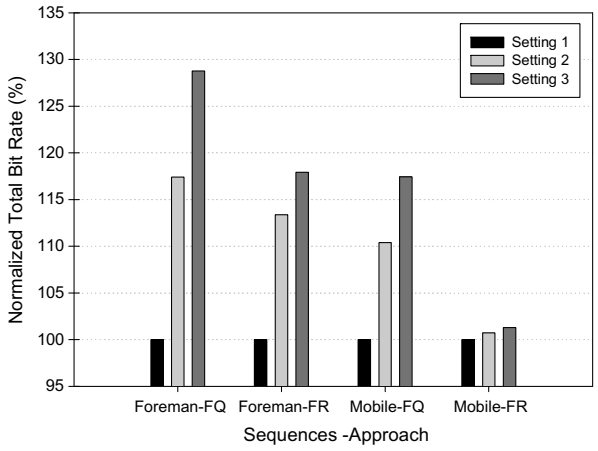

Fig. 11. Comparison of total bit rate with different dependency settings. Both the fixed-quality (FQ) and fixed-rate (FR) configurations were used.

increase in comparison with Setting #1 (default setting). The penalty arises mostly because the layers A1 and A2 are not utilized during the inter-layer prediction of the layer B0 in Settings #2 and #3.

6. Experiments and comparisons

6.1. Analysis of optimal extraction paths

In this section, we present a detailed analysis on the optimal extraction paths in regard to the following factors. The analysis is to understand how these factors may affect the choice of optimal extraction paths.

- *Video contents:* Static vs. Motion.
- *Device types:* QCIF @ 30/15 Hz, CIF @ 30/15 Hz, and 4CIF @ 30/15 Hz.
- *Distortion measures:* Mean-squared error vs. Mean opinion score.
- *Temporal interpolation:* Frame replication (F.R.) vs. B_Direct_16 × 16 (B.Direct).

Table 1 lists our testing conditions, in which the settings of QP values and inter-layer dependencies comply with the guidelines prescribed in Section 4. To simulate the actual use of SVC, extracted videos were interpolated to the highest spatiotemporal resolutions available on all viewing devices. The spatiotemporal interpolation was accomplished by the standard-compliant spatial filtering [10], followed by frame replication (F.R.) or motion field estimation (B.Direct) for temporal interpolation. While sophisticated temporal interpolation schemes could be used, we chose the two straightforward implementations not only because of their simplicity but also their popularity. In addition, in the experiments comparing subjective and objective distortion measures, we adopted the VQM software [1,9] to predict the subjective quality of decoded videos and computed the Mean Square Error (MSE) between the original and the decoded videos as an objective criterion.

6.1.1. Optimal paths versus video contents

The optimal extraction paths depend heavily on video contents. This is because the spatiotemporal characteristics of video contents crucially affect the efficiency of interpolation algorithms performed on viewing devices. Refining temporal quality normally results in better R-D performance in fast-motion sequences, whereas maintaining spatial or SNR quality is more beneficial in slow-motion sequences. The results can be seen by comparing the optimal paths in Fig. 12, where the MSE and the F.R. are used for distortion measure and temporal interpolation, respectively. Interestingly, most of the optimal paths shown in Fig. 12 are found to preferentially improve temporal quality except the ones for the sequence "Akiyo". The reasons are twofold. First, the MSE has difficulties in appreciating temporal quality. Second, the F.R. can very often yield erroneous results in video frames undergoing rapid temporal changes. The two facts together explain the dramatic increase in MSE if video frames are skipped, and thereby justify the tendency of optimal paths to improve temporal quality.

6.1.2. Optimal paths versus distortion measures

In addition to video contents, distortion measures also influence the choice of optimal paths. Fig. 13 compares the paths found by using the MSE and MOS criteria. It can be readily seen that the MSE-based extraction paths are biased towards temporal quality in comparison with the MOS-based solutions. The observation agrees with the general fact that the MSE is likely to overestimate the quality degradation caused by temporal jerkiness even if the impairment in perceptual quality is insignificant. On the other hand, the results using the MOS, although correlate much well with perceptual quality, are generally less analytical owing to the unpredictable nature of MOS. In view of the pros and cons of each measure, experimental results that follow are provided with both distortion criteria.

Table 1
Testing conditions and encoder parameters

Software		JSVM 9
Spatial scalability		QCIF (176 × 144), CIF (352 × 288), 4CIF (704 × 576)
Temporal scalability		GOP size = 8, frame rate = 3.75–30 Hz, hierarchical B pictures
SNR scalability		Coarse granularity scalability (CGS)
Inter-layer encoding		Adaptive motion, residual, textural predictions
Sequences	Inter-layer dependency	QP settings
Akiyo	QCIF(A0 ← A1←A2), CIF(A1 ← B0 ← B1)	QCIF(50,43,37), CIF(44,40)
Foreman	QCIF(A0 ← A1 ← A2), CIF(A1 ← B0 ← B1)	QCIF(46,40,34), CIF(41,34)
Football	QCIF(A0 ← A1 ← A2), CIF(A0 ← B0 ← B1)	QCIF(41,35,30), CIF(36,30)
Mobile	QCIF(A0 ← A1 ← A2), CIF(A0 ← B0 ← B1)	QCIF(41,35,30), CIF(34,28)
Harbor	CIF(A0 ← A1 ← A2), 4CIF(A0 ← B0 ← B1)	CIF(41,36,31), 4CIF(37,29)
ICE	CIF(A0 ← A1 ← A2), 4CIF(A1 ← B0 ← B1)	CIF(45,40,35), 4CIF(41,33)

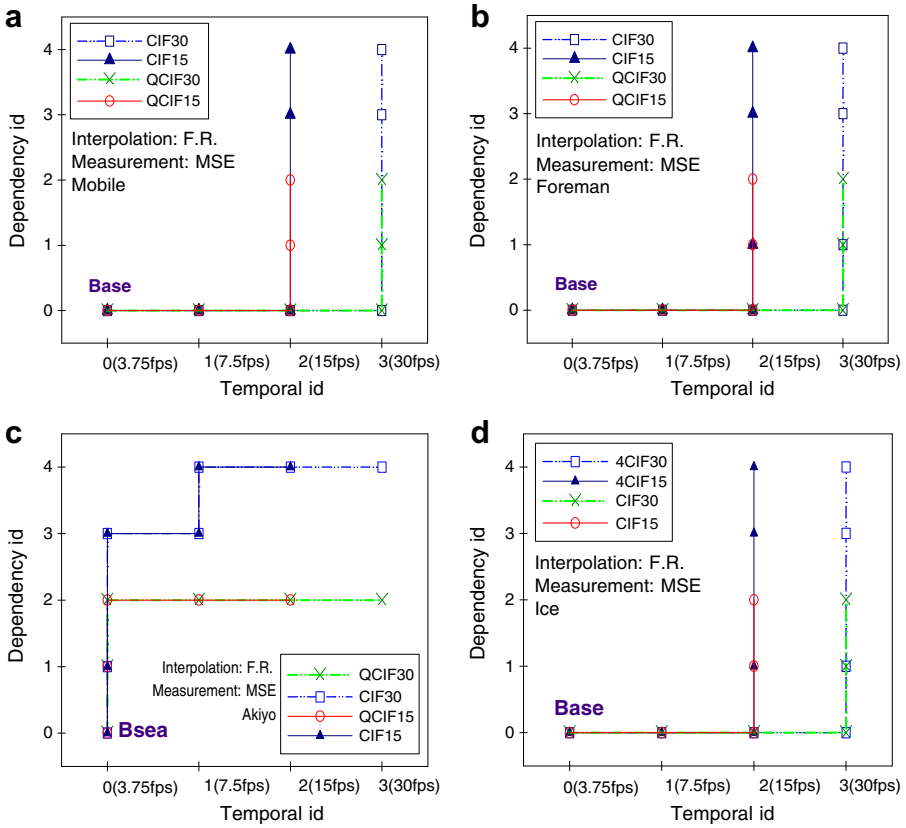

Fig. 12. Comparison of optimal extraction paths for different viewing devices: (a) Mobile, (b) Foreman, (c) Akiyo, and (d) ICE. The B.Direct and MSE are used for temporal interpolation and distortion measure, respectively.

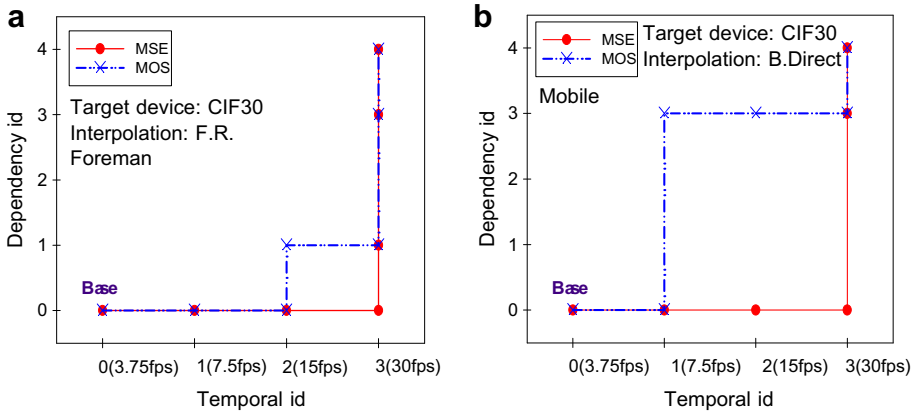

Fig. 13. Comparison of optimal extraction paths found by using MSE and MOS for distortion measure: (a) Foreman CIF@30 and (b) Mobile CIF@30.

6.1.3. Optimal paths versus spatiotemporal interpolation

Besides video contents and distortion measures, the interpolation algorithms performed by viewing devices also have a significant effect on the optimal paths. Moreover, the temporal interpolation is more critical than the spatial interpolation because poor performance could easily give rise to significant distortion and visible artifacts. In Fig. 14 the influences of temporal interpolation are analysed by comparing the optimal paths found by using the frame replication (F.R.) and the B_Direct_16 × 16 (B.Direct) on viewing devices. In general, the B.Direct method provides better quality than the straightforward F.R. owing to better estimation of motion fields. The fact also explains why the B.Direct method allows the extraction to improve more in spatial quality, while the F.R. causes it to extract more temporal layers. The results also confirm that further optimization would be made possible if the interpolation algorithms performed by viewing devices are provided.

Summarizing, in this section, we have shown that the choice of optimal extraction paths is determined by several factors: the visual characteristics of video contents, the distortion measures, and the spatiotemporal interpolation algorithms performed by viewing devices. All these factors are related directly or indirectly to the final playback quality and should be considered jointly during the optimization process of bitstream extractions.

6.2. Performance of steepest-descent search strategy

After the optimal extraction paths have been studied in details, this section further evaluates the performance of the steepest-descent strategy in search for optimal paths. Exhaustive search is used as baseline for comparison.

6.2.1. Extraction paths and R-D performance

Based on the MSE criterion and well-adapted bitstreams, Table 2 compares the optimal extraction paths found by the steepest-descent method and exhaustive search. The differences in path indices are contrasted utilizing exclusive-OR operation.

Clearly, from the table the two methods produce almost identical results. It has been found from the R-D trellis diagrams that both the *global* and the *strong local* conditions were satisfied in most of the test sequences, which explains the fairly good performance of the steepest-descent method. The global condition results largely from the well-adapted settings. The local condition, on the other hand, is more intricate in that it represents local R-D variations and may not be precisely controlled. In fact, no effort was made to adapt the QP or coding to take into account the local condition. The reason that the *strong local condition* holds in this particular set of experiments is mostly due to the MSE effect. Most

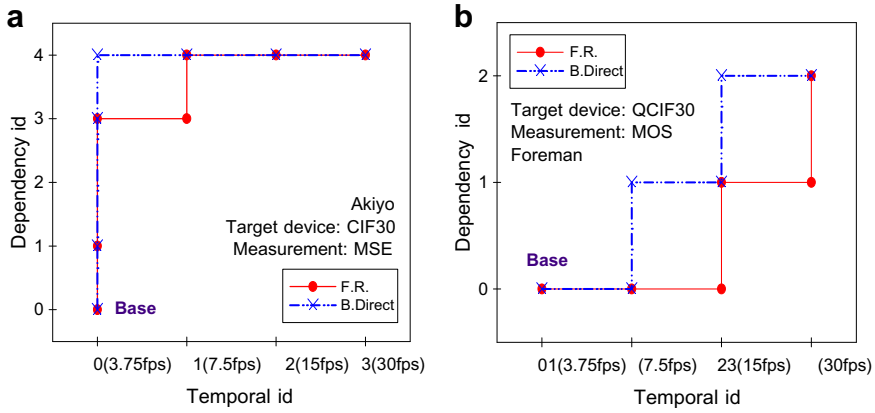

Fig. 14. Comparison of optimal extraction paths found by using frame replication (F.R.) and B_Direct_16 × 16 (B.Direct) for temporal interpolation: (a) Akiyo CIF@30 and (b) Foreman QCIF@30.

Table 2
Comparison of extraction paths with the MSE

	CIF30 Exh. MSE + F.R.	S.D.	XOR	CIF15 Exh.	S.D.	XOR	QCIF30 Exh.	S.D.	XOR	QCIF15 Exh.	S.D.	XOR
Akiyo	110100	110100	0	11010	11010	0	11000	10100	**01100**	1100	1100	0
Foreman	000111	000111	0	00111	00111	0	00011	00011	0	0011	0011	0
Mobile	00011	00011	0	0011	0011	0	00011	00011	0	0011	0011	0
Football	00011	00011	0	0011	0011	0	00011	00011	0	0011	0011	0
	4CIF30			4CIF15			CIF30			CIF15		
Harbor	00011	00011	0	0011	0011	0	00011	00011	0	0011	0011	0
ICE	000111	000111	0	001111	001111	0	00011	00011	0	0011	0011	0
	MSE + B.Direct											
Akiyo	111000	111000	0	11100	11100	0	11000	11000	0	1100	1100	0
Foreman	000111	000111	0	00111	00111	0	00011	00011	0	0011	0011	0
Mobile	00011	00011	0	0011	0011	0	00011	00011	0	0011	0011	0
Football	00011	00011	0	0011	0011	0	00011	00011	0	0011	0011	0
	4CIF30			4CIF15			CIF30			CIF15		
Harbor	00011	00011	0	0011	0011	0	00011	00011	0	0011	0011	0
ICE	000111	000111	0	001111	001111	0	00011	00011	0	0011	0011	0

of the local trellises in these test sequences are found to have a much higher preference for temporal quality.

The *strong local condition*, however, may be violated. One such example is the extraction of the sequence "Akiyo" for QCIF30 devices, in which *only* the *weak local condition* is satisfied. It has been shown in our theoretical framework that the steepest-descent method may fail to find the optimal solution in such a case. This can also be seen practically from Fig. 15 (a), where a wrong decision was made when two convex R-D segments appeared at the upper-left corner. However, even if the optimal solution cannot be found by using the steepest-descent method, we usually end up with a suboptimal path having very similar R-D performance to the optimal one (see the R-D comparison in Fig. 15(b)). This is because the greedy nature of the steepest-descent method causes it to always pick the R-D points that are closer to the convex hull.

Before closing this section, it is worth remarking on a few phenomena exhibited by Fig. 15(b). First, there are R-D points violating the general expectation that distortion should decrease as the bit rate increases, which is usually true when considering the R-D performance of video codec alone. However, Fig. 15(b) describes the *true* R-D behavior when decoded videos are presented on viewing devices, i.e., the distortion is measured against the interpolated videos rather than the decoded videos. Apparently, the results depend not only on the encoding algorithm, but also on the interpolation schemes performed by viewing devices. Second, the R-D optimized extraction offers significant improvement in playback picture quality. Without optimization, one may possibly choose an extraction path with extremely poor R-D performance. An example of such a path is illustrated by the dash curve in Fig. 15(b).

6.2.2. Computational complexity

The computationally most demanding part in search for optimal extraction paths is to collect the R-D data associated with each scalable layer representation. While the exhaustive search needs to actually decode every scalable layer representation, supported in an SVC bitstream the steepest-descent method reduces the computation by lazy evaluation and performs only half (42–58%) of the computation necessary for an exhaustive search. The gain is even more obvious when an SVC bitstream contains a large number of scalable layers.

6.3. Comparisons with other extraction schemes

This section compares and contrasts the major differences of our proposed scheme with other previous works, including the Basic Extraction in JSVM [10,7], the Quality Information Table [4], the Quality Index [6], as well as the Quality-Layers-based approach [2].

- *Applications:* The Quality-Layers-based extraction [2] aims at *medium-grain quality adaptation*, while the other schemes focus on *multi-dimensional adaptation* with *combined scalability*. In particular, the Quality-Layers-based approach [2] is conditioned on the full extraction of the base layer, whereas the others allow performing R-D optimal extraction without the presence of the entire base layer.
- *Extraction constraints:* Both our scheme and the Quality-Layers-based extraction must perform incremental bitstream extraction for *successive refinement*, while the others allow discretionary extraction. Through *successive refinement*, coarser representations are always embedded in finer ones, which leads to more efficient use and share of extracted bitstreams among viewing devices. The differences in bitstream extraction with and without *successive refinement* are shown in Fig. 16 using Venn diagram.
- *Extraction criteria:* Most of the schemes perform bitstream extraction based on the R-D performance of extracted bitstreams except the Basic Extraction approach, which carries out extraction in such a way that the resultant bitstream must have a bit rate that is closest to but not greater than a target bit rate. As it has been shown in our R-D analysis, decoding a substream with a higher bit rate does not necessarily produce better playback quality, especially when spatiotemporal interpolation is involved.
- *Distortion measurement:* Both our scheme and the Quality-Index-based approach compute the R-D data using *interpolated videos* rather than *decoded videos*. Also, as indicated in our analysis, the *interpolated videos* can more realistically reflect the playback quality on viewing devices. An even more direct approach is to

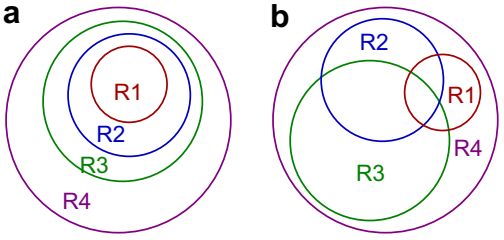

Fig. 16. Bitstream extraction (a) with and (b) without successive refinement. R1–R4 indicate the extracted bitstreams associated with the increasing bit rate.

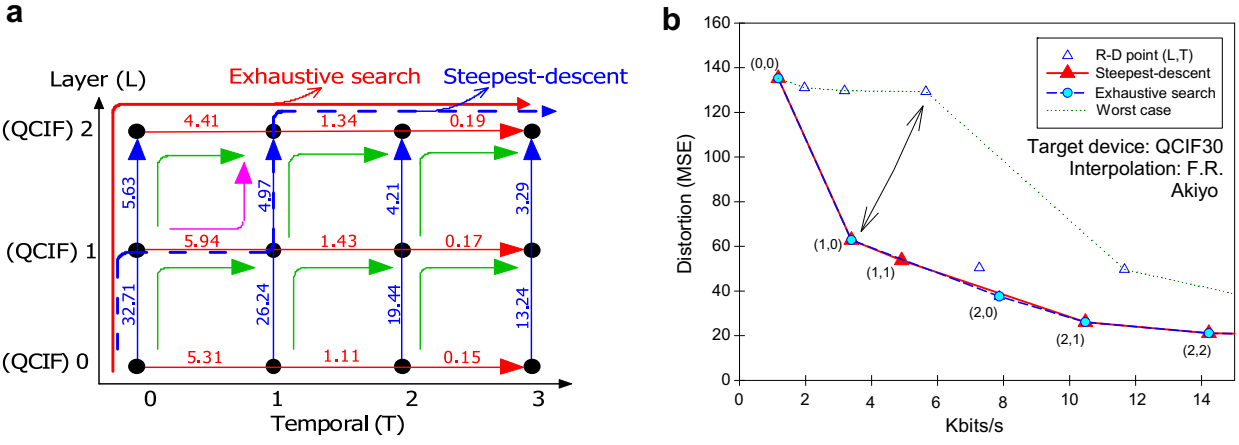

Fig. 15. Comparison of extraction paths found by using the steepest-descent method and the exhaustive search: (a) R-D trellis diagram and (b) R-D curves.

acquire the perceptual preference, as used in the Quality Information Table. However, it would be impossible to have subjective evaluation for every video sequence.

- *Rate-distortion performance:* While most previous works simply try to find an R-D optimized extraction path for pre-encoded SVC bitstreams, in this paper we further recommended a set of criteria for generating well-adapted bitstreams, which promise the R-D convexity along optimal extraction paths.
- *Search strategy and complexity:* Through the use of well-adapted settings, our steepest-descent method can very often find the optimal/near-optimal candidates while reducing the complexity by 50% or more in comparison with the exhaustive search that was adopted by most previous works.

7. Conclusions

In our work, we attempted to approach the task of rate-distortion (R-D) optimized SVC bitstream extraction from a new direction. Our approach was characterized by three unique considerations: (1) the combined effect of proper encoder setting coupled with matching bitstream extraction and decoding mechanisms, (2) the computation efficiency of search strategies for R-D optimized extraction paths, and (3) the choice of extraction paths amenable to successive refinement of SVC bitstreams.

Through theoretical analysis of SVC inter-layer dependence relations and empirical study of the R-D performance of different encoded/extracted bitstreams, we obtain the following discoveries:

(1) An optimal extraction path (corresponding to a convex R-D curve with minimal underlying area) can be found for an SVC bitstream if convex R-D performance can be maintained at every spatial/quality layer as well as temporal layers (referred as the global R-D conditions) and in every pair of successive refinement steps (referred as the local R-D conditions). If the convexity of R-D performance is violated only by minor deviations occur in a small fraction of all refinement steps then a near-optimal extraction path can be found.

(2) Convex R-D performance can be maintained across spatial/quality layers by adapting the inter-layer dependencies between different layers and the quantization parameter QP of individual layer during SVC encoding. The R-D convexity of SVC layers (especially the spatial layers) can be predicted by referring to the R-D performance of corresponding H.264/AVC bitstreams encoded with fixed-quality or fixed-rate settings. On the other hand, convex R-D performance across temporal layers can be ensured by the proper cascade of QP values over the hierarchy of temporal layers.

(3) The greedy steepest-descent strategy can be employed to search for the unique optimal extraction path if the SVC bitstream can satisfy both global R-D conditions and strong local R-D conditions. The steepest-descent strategy is most computationally efficient as it decodes only half of the scalable layer representations in comparison with the exhaustive search strategy that was adopted by most previous works. Beside of being efficient, our experiments showed that the steepest-descent strategy is also relatively robust with respect to its search results. The strategy can always find a sub-optimal extraction path close to the optimal path even under weak local R-D conditions. The strategy can even find the near-optimal extraction path when the global and local R-D conditions are violated in parts as when a subjective quality measure such as mean opinion scores (MOS) is used to quantify R-D performance.

Our work is still in its early stage, we plan to extend our investigation in several directions: (1) to study R-D optimized encoding and bitstream extractions for the SVC bitstreams with medium-grain scalability (MGS) support, (2) to conduct experiments with error concealment techniques and finally, and (3) to devise computationally efficient strategies to search for optimal/near-optimal extraction paths under weak or fractional violation of global and local R-D conditions.

References

[1] ITS Video Quality Research. Available from: <http://www.its.bldrdoc.gov/n3/video/index.php>.
[2] I. Amonou, N. Cammas, S. Kervadec, 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 17 (September) (2007) 1186–1193.
[3] H.C. Huang, W.H. Peng, T. Chiang, H.M. Hang, Advances in the scalable amendment of H.264/SVC, IEEE Communications Magazine 45 (2007) 68–76.
[4] Y.S. Kim, Y.J. Jung, T.C. Thang, Y.M. Ro, Bit-stream extraction to maximize perceptual quality using quality information table in SVC, in: SPIE Conference on Visual Communications and Image Processing (VCIP), vol. 6077, January 2006.
[5] Z.G. Li, S. Rahardja, H. Sun, Implicit bit allocation for combined coarse granular scalability and spatial scalability, IEEE Transactions on Circuits and Systems for Video Technology 16 (12) (2006) 1449–1459.
[6] J. Lim, M. Kim, S. Hahm, K. Lee, K. Park, An optimization-theoretic approach to optimal extraction of SVC Bitstreams, in: ISO/IEC JTCI/SC29/WG11 and ITU-T SG16 Q.6, JVT-U081, October 2006.
[7] H. Liu, H. Li, Y.K. Wang, Showcase of scalability information sei message, in: ISO/IEC JTCI/SC29/WG11 and ITU-T SG16 Q.6, JVT-Q067, October 2005.
[8] W.H. Peng, J.K. Zao, T.W. Wang, H.T. Huang, Multidimensional SVC bitstream adaptation and extraction for rate-distortion optimized heterogeneous multicasting and playback, in: IEEE International Conference on Image Processing (ICIP), October 2008.
[9] M. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Transactions on Broadcasting 50 (3) (2004) 312–322.
[10] J. Reichel, H. Schwarz, M. Wien, Joint scalable video model JSVM-9, in: ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, JVT-V202, January 2007.
[11] H. Schwarz, D. Marpe, T. Wiegand, Overview of the scalable video coding extension of the H.264/AVC standard, in: IEEE International Conference on Image Processing (ICIP), October 2006.
[12] H. Schwarz, T. Wiegand, Further results for an RD-optimized multi-loop SVC encoder, in: ISO/IEC JTCI/SC29/WG11 and ITU-T SG16 Q.6, JVT-W071, April 2007.
[13] D. Taubman, High performance scalable image compression with EBCOT, in: IEEE International Conference on Image Processing (ICIP), October 1999.
[14] Y.-K. Wang, M. Hannuksela, S. Pateux, A. Eleftheriadis, S. Wenger, System and transport interface of SVC, IEEE Transactions on Circuits and Systems for Video Technology 17 (9) (2007) 1149–1163.
[15] T. Wiegand, G. Sllivan, J. Reichel, H. Schwarz, M. Wien, Joint Draft ITU-T Rec. H.264—ISO/IEC 14496-10/Amd.3 scalable video coding, in: ISO/IEC JTCI/SC29/WG11 and ITU-T SG16 Q.6, JVT-X201, July 2007.
[16] W. Yao, Z. G. Li, S. Rahardja, Balanced inter-layer prediction for combined coarse granular scalability and spatial scalability, in: IEEE International Symposium on Circuits and Systems (ISCAS), May 2007.