5. Scalable Multimedia Communications
Contents

- Scalable Coding and Rate Control
- Scalable Coding Techniques
- Fine Granularity Scalability in MPEG-4
- Temporally Subband Scalable Coding
References

A. Scalable Coding and Rate Control
Streaming systems need to provide service to clients with different constraints.
When encoder encodes a multimedia stream, it has no idea to what client this stream will be sent to.

What is the rate when encoding this file?
Even for a single client, due to the fluctuation of Internet bandwidth, server needs to adjust the data rate adaptively.

It is not possible for the encoder to encode the multimedia data once for each rate:
- Due to storage and computation time.

We need scalable coding such that a single stream can suit to all clients and network conditions.

Streams of different rates can be extracted from a single stream when required.

While video requires the most computation time and storage in a multimedia stream, scalable video coding techniques are particularly demanded.
B. Scalable Coding Techniques
Scalable video coding techniques have been included in MPEG-2 standard

Introduce an enhancement layer to achieve a two-layer scalable video coding

If low bandwidth, send only the base layer data. Or otherwise, send both layer data
3 major kinds of scalable video coding technique

- **Temporal Scalability Coding**
  - Allow video of different frame rates to be extracted from a single stream
  - e.g. provide video stream that can be displayed at 25 fps or 10 fps

- **Resolution Scalability Coding**
  - Allow video of different resolutions to be extracted from a single stream
  - e.g. provide video stream that can be displayed at CIF (352x288) or QCIF (176x144)

- **SNR Scalability Coding**
  - Allow video of different signal-to-noise ratios to be extracted from a single stream
I. SNR Scalability

- **Signal-to-noise ratio (SNR) scalability** is a technique to code a video into two layers at different quantization accuracy.
Decoder simply does the reverse

With the data from base layer and enhancement layer, the resulted DCT coefficients have the accuracy as if all of them are quantized using $Q_2$. 

$Q_1 > Q_2$
II. Resolution Scalability

- **Resolution (spatial) Scalability** is a technique to code a video into two layers at different spatial resolutions.
- Resolution of an image can be changed by **downsampling** and **upsampling**.

![Downsampling and Upsampling Diagram]

Although retaining the size, information is loss.
- Encoder architecture:

```
Image or error residue
      ▼
      |     |
  ▼     ▼
Downsample ▼ DCT/Q₁/VLC ▼ Base Layer
  ▼     ▼
    |     |
Upsample  ▼ VLD/Q⁻¹/DCT⁻¹ ▼ Enhancement Layer
  ▼     ▼
      |     |
DCT/Q₁/VLC ▼ ▼
```

14
Decoder architecture:

Base Layer

VLD /Q₁⁻¹/DCT⁻¹

Original image or error residue (with loss due to quantization)

Enhancement Layer

VLD /Q₁⁻¹/DCT⁻¹

Upsample

Base layer image or error residue (optional)
III. Temporal Scalability

- For each group-of-picture (GOP) of MPEG-2, encoded video frames are divided into 3 different categories:
  - I: INTRA
  - B: Bi-directional
  - P: INTER
For temporal scalability, the enhancement layer introduces a layer of P and B frames to increase the frame rate.

Enhancement layer frames + Base layer frames = doubled frame rate
C. Fine Granularity Scalability in MPEG-4
A major **deficiency** of MPEG-2 scalable coding is it only **provides two layers**

Enhancement layer gives enhancement only once, but not further even if the constraint of client is further reduced.

---

**Rate-distortion curve of the video source**

- **Non-scalable**
- **2 layer-scalable**
To achieve multi-level scalable coding, encoder needs to give multi-level embedded bitstream.

- Embedded bitstream over a wide band channel
- Base layer + part of layer 2

Media Server

Clients
- Idea of **Fine Granularity Scalability (FGS)** is proposed in the **Amendment of MPEG-4**
- **Major difference** from layer scalable coding is
  - although FGS still divides the code into two layers, enhancement bitstream can be truncated into any number of bits within each frame
- Can achieve **partial enhancement proportional to the number of bits decoded for each frame**
- Can give **continuous scalability curve theoretically**
- One of the approaches to achieve FGS is by using **bit-plane coding**
Codec using Bit-plane Coding

- $Q_1$ and $Q_2$ are set to give the minimum and maximum, respectively, information required for presenting the multimedia data
- Bit-plane coding enables continuous enhancement over the base layer bitstream
Bit-plane coding considers each quantized DCT coefficient as a binary number of bits.

Coefficients are coded bit-plane by bit-plane rather than coefficient-by-coefficient.

For example:

```
10 0 6 0 0 3 0 2 2 0 0 2 0 0 1 0 .. 0 0 (absolute)
0 x 1 x x 1 x 0 0 x x 1 x x 0 x .. x x (sign bits)
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 .. 0 0 (MSB)
0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 .. 0 0 (MSB-1)
1 0 1 0 0 1 0 1 1 0 0 1 0 0 0 0 .. 0 0 (MSB-2)
0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 .. 0 0 (MSB-3)
```
Converting the four bit-planes into (RUN,EOP) symbols, we have

<table>
<thead>
<tr>
<th>Bit Plane</th>
<th>Symbols</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSB-3</td>
<td>0000000000000000 .. 00 (MSB)</td>
</tr>
<tr>
<td>0,1</td>
<td>RUN, EOP of MSB</td>
</tr>
<tr>
<td>MSB-2</td>
<td>0000000000000000 .. 00 (MSB-1)</td>
</tr>
<tr>
<td>2,1</td>
<td>(RUN,EOP) of MSB-1</td>
</tr>
<tr>
<td>MSB-1</td>
<td>1010110010010010 .. 00 (MSB-2)</td>
</tr>
<tr>
<td>0,0,1,0,2,0,1,0</td>
<td>(RUN,EOP) of MSB-2</td>
</tr>
<tr>
<td>MSB</td>
<td>0000000000000001 .. 00 (MSB-3)</td>
</tr>
<tr>
<td>5,0,8,1</td>
<td>(RUN,EOP) of MSB-3</td>
</tr>
</tbody>
</table>
Each sign bit is put into the bitstream only once right after the VLC code that contains the MSB of the nonzero absolute value associated with the sign bit.

<table>
<thead>
<tr>
<th>Sign Bits</th>
<th>MSB</th>
<th>MSB-1</th>
<th>MSB-2</th>
<th>MSB-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 x 1 x x 1 x 0 0 x x 1 x x 0 x .. x x</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 .. 0 0</td>
<td>0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 .. 0 0</td>
<td>0 1 0 0 0 1 0 1 1 0 0 1 0 0 0 0 .. 0 0</td>
<td>0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 .. 0 0</td>
</tr>
</tbody>
</table>

(RUN, EOP) of MSB, (RUN, EOP) of MSB-1, (RUN, EOP) of MSB-2, (RUN, EOP) of MSB-3
● Up to 20% of bit saving can be achieved by using bit-plane coding as compared with run-length coding

● Because bit-plane statistics are independent of the Q value

● For run-length coding, it is impossible for a single VLC table to be optimal for all Q values (total 31)

● Nevertheless, statistics of the first three bit-planes (MSB, MSB-1 and MSB-2) are very different from each other and others lower bit-planes

● Four different VLC tables are designed for MSB, MSB-1, MSB-2 and the rest
● When constructing the bitstream data from more significant bit-planes are placed first

● For example: \((0,1), 0, (2,1), 1, (0,0), (1,0), (2,0), 1, (1,0), 0, (0,0), 0, (2,1), 1, (5,0), (8,1), 0\)

● Transmission can stop at any point of the bitstream depending on the channel bandwidth

● For example

\[(0,1), 0, (2,1), 1, (0,0), (1,0)\] Extract part of the bitstream

Decode result

| 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ..., 0, 0 | (MSB) |
| 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ..., 0, 0 | (MSB-1) |
| 1, 0, 1 | (MSB-2) |
| 0, x, 1, x, x, x, x, x, x, x, x, x, x, x, x, x, x, ..., x, x | |
| 10, 0, 6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ..., 0, 0 | (Decoded) |
| 10, 0, 6, 0, 0, 3, 0, 2, 2, 0, 0, 2, 0, 0, 1, 0, ..., 0, 0 | (Original) |
D. Temporal Subband Scalable Coding
Intuitively, temporally scalability can be implemented with less effort than SNR and Resolution Scalable coding.

Due to the predictive nature, we cannot arbitrarily delete frames to achieve a specific frame rate.

**Full frame rate**

![Diagram of Group of pictures (GOF)](image)
Temporal prediction techniques such as telescoping and recursive prediction may apply.
• Recursive prediction approach to achieve 1/4 and 1/8 frame rate:
Great computational effort is required to generate the required motion vectors and residue error.

It has been demonstrated, particularly for error prone channel, predictive coding may not be good.

By sacrificing predictive coding, temporal scalability can be easily implemented.

Without prediction, any frame can be removed to achieve a certain frame rate.
- However, low compression efficiency may be resulted
- An intermediate approach ⇒ Temporal Subband Coding
Haar Subband Filtering Scheme

FIR filters

Moving average

Moving difference
Example

Input: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 0

Haar low pass:  0, 1, 3, 5, 7, 9, 11, 13, 15, 17, 9
Haar high pass: 0, -1, -1, -1, -1, -1, -1, -1, -1, -1, 9

Low pass downsampling: 0, 3, 7, 11, 15, 9
High pass downsampling: 0, -1, -1, -1, -1, 9

Low pass upsampling: 0, 0, 3, 0, 7, 0, 11, 0, 15, 0, 9, 0
High pass upsampling: 0, 0, -1, 0, -1, 0, -1, 0, -1, 0, 9, 0

Low pass reconstruct: 0, 0, 3, 3, 7, 7, 11, 11, 15, 15, 9, 9
High pass reconstruct: 0, 0, -1, 1, -1, 1, -1, 1, -1, 1, 9, -9

Output: 0, 0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 0
- The perfect reconstruction is not accidental. Can be formally proved.
- By discrete Fourier series,

\[ Y'(z) = 0.5(Y(z) + Y(-z)) \]
\[
Y(z) = 0.5\left(\tilde{L}(z)L(z) + \tilde{H}(z)H(z)\right)X(z) + \\
0.5(\tilde{L}(-z)L(z) + \tilde{H}(-z)H(z))X(-z)
\]  
(Eqn a)
Haar filters

\[ \mathcal{L}(z) = 1 + z^{-1}; \quad \mathcal{L}(-z) = 1 - z^{-1} \]

\[ \mathcal{H}(z) = -1 + z^{-1}; \quad \mathcal{H}(-z) = -1 - z^{-1} \]

\[ L(z) = 1 + z^{-1}; \quad L(-z) = 1 - z^{-1} \]

\[ H(z) = 1 - z^{-1}; \quad H(-z) = 1 + z^{-1} \]
Exercise I

- By substituting the z-domain representation of the haar filters to eqn.a, show that

\[ Y(z) = 2z^{-1}X(z) \]

That is, the output \( Y(z) \) is the scaled up and delayed version of the input \( X(z) \)
A set of subband filters

Moving average of video frame

Base layer bitstream

Moving difference of video frame

Enhancement layer bitstream (Transmit when bandwidth allow)
By means of **downsampling**, video frames with reduced frame rate is generated.

The high pass branch gives a **moving difference** between frames.

It is seen that if the difference is not great, the data at the high pass branch will be small. Hence suitable for subsequent spatial coding.

However if motion is large, the data at the high pass branch is large.

Hence **temporal redundancy is partially reduced.**
• Multi-level temporal scalability can be achieved by cascading a few stages of subband filters
The major problem of the temporal subband coding scheme is at the low pass branches. It in general gives moving averages of video frame. In case of high motion video, blurred output will be produced.
To solve this problem, **interpolative wavelet transform** is used.

Proposed by Donoho in 90’s.

One of the possible structures **allows all pass filtering at its low pass branch**.

Hence the **low pass branch gives exactly the original frames with lowered frame rate**.

However, the filters need to be carefully designed.
One of the possible choices is to use the Deslauriers-Dubuc filters with 3 tags

\[
\tilde{L}(z) = 1 \\
\tilde{H}(z) = -0.5z^2 + z - 0.5 \\
H(z) = z^{-1} \\
L(z) = 0.5z + 1 + 0.5z^{-1}
\]
Exercise II

- By substituting the z-domain representation of the Deslauriers-Dubuc filters to eqn.a, show that

\[ Y(z) = X(z) \]