Scalable Video Coding Extension of H.264/AVC

Wen-Hsiao Peng, Ph.D

Multimedia Architecture Processing Laboratory (MAPL)
Department of Computer Science, National Chiao Tung University

May 2008
Outline

- Introduction
- Scalable Video Coding Extension of H.264/AVC
  - Standardization Activities
  - Coding Technologies
  - Transport Interface
- R-D Performance
- Complexity Analysis
- Concluding Remarks
Introduction
Scalable Video Coding (SVC)

- **Embedded** Video Representation
- **Spatial**, **Temporal**, **SNR**, and **Combined** Scalability
Scalability

- QCIF
  - QCIF@75Hz,L
  - QCIF@15Hz,M
  - QCIF@30Hz,H

- CIF
  - CIF@15Hz,L
  - CIF@30Hz,H
Why Do We Need SVC?

- Diversified Clients
  - Different Computation Power and Display Capabilities
- Heterogeneous Networks
  - Different Types of Networks
  - Time-varying Channel Bandwidth
- Servers
  - Limited Storage Spaces - compression is required
  - Limited Bandwidth - unicast may not be applicable
Why Do We Need SVC?

- **Limited Resources**
  - Broadcasting
  - Server
  - Router
  - Wireless
  - Point-to-Point Transmission

- **Diversified Clients**
  - Ethernet
  - 384 kbps
  - 512 kbps

- **Time-Varying Bandwidth**
  - 3 Mbps
  - 1.5 Mbps
  - 384 kbps
  - 64 kbps
  - 512 kbps
  - 128 kbps
  - 64 kbps
  - 32 kbps
  - 256 kbps
  - 64 kbps

Wen-Hsiao Peng, Ph.D (NCTU CS)
Graceful Degradation

Foreman
How is SVC Useful?

Multiple Adaptation Using Application Layer Multicast!
SVC Standardization Activities
Roadmap

- MPEG-4 Part 10 Amd. 3
  - Started in MPEG-21 Part 13
  - Moved to Joint Video Team (JVT) since 2004/10

Roadmap

- 2003/03, Call for Evidence (CfE)
- 2004/04, Call for Proposal (CfP) - 12 Wavelet-based + 2 DCT-based
- 2004/10, Activities Moved to JVT (Palma Meeting)
- 2005/01, Wording Draft (WD) - HHI Proposal
- 2006/01, Proposed Draft of Amendment (PDAM)
- 2006/07, Final Proposed Draft of Amendment (FPDAM)
- 2007/01, Final Draft of Amendment (FDAM)
- 2007/07, Amendment (AMD)
SVC Call-for-Proposal

- AVC/H.264-based technology
  - HHI, NCTU

- Wavelet-based schemes
  - University of Brescia
  - Danae+ Thomson
  - Microsoft Research Asia (MSRA)
  - University of South Wales (UNSW)
AVC/H.264-based Technology

- Scalable Extension of AVC/H.264

![Flowchart Diagram]
Wavelet-based Technology

- **t+2D**

  - Temporal Wavelet
  - Spatial Wavelet
  - Scalable Entropy Coding

- **2D+t**

  - In-Band Temporal Wavelet
  - Spatial Wavelet
  - Scalable Entropy Coding
Subjective Quality Evaluation

- Single Stimulus Multimedia Test (SSMT)
  - Absence of unimpaired reference
  - 11 grade scale
Subjective Quality Comparison

<table>
<thead>
<tr>
<th>Bit Rate</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>96 kbps</td>
<td>Brescia</td>
</tr>
<tr>
<td>128 kbps</td>
<td>Danae + Thomson</td>
</tr>
<tr>
<td>192 kbps</td>
<td>HHI</td>
</tr>
<tr>
<td></td>
<td>MSRA</td>
</tr>
<tr>
<td></td>
<td>UNSW</td>
</tr>
</tbody>
</table>

Harbour - QCIF 15Hz

MOS

AVC/H.264-based

Wavelet-based
Summary

- **AVC/H.264-based Technologies**
  - Better Viewing Quality but Worse PSNR

- **Wavelet-based Schemes**
  - Worse Viewing Quality but Better PSNR

- 2004, AVC/H.264-based approach formally became starting point
SVC Technology Overview
SVC Technology Overview

- Temporal Scalability
  - Motion Compensated Temporal Filtering (MCTF)
  - Hierarchical B Pictures

- Spatial Scalability
  - Pyramid and Stack Structure
  - Separate MCTF/Motion Descriptions

- SNR Scalability
  - Coarse Granularity Scalability (CGS)
  - Medium Granularity Scalability (MGS)

- Adaptive Inter-layer Prediction
  - No need to decode multiple prediction loops
  - Comparable decoding complexity than single layer
SVC Encoder Block Diagram
SVC Syntax Structure

NAL unit header → Slice header → Slice data → Slice trail

Slice data:
- cabac_alignment
- mb_skip_run
- mb_skip_flag
- mb_field_decoding_flag → Macroblock layer → End of slice

Macroblock layer:
- base_mode_flag
- mb_type
- pcm sample
- Sub_mb_pred
- residual_pred_flag
- Mb_pred
- coded_block_pattern
- mb_qp_delta
- Residual

End of slice
Temporal Scalability
Dyadic Temporal Scalability

MCTF

Hierarchical B Prediction
Motion Compensated Temporal Filtering (MCTF)

- Wavelet Transform along Motion Trajectory

- Prediction Step, \( P_k = \left[ -\frac{1}{2}, 1, \frac{1}{2} \right] \)

\[
\hat{h}_k = \underbrace{S_{2k+1}}_{\text{HighPassOutputs}} - \underbrace{P_k \otimes S_{2k}}_{\text{EvenFrameFiltering}}
\]

- Update Step, \( U_k = \left[ \frac{3}{4}, 1, \frac{3}{4} \right] \)

\[
\hat{l}_k = \underbrace{S_{2k}}_{\text{LowPassOutputs}} + \underbrace{U_k \otimes h_k}_{\text{HighPassFrameFiltering}}
\]
MCTF Example
Hierarchical B Predictions

- Hierarchical B Predictions
  - Remove Update Step + Close-loop Prediction
  - Enabled by H.264/AVC Syntax

- Hierarchical B vs. P (Low Delay)
Efficiency and Mismatch Trade-off

- **Closed Loop (Hierarchical B)**
  - Use reconstructed frames for prediction
  - **No Mismatch** but **Worse** Prediction Efficiency

- **Open Loop (MCTF)**
  - Employ original pictures for prediction
  - **Mismatch** with **Better** Prediction Efficiency

![Diagram of closed and open loop encoding processes]
Encoder and Decoder Mismatch

- Drifting Errors

![Diagram showing encoder and decoder mismatch with predictor mismatch highlighted]
Loop Control Performance

- Closed-loop is better in most cases
Spatial Scalability
Spatial Scalability

- Stack Structure with Adaptive Inter-layer Predictions
  - Residual prediction performed in **Spatial** domain
SNR Scalability
SNR Scalability

- Coarse Granularity Scalability (CGS)
  - Distinctive Quality Levels
  - Stack Structure of Spatial Scalability
- Medium Granularity Scalability (MGS)
  - Packet-based Quality Scalable Coding
  - Key Pictures
Coarse Granularity Scalability

- Stack Structure (Similar to Spatial Scalability)
- Residual prediction performed in Transform domain
Medium Granularity Scalability

- Packet-based Quality Scalable Coding
  - Distribute Transform Coefficients among Several Slices

- Key Picture Concepts
  - Trade-off between Drift and Coding Efficiency
  - Resynchronization Points between Encoder and Decoder
Drift Control

- BL-only, EL-only, Two-Loop, and Key Pictures

(A) BL-only

(B) EL-only

(C) Two-Loop

(D) Key Pictures
Viewing Quality Comparison

- Two Loops
- Key Pictures

TwoLoops

KeyPictures
Performance Comparison of Prediction Structures

- BL-only, EL-only, Two-Loop, and Key Pictures

![Graph showing PSNR vs. Bit-rate for different prediction structures.]
Performance Comparison of SNR Scalability

- CGS performance decreases with increasing number of layers

![Graph showing SNR scalability comparison]

- Crew, CIF 15Hz, GOP16
Adaptive Inter-layer Prediction
Adaptive Inter-Layer Prediction

- Motion Prediction
  - Macroblock Type, Reference Index, Motion Vector
- Residual Prediction
  - Subtraction of Base Layer from Enhancement Layer
- Textural Prediction
  - Inter-layer Intra Prediction
- Remarks
  - No Base Layer Reconstruction
  - Approximately 0-0.5dB Coding Loss
Multi-level Inter-layer Prediction

GOP Boundary

Spatial Enh. Layer 2 with MCTF (Frame Rate = FR)

Spatial Enh. Layer 1 with MCTF (Frame Rate = FR)

Base Layer with Hierarchical B Pictures (Frame Rate = FR/2)
Inter-layer Motion Prediction

- Macroblock Type, Reference Index, Motion Vector

```
<table>
<thead>
<tr>
<th>base_mode_flag</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>co-located 8x8 block</td>
<td>Intra-coded</td>
<td>Inter-coded</td>
</tr>
<tr>
<td>inter-layer intra-prediction</td>
<td>derive MB partition, mv, ref index</td>
<td></td>
</tr>
<tr>
<td>mb_type</td>
<td>motion_prediction_flag</td>
<td>1</td>
</tr>
<tr>
<td>derive mvp, ref index mvd</td>
<td>mv, ref index</td>
<td></td>
</tr>
</tbody>
</table>

residual signal
```
Macroblock Partition Derivation

Partition Derivation across Spatial Resolutions

Spatial Enhancement Layer

Subordinate Layer

Direct, 16x16, 16x8, 8x16

Intra
Inter-layer Residual and Textural Prediction

- **Textural Prediction - Intra-coded Macroblocks**
  - De-blocking Filtering before Spatial Interpolation
  - Spatial Interpolation across Submacroblocks (8x8)

- **Residual Prediction - Inter-coded Macroblocks**
  - Independent from Motion Prediction
  - No Spatial Interpolation across Transform Blocks
Gain of Textural Prediction

- Simulcast vs. SVC+Textural

Spatial Scalability: Foreman, 150 pics, QCIF 15Hz @ 48 kbit/s -> CIF 15 Hz
Gain of Motion Prediction

- SVC+Textural vs. SVC+Textural/Motion
Gain of Residual Prediction

- SVC+Textural/Motion vs. SVC+Textural/Motion/Residual

Spatial Scalability: Foreman, 150 pics, QCIF 15Hz @ 48 kbit/s -> CIF 15 Hz

- QCIF Base Layer
- CIF Single Layer
- QCIF & CIF Simulcast
- Bit rate used for Residuals
- Achieved RD curve for Residual prediction
- Target RD curve for Residual prediction
Combined Scalability
Combined Scalability

Layer 0: QCIF 15 Hz
SNR base 41 kbit/s
QCIF, 15 Hz 41 - 80 kbit/s
QCIF, 7.5 Hz 32 - 66 kbit/s
QCIF, 3.75 Hz 24 - 55 kbit/s
QCIF, 1.875 Hz 18 - 48 kbit/s

Layer 1: CIF 30 Hz
SNR base 115 kbit/s
CIF, 30 Hz, 115 - 256 kbit/s
CIF, 15 Hz, 88 - 222 kbit/s
CIF, 7.5 Hz, 65 - 189 kbit/s
CIF, 3.75 Hz, 55 - 165 kbit/s
CIF, 1.875 Hz, 48 - 139 kbit/s

Spatial upsampling
FGS refinement
Prediction

FGS 41 - 80 kbit/s
Layer 0: QCIF 15 Hz
SNR base 41 kbit/s

FGS 115 - 256 kbit/s
Layer 1: CIF 30 Hz
SNR base 115 kbit/s
Encoder Optimization
**Bottm-up Encoder Control**

- **Current JSVM**
  - Bottom-up Encoding Process
- **Sequential Mode Selection**
  - **Base Layer**
    
    \[ p_B^* = \arg \min_{\{p_B\}} [D_B(p_B) + \lambda_B R_B(p_B)] \]
  - **Enhancement Layer**
    
    \[ p_E^* = \arg \min_{\{p_E|p_B^*\}} [D_E(p_E|p_B^*) + \lambda_E R_E(p_E|p_B^*)] \]

- **Remarks**
  - Base layer is comparable to H.264/AVC
  - Enhancement layer is much worse than H.264/AVC
Multi-loop Encoder Control

- Joint Base and Enhancement Layer Optimization
  - Base Layer Encoding

\[
\mathbf{p}_B^* = \arg \min_{\{\mathbf{p}_E | \mathbf{p}_B, \mathbf{p}_B\}} \left( \omega \times [D_B(\mathbf{p}_B) + \lambda_B R_B(\mathbf{p}_B)] + (1 - \omega) \times [D_E(\mathbf{p}_E) + \lambda_E R_E(\mathbf{p}_E)] \right)
\]

- Enhancement Layer Encoding (not modified)

\[
\mathbf{p}_E^* = \arg \min_{\{\mathbf{p}_E | \mathbf{p}_B^*\}} \left[D_E(\mathbf{p}_E | \mathbf{p}_B^*) + \lambda_E R_E(\mathbf{p}_E | \mathbf{p}_B^*) \right]
\]

- Remarks
  - Trade-off between Base and Enhancement Coding Efficiency
Multi-loop Encoder Control

- 10% bit rate increase relative to H.264/AVC
SVC Transport Interface
**SVC NAL HEADER**

- **Traditional NAL Header**
- **IDR Flag**
- **(D, T, Q) Information**
- **Discardable Flag**
- **Priority ID of a NAL**
- **No Inter-Layer Prediction Flag**
- **Use Base Reference**

![Diagram of SVC NAL Header](image-url)
SVC NAL Grouping

- Layer Representation, Dependency Layer, Scalable Layer, Scalable Layer Representation, Access Unit
SVC Inter-layer Dependency Hierarchy

(Layer Type, Dependency_id, Quality_id)

Spatial layer 0
- CGS_2_0
  - CGS_1_0
  - BASE_0_0

Spatial layer 1
- MGS_3_3
  - MGS_3_2
  - MGS_3_1
  - BASE_3_0

Spatial layer 2
- MGS_4_2
  - MGS_4_1
  - BASE_4_0 
Scalability Information SEI

- Number of Scalable Layers
- Information of Each Scalable Layer
  - Layer Identifier + Decoding Dependency Information
  - \( \text{layer}_id = (\text{Dependency}_id \times 16 + \text{Quality}_id) \times 8 + \text{Temporal}_id \)
  - Bit Rate, Frame Size, Frame Rate
  - Initial Parameter Sets
  - Region-of-interest (ROI) Information
- Priority Information
Layer Switching

- Switching between Quality Refinement Layers
  - Possible in each access unit
- Switching between Dependency Layers
  - Possible at IDR access units
- Down-switching
  - Possible in virtually any access unit
  - Require multiple-loop decoding
- Up-switching
  - Wait for next IDR access unit
R-D Optimized Layer Switching

- What to produce and what to use?
  - QCIF@30Hz or CIF@10Hz?
  - Spatial quality vs. Temporal quality
H.264/AVC vs. SVC
R-D Performance

![Graph showing the performance comparison between AVC/H.264 and SVC at different bitrates. The graph plots PSNR (dB) against bitrate (kbps). The graph includes data points for AVC/H.264 @CIF, SVC @CIF, and AVC/H.264 @4CIF.](image-url)

**Axes:**
- Y-axis: PSNR (dB)
- X-axis: Bitrate (kbps)

**Data Points:**
- **AVC/H.264 @CIF**
- **SVC @CIF**
- **AVC/H.264 @4CIF**
- **SVC @4CIF**
SVC Complexity Analysis
How Complex is SVC?

- Complexity of SVC depends on \# of prediction loops
- Encoder is slower than real-time by
  - 60x for temporal scalability in CIF resolution
  - 146x for spatial scalability
  - 27x for CGS, 75x for FGS in CIF resolution
  - 1000x for combined scalability
- Decoder is slower than real-time by
  - 16x for spatial scalability
  - 1.7x for CGS, 5x for FGS in QCIF resolution
  - 92x for combined scalability
- Decoder is faster than real-time for temporal scalability in QCIF
- Ratio of memory access over computation is $\geq 200$
Concluding Remarks
Concluding Remarks

- Scalable Video Coding (SVC)
  - Spatial, Temporal, SNR, and Combined Scalability
  - Power and Format Adaptation with Graceful Degradation

- What is Next?
  - Encoder Optimization
  - Decoder Implementation
  - Transport Mechanism
  - Bit-depth and Color Scalability
  - Interlaced Videos
References


