Video Compression
Conferencing & Internet Video

Hamid R. Rabiee, PhD
Intel Corporation &
Oregon Graduate Institute
Objectives

The student should be able to:

☑️ Describe the basic components of the H.263 video codec and how it differs from H.261.
☑️ Describe and understand the improvements of H.263+ over H.263.
☑️ Understand enough about Internet and WWW protocols to see how they affect video.
☑️ Understand the basics of streaming video over the Internet as well as error resiliency and concealment techniques.
Outline

Section 1: Conferencing Video
Section 2: Internet Review
Section 3: Internet Video
Section 1: Conferencing Video

- Video Compression Review
- Chronology of Video Standards
- The Input Video Format
- H.263 Overview
- H.263+ Overview
Video Compression Review
Garden Variety Video Coder

Video codecs have three main functional blocks: Motion Estimation & Compensation, Transform, Quantization, Zig-Zag Scan & Run-Length Encoding, Symbol Encoder. The process starts with Frames of Digital Video and ends with a Bit Stream.
The symbol encoder exploits the statistical properties of its input by using shorter code words for more common symbols.

Examples: Huffman & Arithmetic Coding
This block is the basis for most lossless image coders (in conjunction with DPCM, etc.)
A transform (usually DCT) is applied to the input data for better energy compaction which decreases the entropy and improves the performance of the symbol encoder.
Transform & Quantization

The DCT also decomposes the input into its frequency components so that perceptual properties can be exploited. For example, we can throw away high frequency content first.
Quantization lets us reduce the representation size of each symbol, improving compression but at the expense of added errors. It’s the main tuning knob for controlling data rate.
Zig-zag scanning and run-length encoding orders the data into 1-D arrays and replaces long runs of zeros with run-length symbols.
Still Image Compression

These two components form the basis for many still image compression algorithms such as JPEG, PhotoCD, M-JPEG and DV.
Finally, because video is a sequence of pictures with high temporal correlation, we add motion estimation/compensation to try to predict as much of the current frame as possible from the previous frame.
Motion Estimation/Compensation

Most common method is to predict each block in the current frame by a (possibly translated) block of the previous frame.
These three components form the basis for most of the standard video compression algorithms: MPEG-1, -2, & -4, H.261, H.263, H.263+. 
Section 1: Conferencing Video

- Video Compression Review
- Chronology of Video Standards
- The Input Video Format
- H.263 Overview
- H.263+ Overview
Chronology of Video Standards


H.261  H.263  H.263++  H.263+  H.263L

MPEG 1  MPEG 2  MPEG 4  MPEG 7

ITU-T  ISO

ISO/ITU-T
Chronology of Video Standards

- **(1990) H.261, ITU-T**
  - Designed to work at multiples of 64 kb/s (px64).
  - Operates on standard frame sizes CIF, QCIF.
- **(1992) MPEG-1, ISO “Storage & Retrieval of Audio & Video”**
  - Evolution of H.261.
  - Main application is CD-ROM based video (~1.5 Mb/s).
Chronology continued

- (1994-5) MPEG-2, ISO “Digital Television”
  - Evolution of MPEG-1.
  - Main application is video broadcast (DirecTV, DVD, HDTV).
  - Typically operates at data rates of 2-3 Mb/s and above.
Chronology continued

• (1996) H.263, ITU-T
  – Evolution of all of the above.
  – Supports more standard frame sizes (SQCIF, QCIF, CIF, 4CIF, 16CIF).
  – Targeted low bit rate video <64 kb/s. Works well at high rates, too.

• (1/98) H.263 Ver. 2 (H.263+), ITU-T
  – Additional negotiable options for H.263.
  – New features include: deblocking filter, scalability, slicing for network packetization and local decode, square pixel support, arbitrary frame size, chromakey transparency, etc…
 Chronology continued

• (1/99) MPEG-4, ISO “Multimedia Applications”
  – MPEG4 video based on H.263, similar to H.263+
  – Adds more sophisticated binary and multi-bit transparency support.
  – Support for multi-layered, non-rectangular video display.

• (2H/’00) H.263++ (H.263V3), ITU-T
  – Tentative work item.
  – Addition of features to H.263.
  – Maintain backward compatibility with H.263 V.1.
Chronology continued

• (2001) MPEG7, ISO “Content Representation for Info Search”
  – Specify a standardized description of various types of multimedia information. This description shall be associated with the content itself, to allow fast and efficient searching for material that is of a user’s interest.

• (2002) H.263L, ITU-T
  – Call for Proposals, early ‘98.
  – Proposals reviewed through 11/98, decision to proceed.
  – Determined in 2001
Section 1: Conferencing Video

- Video Compression Review
- Chronology of Video Standards
- The Input Video Format
- H.263 Overview
- H.263+ Overview
Video Format for Conferencing

- Input color format is YCbCr (a.k.a. YUV). Y is the luminance component, U & V are chrominance (color difference) components.
- Chrominance is subsampled by two in each direction.
- Input frame size is based on the Common Intermediate Format (CIF) which is 352x288 pixels for luminance and 176x144 for each of the chrominance components.
**YCbCr (YUV) Color Space**

- Defined as input color space to H.263, H.263+, H.261, MPEG, etc.
- It’s a 3x3 transformation from RGB.

\[
\begin{bmatrix}
Y \\
Cb \\
Cr
\end{bmatrix} = \begin{bmatrix}
0.299 & 0.587 & 0.114 \\
-0.169 & -0.331 & 0.500 \\
0.500 & -0.419 & -0.081
\end{bmatrix}\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\]

Y represents the luminance of a pixel.
Cr, Cb represents the color difference or chrominance of a pixel.
Subsampled Chrominance

- The human eye is more sensitive to spatial detail in luminance than in chrominance.
- Hence, it doesn’t make sense to have as many pixels in the chrominance planes.
Spatial relation between luma and chroma pels for CIF 4:2:0

Different than MPEG-2 4:2:0

- luminance pel
- chrominance pel
- block edge
Common Intermediate Format

• The input video format is based on Common Intermediate Format or CIF.

• It is called Common Intermediate Format because it is derivable from both 525 line/60 Hz (NTSC) and 625 line/50 Hz (PAL) video signals.

• CIF is defined as 352 pels per line and 288 lines per frame.

• The picture area for CIF is defined to have an aspect ratio of about 4:3. However,

\[
\frac{352}{4 \times 3} = 264 \neq 288
\]
Picture & Pixel Aspect Ratios

Pixels are not square in CIF.

Picture
4:3

Pixel
12:11
Hence on a square pixel display such as a computer screen, the video will look slightly compressed horizontally. The solution is to spatially resample the video frames to be

**384 x 288 or 352 x 264**

This corresponds to a 4:3 aspect ratio for the picture area on a square pixel display.
The luma and chroma planes are divided into 8x8 pixel blocks. Every four luma blocks are associated with a corresponding Cb and Cr block to create a macroblock.
Section 1: Conferencing Video

- Video Compression Review
- Chronology of Video Standards
- The Input Video Format
- H.263 Overview
- H.263+ Overview
ITU-T Recommendation
H.263
ITU-T Recommendation

H.263

• H.263 targets low data rates (< 28 kb/s). For example it can compress QCIF video to 10-15 fps at 20 kb/s.
• For the first time there is a standard video codec that can be used for video conferencing over normal phone lines (H.324).
• H.263 is also used in ISDN-based VC (H.320) and network/Internet VC (H.323).
ITU-T Recommendation H.263

Composed of a baseline plus four negotiable options

- Baseline Codec
- Unrestricted/Extended Motion Vector Mode
- Advanced Prediction Mode
- PB Frames Mode
- Syntax-based Arithmetic Coding Mode
Frame Formats

<table>
<thead>
<tr>
<th>Format</th>
<th>Y</th>
<th>U,V</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQCIF</td>
<td>128x96</td>
<td>64x48</td>
</tr>
<tr>
<td>QCIF</td>
<td>176x144</td>
<td>88x72</td>
</tr>
<tr>
<td>CIF</td>
<td>352x288</td>
<td>176x144</td>
</tr>
<tr>
<td>4CIF</td>
<td>704x576</td>
<td>352x288</td>
</tr>
<tr>
<td>16CIF</td>
<td>1408x1152</td>
<td>704x576</td>
</tr>
</tbody>
</table>

Always 12:11 pixel aspect ratio.
Picture & Macroblock Types

- Two picture types:
  - INTRA (I-frame) implies no temporal prediction is performed.
  - INTER (P-frame) may employ temporal prediction.

- Macroblock (MB) types:
  - INTRA & INTER MB types (even in P-frames).
    - INTER MBs have shorter symbols in P frames
    - INTRA MBs have shorter symbols in I frames
  - Not coded - MB data is copied from previous decoded frame.
Motion Vectors

- Motion vectors have 1/2 pixel granularity. Reference frames must be interpolated by two.
- MV’s are not coded directly, but rather a median predictor is used.
- The predictor residual is then coded using a VLC table.

\[
MV_{\Delta} = MV_X - \text{median}(MV_A, MV_B, MV_C)
\]
Motion Vector Delta (MVD) Symbol Lengths

Code length in bits

MVD Absolute Value

0 0.5 1 1.5 2 2.5 - 3.5 4.0 - 5.0 5.5 - 12.0 12.5 - 15.5
Assign a variable length code according to three parameters (3-D VLC):

1. Length of the run of zeros preceding the current nonzero coefficient.
2. Amplitude of the current coefficient.
3. Indication of whether current coefficient is the last one in the block.

3. The most common are variable length coded (3-13 bits), the rest are coded with escape sequences (22 bits).
Quantization

- H.263 uses a scalar quantizer with center clipping.
- Quantizer varies from 2 to 62, by 2’s.
- Can be varied ±1, ±2 at macroblock boundaries (2 bits), or 2-62 at row and picture boundaries (5 bits).
Bit Stream Syntax

Hierarchy of three layers.

- Picture Layer
- GOB* Layer
- MB Layer

* A GOB is usually a row of macroblocks, except for frame sizes greater than CIF.
Picture Layer Concepts

<table>
<thead>
<tr>
<th>Picture Start Code</th>
<th>Temporal Reference</th>
<th>Picture Type</th>
<th>Picture Quant</th>
</tr>
</thead>
</table>

- **PSC** - sequence of bits that can not be emulated anywhere else in the bit stream.
- **TR** - 29.97 Hz counter indicating time reference for a picture.
- **PType** - Denotes INTRA, INTER-coded, etc.
- **P-Quant** - Indicates which quantizer (2…62) is used initially for the picture.
GOB Layer Concepts

**GOB Headers are Optional**

<table>
<thead>
<tr>
<th>GOB Start Code</th>
<th>GOB Number</th>
<th>GOB Quant</th>
</tr>
</thead>
</table>

- GSC - Another unique start code (17 bits).
- GOB Number - Indicates which GOB, counting vertically from the top (5 bits).
- GOB Quant - Indicates which quantizer (2…62) is used for this GOB (5 bits).

GOB can be decoded independently from the rest of the frame.
Macroblock Layer Concepts

<table>
<thead>
<tr>
<th>Coded Flag</th>
<th>MB Type</th>
<th>Code Block Pattern</th>
<th>DQuant</th>
<th>MV Deltas</th>
<th>Transform Coefficients</th>
</tr>
</thead>
</table>

- COD - if set, indicates empty INTER MB.
- MB Type - indicates INTER, INTRA, whether MV is present, etc.
- CBP - indicates which blocks, if any, are empty.
- DQuant - indicates a quantizer change by +/- 2, 4.
- MV Deltas - are the MV prediction residuals.
- Transform coefficients - are the 3-D VLC’s for the coefficients.
Unrestricted/Extended Motion Vector Mode

- Motion vectors are permitted to point outside the picture boundaries.
  - non-existent pixels are created by replicating the edge pixels.
  - improves compression when there is movement across the edge of a picture boundary or when there is camera panning.
- Also possible to extend the range of the motion vectors from [-16,15.5] to [-31.5,31.5] with some restrictions. This better addresses high motion scenes.
Motion Vectors Over Picture Boundaries

Reference Frame N-1

Target Frame N

Edge pixels are repeated.
Extended MV Range

Extended motion vector range, [-16,15.5] around MV predictor.

Base motion vector range.
Advanced Prediction Mode

• Includes motion vectors across picture boundaries from the previous mode.

• Option of using four motion vectors for 8x8 blocks instead of one motion vector for 16x16 blocks as in baseline.

• Overlapped motion compensation to reduce blocking artifacts.
Overlapped Motion Compensation

- In normal motion compensation, the current block is composed of
  - the predicted block from the previous frame (referenced by the motion vectors), *plus*
  - the residual data transmitted in the bit stream for the current block.

- In overlapped motion compensation, the prediction is a weighted sum of *three* predictions.
Overlapped Motion Compensation

• Let \((m, n)\) be the column & row indices of an 8×8 pixel block in a frame.
• Let \((i, j)\) be the column & row indices of a pixel within an 8×8 block.
• Let \((x, y)\) be the column & row indices of a pixel within the entire frame so that:

\[(x, y) = (m \times 8 + i, n \times 8 + j)\]
Overlapped Motion Comp.

- Let \((MV^0_x, MV^0_y)\) denote the motion vectors for the current block.
- Let \((MV^1_x, MV^1_y)\) denote the motion vectors for the block above (below) if the current pixel is in the top (bottom) half of the current block.
- Let \((MV^2_x, MV^2_y)\) denote the motion vectors for the block to the left (right) if the current pixel is in the left (right) half of the current block.
Overlapped Motion Comp.

Then the summed, weighted prediction is denoted:

\[ P(x, y) = \frac{(q(x, y) H_0(i, j) + r(x, y) H_1(i, j) + s(x, y) H_2(i, j) + 4)}{8} \]

Where,

\[ q(x, y) = (x + M V^0_x, y + M V^0_y), \]
\[ r(x, y) = (x + M V^1_x, y + M V^1_y), \]
\[ s(x, y) = (x + M V^2_x, y + M V^2_y) \]
$$H^0(i, j) =$$

<table>
<thead>
<tr>
<th></th>
<th>4</th>
<th>5</th>
<th>5</th>
<th>5</th>
<th>5</th>
<th>5</th>
<th>5</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>
Overlapped Motion Comp.

\[
H^1(i, j) =
\]

\[
H^2(i, j) = (H^1(i, j))^T
\]
PB Frames Mode

• Permits two pictures to be coded as one unit: a P frame as in baseline, and a bi-directionally predicted frame or B frame.

• B frames provide more efficient compression at times.

• Can increase frame rate 2X with only about 30% increase in bit rate.

• Restriction: the backward predictor cannot extend outside the current MB position of the future frame. See diagram.
PB Frames

Picture 1
P or I Frame

Picture 2
B Frame

Picture 3
P or I Frame

PB

2X frame rate for only 30% more bits.
Syntax based Arithmetic Coding Mode

- In this mode, all the variable length coding and decoding of baseline H.263 is replaced with arithmetic coding/decoding. This removes the restriction that each symbol must be represented by an integer number of bits, thus improving compression efficiency.

- Experiments indicate that compression can be improved by up to 10% over variable length coding/decoding.

- Complexity of arithmetic coding is higher than variable length coding, however.
H.263 Improvements over H.261

- H.261 only accepts QCIF and CIF format.
- No 1/2 pel motion estimation in H.261, instead it uses a spatial loop filter.
- H.261 does not use median predictors for motion vectors but simply uses the motion vector in the MB to the left as predictor.
- H.261 does not use a 3-D VLC for transform coefficient coding.
- GOB headers are mandatory in H.261.
- Quantizer changes at MB granularity requires 5 bits in H.261 and only 2 bits in H.263.
Demo: QCIF, 8 fps @ 28 Kb/s
Video Conferencing Demonstration
Section 1: Conferencing Video

- Video Compression Review
- Chronology of Video Standards
- The Input Video Format
- H.263 Overview
- □ H.263+ Overview
ITU-T Recommendation
H.263 Version 2
(H.263+)
H.263 Ver. 2 (H.263+)

- H.263+ was standardized in January, 1998.
- H.263+ is the working name for H.263 Version 2.
- Adds negotiable options and features while still retaining a backwards compatibility mode.
H.263: Overview

H.263 “plus” more negotiable options

- Arbitrary frame size, pixel aspect ratio (including square), and picture clock frequency
- Advanced INTRA frame coding
- Loop de-blocking filter
- Slice structures
- Supplemental enhancement information
- Improved PB-frames
H.263: Overview

*H.263 “plus” more negotiable options*

- Reference picture selection
- Temporal, SNR, and Spatial Scalability Mode
- Reference picture resampling
- Reduced resolution update mode
- Independently segmented decoding
- Alternative INTER VLC
- Modified quantization
In addition to the multiples of CIF, H.263+ permits any frame size from 4x4 to 2048x1152 pixels in increments of 4.

Besides the 12:11 pixel aspect ratio (PAR), H.263+ supports square (1:1), 525-line 4:3 picture (10:11), CIF for 16:9 picture (16:11), 525-line for 16:9 picture (40:33), and other arbitrary ratios.

In addition to picture clock frequencies of 29.97 Hz (NTSC), H.263+ supports 25 Hz (PAL), 30 Hz and other arbitrary frequencies.
Advanced INTRA Coding Mode

- In this mode, either the DC coefficient, 1st column, or 1st row of coefficients are predicted from neighboring blocks.
- Prediction is determined on a MB-by-MB basis.
- Essentially DPCM of INTRA DCT coefficients.
- Can save up to 40% of the bits on INTRA frames.
Advanced INTRA Mode

Row Prediction

Column Prediction

DCT Blocks
Deblocking Filter Mode

- Filter pixels along block boundaries while preserving edges in the image content.
- Filter is in the coding loop which means it filters the decoded reference frame used for motion compensation.
- Can be used in conjunction with a post-filter to further reduce coding artifacts.
Deblock Filter Mode

Block Boundary

Block Boundary
Deblocking Filter Mode

- A, B, C and D are replaced by new values, A1, B1, C1, and D1 based on a set of non-linear equations.
- The strength of the filter is proportional to the quantization strength.
Deblocking Filter Mode

A,B,C,D are replaced by A1,B1,C1, D1:

\[
\begin{align*}
B1 &= \text{clip}(B + d1) \\
C1 &= \text{clip}(C - d1) \\
A1 &= A - d2 \\
D1 &= D + d2 \\
d1 &= \text{Filter}\left(\frac{(A - 4B + 4C - D)}{8}, \text{Strength(QUANT)}\right) \\
d2 &= \text{clipd1}\left(\frac{(A - D)}{4}, \frac{d1}{3}\right)
\end{align*}
\]

\[
\text{Filter}(x, \text{Strength}) = \text{SIGN}(x) \times (\text{MAX}(0, \text{abs}(x)) - \text{MAX}(0, 2 \times (\text{abs}(x) - \text{Strength})))
\]

H.263+
Post-Filter

- Filter the decoded frame first horizontally, then vertically, using a 1-D filter.
- The post-filter strength is proportional to the quantization: \( \text{Strength(QUANT)} \)

\[
D_1 = D + \text{Filter}((A+B+C+E+F+G-6D)/8,\text{Strength})
\]
Deblocking Filter Demo

No Filter

Deblocking Loop Filter
Deblocking Filter Demo

No Filter

Loop & Post Filter
Filter Demo Videos

No Filter

Loop Filter

Loop & Post Filter
Slice Structured Mode

- Allows insertion of resynchronization markers at macroblock boundaries to improve network packetization and reduce overhead. More on this later.

- Allows more flexible tiling of video frames into independently decodable areas to support “view ports”, a.k.a. “local decode.”

- Improves error resiliency by reducing intra-frame dependence.

- Permits out-of-order transmission to reduce latency.
Slice Structured Mode

Slices start and end on macroblock boundaries.

No INTRA or MV Prediction across slice boundaries.
Slice Structured Mode

*Independent Segments*

Slice sizes remain fixed between INTRA frames.

No INTRA or MV Prediction across slice boundaries.
Supplemental Enhancement Information

Backwards compatible with H.263 but permits indication of supplemental information for features such as:

- Partial and full picture freeze requests
- Partial and full picture snapshot tags
- Video segment start and end tags for off-line storage
- Progressive refinement segment start and end tags
- Chroma keying info for transparency
Reference Picture Resampling

• Allows frame size changes of a compressed video sequence without inserting an INTRA frame.

• Permits the warping of the reference frame via affine transformations to address special effects such as zoom, rotation, translation.

• Can be used for emergency rate control by dropping frame sizes adaptively when bit rate get too high.
Reference Picture Resampling 

*with Warping*

Specify arbitrary warping parameters via displacement vectors from corners.
Reference Picture Resampling

**Factor of 4 Size Change**

No INTRA Frame Required when changing video frame sizes
Scalability Mode

<table>
<thead>
<tr>
<th>Enhancement Layer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enhancement Layer 1</td>
</tr>
<tr>
<td>Base Layer</td>
</tr>
</tbody>
</table>

- A scalable bit stream consists of layers representing different levels of video quality.
- Everything can be discarded except for the base layer and still have reasonable video.
- If bandwidth permits, one or more enhancement layers can also be decoded which refines the base layer in one of three ways:
  - temporal, SNR, or spatial
Layered Video Bitstreams

H.263+ Encoder

H.263+

Enhancement Layer 4
Enhancement Layer 3
Enhancement Layer 2
Enhancement Layer 1
Base Layer

20 kb/s
40 kb/s
90 kb/s
200 kb/s
320 kb/s

40 kb/s
90 kb/s
200 kb/s
Scalability Mode

- Scalability is typically used when one bit stream must support several different transmission bandwidths simultaneously, or some process downstream needs to change the data rate unbeknownst to the encoder.

- Example: Conferencing Multipoint Control Unit (we’ll see another example in Internet Video)
Layered Video Bit Streams in multipoint conferencing
Temporal Enhancement

Base Layer + B Frames

Higher Frame Rate!
Temporal scalability means that two or more frame rates can be supported by the same bit stream. In other words, frames can be discarded (to lower the frame rate) and the bit stream remains usable.
Temporal Scalability

- The discarded frames are never used as prediction.
- In the previous diagram the I and P frames form the base layer and the B frames from the temporal enhancement layer.
- This is usually achieved using bidirectional predicted frames or B-frames.
B Frames

2X frame rate for only 30% more bits
Temporal Scalability Demonstration

- layer 0, 3.25 fps, P-frames
- layer 1, 15 fps, B-frames
SNR Enhancement

Base Layer + SNR Layer

Better Spatial Quality!
SNR Scalability

- Base layer frames are coded just as they would be in a normal coding process.
- The SNR enhancement layer then codes the difference between the decoded base layer frames and the originals.
- The SNR enhancement MB’s may be predicted from the base layer or the previous frame in the enhancement layer, or both.
- The process may be repeated by adding another SNR enhancement layer, and so on...
SNR Scalability

Legend:

- **I** - Intracoded or Key Frame
- **P** - Predicted Frame
- **EI** - Enhancement layer key frame
- **EP** - Enhancement layer predicted frame

Base Layer (15 kbit/s)

Enhancement Layer (40 kbit/s)
SNR Scalability Demonstration

- layer 0, 10 fps, 40 kbps
- layer 1, 10 fps, 400 kbps
Spatial Enhancement

Base Layer + Spatial Layer

More Spatial Resolution!!
Spatial Scalability

- For spatial scalability, the video is down-sampled by two horizontally and vertically prior to encoding as the base layer.
- The enhancement layer is 2X the size of the base layer in each dimension.
- The base layer is interpolated by 2X before predicting the spatial enhancement layer.
Spatial Scalability

Enhancement Layer

Base Layer

EI  EP  EP
I   P   P
Spatial Scalability Demonstration

- layer 0, QCIF, 10 fps, 60 kbps
- layer 1, CIF, 10 fps, 300 kbps
Hybrid Scalability

It is possible to combine temporal, SNR and spatial scalability into a flexible layered framework with many levels of quality.
Hybrid Scalability

Enhancement Layer 2

Enhancement Layer 1

Base Layer
Scalability Demonstration

- SNR/Spatial Scalability, 10 fps
  - layer 0, 88x72, ~5 kbit/s
  - layer 1, 176x144, ~15
  - layer 2, 176x144, ~40
  - layer 3, 352x288, ~80
  - layer 4, 352x288, ~200
Other Miscellaneous Features

• Improved PB-frames
  – Improves upon the previous PB-frame mode by permitting forward prediction of “B” frame with a new vector.

• Reference picture selection (discussed later)
  – A lower latency method for dealing with error prone environments by using some type of back-channel to indicate to an encoder when a frame has been received and can be used for motion estimation.

• Reduced resolution update mode
  – Used for bit rate control by reducing the size of the residual frame adaptively when bit rate gets too high.
Other Miscellaneous Features

• Independently decodable segments
  – When signaled, it restricts the use of data outside of a current Group-of-Block segment or slice segment. Useful for error resiliency.

• Alternate INTER VLC
  – Permits use of an alternative VLC table that is better suited for INTRA coded blocks, or blocks with low quantization.
Other Miscellaneous Features

- Modified Quantization
  - Allows more flexibility in adapting quantizers on a macroblock by macroblock basis by enabling large quantizer changes through the use of escape codes.
  - Reduces quantizer step size for chrominance blocks, compared to luminance blocks.
  - Modifies the allowable DCT coefficient range to avoid clipping, yet disallows illegal coefficient/quantizer combinations.
Outline

Section 1: Conferencing Video
Section 2: Internet Review
Section 3: Internet Video
The Internet
Phone lines are “circuit-switched”. A (virtual) circuit is established at call initiation and remains for the duration of the call.
Computer networks are “packet-switched”. Data is fragmented into packets, and each packet finds its way to the destination using different routes. Lots of implications...
The Internet is heterogeneous [V. Cerf]
Network Access Layer consists of routines for accessing physical networks.

Internet Layer defines the datagram and handles the routing of data.

Host-to-Host Transport Layer provides end-to-end data delivery services.

Application Layer consists of applications and processes that use the network.

Network Access Layer consists of routines for accessing physical networks.
Data Encapsulation

- Application Layer
  - Data

- Transport Layer
  - Header
  - Data

- Internet Layer
  - Header
  - Header
  - Data

- Network Access Layer
  - Header
  - Header
  - Header
  - Data
Internet Protocol Architecture

Utility/Application
Host-Host Transport
Internet
Network Access Layer

TCP
UDP
SNMP
DNS
RTP

MIME
VIC/VAT
MBone

TELNET
FTP
SMTP
SMTP
SNMP

Internet Protocol Architecture

Internet Review

INTERNET
ACCESS
LAYER
Specific Protocols for Multimedia

Data

RTP

TCP

UDP

IP

Physical Network

Payload header

payload

payload

payload
The Internet Protocol (IP)

- IP implements two basic functions
  - addressing & fragmentation
- IP treats each packet as an independent entity.
- Internet routers choose the best path to send each packet based on its address. Each packet may take a different route.
- Routers may fragment and reassemble packets when necessary for transmission on smaller packet networks.
The Internet Protocol (IP)

- IP packets have a Time-to-Live, after which they are deleted by a router.
- IP does not ensure secure transmission.
- IP only error-checks headers, not payload.
- Summary: no guarantee a packet will reach its destination, and no guarantee of when it will get there.
Transmission Control Protocol (TCP)

- TCP is connection-oriented, end-to-end reliable, in-order protocol.
- TCP does not make any reliability assumptions of the underlying networks.
- Acknowledgment is sent for each packet.
- A transmitter places a copy of each packet sent in a timed buffer. If no "ack" is received before the time is out, the packet is re-transmitted.
- TCP has inherently large latency - not well suited for streaming multimedia.
Universal Datagram Protocol (UDP)

- UDP is a simple protocol for transmitting packets over IP.
- Smaller header than TCP, hence lower overhead.
- Does not re-transmit packets. This is OK for multimedia since a late packet usually must be discarded anyway.
- Performs check-sum of data.
Real time Transport Protocol (RTP)

- RTP carries data that has real time properties
- Typically runs on UDP/IP
- Does not ensure timely delivery or QoS.
- Does not prevent out-of-order delivery.
- Profiles and payload formats must be defined.
- Profiles define extensions to the RTP header for a particular class of applications such as audio/video conferencing (IETF RFC 1890).
Real-time Transport Protocol (RTP)

- Payload formats define how a particular kind of payload, such as H.261 video, should be carried in RTP.
- Used by Netscape LiveMedia, Microsoft NetMeeting®, Intel VideoPhone, ProShare® Video Conferencing applications and public domain conferencing tools such as VIC and VAT.
Real-time Transport Control Protocol (RTCP)

- RTCP is a companion protocol to RTP which monitors the quality of service and conveys information about the participants in an ongoing session.

- It allows participants to send transmission and reception statistics to other participants. It also sends information that allows participants to associate media types such as audio/video for lip-sync.
Real-time Transport Control Protocol (RTCP)

- Sender reports allow senders to derive round trip propagation times.
- Receiver reports include count of lost packets and inter-arrival jitter.
- Scales to a large number of users since must reduce the rate of reports as the number of participants increases.
- Most products today don’t use the information to avoid congestion, but that will change in the next year or two.
Multicast Backbone (Mbone)

- Most IP-based communication is unicast. A packet is intended for a single destination. For multi-participant applications, streaming multimedia to each destination individually can waste network resources, since the same data may be travelling along sub-networks.
- A multicast address is designed to enable the delivery of packets to a set of hosts that have been configured as members of a multicast group across various sub-networks.
Unicast Example

Streaming media to multi-participants

S1 sends duplicate packets because there’s two participants: D1, D2.

D2 sees excess traffic on this subnet.
Multicasting Example

Streaming media to multi-participants

S1 sends single set of packets to a multicast group.

D2 doesn’t see any excess traffic on this subnet.

Both D1 receivers subscribe to the same multicast group.

S1

S2

D2

D1

D1
Multicast Backbone (MBone)

- Most routers sold in the last 2-3 years support multicast.
- Not turned on yet in the Internet backbone.
- Currently there is an MBone overlay which uses a combination of multicast (where supported) and tunneling.
- Multicast at your local ISP may be 1-2 years away.
ReSerVation Protocol (RSVP)

Internet Draft

- Used by hosts to obtain a certain QoS from underlying networks for a multimedia stream.
- At each node, RSVP daemon attempts to make a resource reservation for the stream.
- It communicates with two local modules: admission control and policy control.
- Admission control determines whether the node has sufficient resources available. “The Internet Busy Signal”
- Policy control determines whether the user has administrative permission to make the reservation.
Real-time Streaming Protocol (RTSP) Internet Draft

- A “network remote control” for multimedia servers.
- Establishes and controls either a single or several time-synchronized streams of continuous media such as audio and video.
- Supports the following operations:
  - Requests a presentation from a media server.
  - Invite a media server to join a conference and playback or record.
  - Notify clients that additional media is available for an existing presentation.
Hyper-Text Transport Protocol (HTTP)

- HTTP generally runs on TCP/IP and is the protocol upon which World-Wide-Web data is transmitted.
- Defines a "stateless" connection between receiver and sender.
- Sends and receives MIME-like messages and handles caching, etc.
- No provisions for latency or QoS guarantees.
Outline

Section 1: Conferencing Video
Section 2: Internet Review
Section 3: Internet Video
Internet Video
How do we stream video over the Internet?

- How do we handle the special cases of unicasting? Multicasting?
- What about packet-loss? Quality of service? Congestion?

We’ll look at some solutions...
HTTP Streaming

- HTTP was not designed for streaming multimedia, nevertheless because of its widespread deployment via Web browsers, many applications stream via HTTP.
- It uses a custom browser plug-in which can start decoding video as it arrives, rather than waiting for the whole file to download.
- Operates on TCP so it doesn’t have to deal with errors, but the side effect is high latency and large inter-arrival jitter.
HTTP Streaming

- Usually a receive buffer is employed which can buffer enough data (usually several seconds) to compensate for latency and jitter.
- Not applicable to two-way communication!
- Firewalls are not a problem with HTTP.
RTP Streaming

• RTP was designed for streaming multimedia.
• Does not resend lost packets since this would add latency and a late packet might as well be lost in streaming video.
• Used by Intel Videophone, Microsoft NetMeeting, Netscape LiveMedia, RealNetworks, etc.
• Forms the basis for network video conferencing systems (ITU-T H.323)
RTP Streaming

• Subject to packet loss, and has no quality of service guarantees.
• Can deal with network congestion via RTCP reports under some conditions:
  – Should be encoding real time so video rate can be changed dynamically.
• Needs a payload defined for each media it carries.
H.263 Payload for RTP

- Payloads must be defined in the IETF for all media carried by RTP.
- A payload has been defined for H.263 and is now an Internet RFC.
- A payload has been defined for H.263+ as an ad-hoc group activity in the ITU and is now an Internet Draft.
- An RTP packet typically consists of...

<table>
<thead>
<tr>
<th>RTP Header</th>
</tr>
</thead>
<tbody>
<tr>
<td>H.263 Payload Header</td>
</tr>
<tr>
<td>H.263 Payload (bit stream)</td>
</tr>
</tbody>
</table>
H.263 Payload for RTP

- The H.263 payload header contains redundant information about the H.263 bit stream which can assist a payload handler and decoder in the event that related packets are lost.
- Slice mode of H.263+ aids RTP packetization by allowing fragmentation on MB boundaries (instead of MB rows) and restricting data dependencies between slices.
- But what do we do when packets are lost or arrive too late to use?
Error Resiliency: Redundancy & Concealment Techniques
Internet Packet Loss

- Depends on network topology.
- On the Mbone
  - 2-5% packet loss
  - single packet loss most common
- For end-to-end transmission, loss rates of 10% not uncommon.
- For ISPs, loss rates may be even higher during high periods of congestion.
Packet Loss Burst Lengths

Distribution of length of loss bursts observed at a receiver

Probability of bursts of length $b$

Length of loss bursts, $b$
Conditional loss probability

Number of consecutive packets lost, $n$

Probability of losing packet $n+1$
First Order Loss Model

2-Stage Gilbert Model

\[ p = 0.083 \quad q = 0.823 \]
Error Resiliency

- Error resiliency and compression have conflicting requirements.
- Video compression attempts to remove as much redundancy out of a video sequence as possible.
- Error resiliency techniques at some point must reconstruct data that has been lost and must rely on extrapolations from redundant data.
Error Resiliency

Errors tend to propagate in video compression because of its predictive nature.

- One block is lost.
- Error propagates to two blocks in the next frame.
Error Resiliency

There are essentially two approaches to dealing with errors from packet loss:

– Error redundancy methods are preventative measures that add extra information at the encoder to make it easier to recover when data is lost. The extra overhead decreases compression efficiency but should improve overall quality in the presence of packet loss.

– Error concealment techniques are the methods that are used to hide errors that occur once packets are lost.

Usually both methods are employed.
Simple INTRA Coding & Skipped Blocks

- Increasing the number of INTRA coded blocks that the encoder produces will reduce error propagation since INTRA blocks are not predicted.
- Blocks that are lost at the decoder are simply treated as empty INTER coded blocks. The block is simply copied from the previous frame.
- Very simple to implement.
Intra Coding Resiliency

Internet Video

Average PSNR vs. Data Rate (kbps)

- Resil 0, Loss 0
- Resil 5, Loss 0
- Resil 10, Loss 0
- Resil 0, Loss 10-20
- Resil 5, Loss 10-20
- Resil 10, Loss 10-20
Reference Picture Selection Mode of H.263+

In RPS Mode, a frame is not used for prediction in the encoder until it’s been acknowledged to be error free.
Reference Picture Selection

- **ACK-based**: a picture is assumed to contain errors, and thus is not used for prediction unless an ACK is received, or…
- **NACK-based**: a picture will be used for prediction unless a NACK is received, in which case the previous picture that didn’t receive a NACK will be used.
Multi-threaded Video

- Reference pictures are interleaved to create two or more independently decodable threads.
- If a frame is lost, the frame rate drops to 1/2 rate until a sync frame is reached.
- Same syntax as Reference Picture Selection, but without ACK/NACK.
- Adds some overhead since prediction is not based on most recent frame.
A video encoder contains a decoder (called the loop decoder) to create decoded previous frames which are then used for motion estimation and compensation.

The loop decoder must stay in sync with the real decoder, otherwise errors propagate.
Conditional Replenishment

• One solution is to discard the loop decoder.
• Can do this if we restrict ourselves to just two macroblock types:
  – INTRA coded and
  – empty (just copy the same block from the previous frame)
• The technique is to check if the current block has changed substantially since the previous frame and then code it as INTRA if it has changed. Otherwise mark it as empty.
• A periodic refresh of INTRA coded blocks ensures all errors eventually disappear.
Error Tracking

Appendix II, H.263

- Lost macroblocks are reported back to the encoder using a reliable back-channel.
- The encoder catalogs spatial propagation of each macroblock over the last M frames.
- When a macroblock is reported missing, the encoder calculates the accumulated error in each MB of the current frame.
- If an error threshold is exceeded, the block is coded as INTRA.
- Additionally, the erroneous macroblocks are not used as prediction for future frames in order to contain the error.
Prioritized Encoding

- Some parts of a bit stream contribute more to image artifacts than others if lost.
- The bit stream can be prioritized and more protection can be added for higher priority portions.

<table>
<thead>
<tr>
<th>Picture Header</th>
<th>Motion Vectors</th>
<th>MB Information</th>
<th>DC Coefficients</th>
<th>AC Coefficients</th>
</tr>
</thead>
</table>

Increasing Error Protection
Prioritized Encoding Demo

Unprotected Encoding

Prioritized Encoding (23% Overhead)

Videos used with permission of ICSI, UC Berkeley
Error Concealment by Interpolation

Lost block

Take the weighted average of 4 neighboring pixels.
Other Error Concealment Techniques

- Error Concealment with Least Square Constraints
- Error Concealment with Bayesian Estimators
- Error Concealment with Polynomial Interpolation
- Error Concealment with Edge-Based Interpolation
- Error Concealment with Multi-directional Recursive Nonlinear Filter (MRNF)

See references for more information...
Example: MRNF Filtering

MPQT@0.5 bpp, block loss: 10%

MRNF-GMLOS, PSNR=34.94dB
Network Congestion

- Most multimedia applications place the burden of rate adaptivity on the source.
- For multicasting over heterogeneous networks and receivers, it’s impossible to meet the conflicting requirements which forces the source to encode at a least-common denominator level.
- The smallest network pipe dictates the quality for all the other participants of the multicast session.
- If congestion occurs, the quality of service degrades as more packets are lost.
Receiver-driven Layered Multicast

- If the responsibility of rate adaptation is moved to the receiver, heterogeneity is preserved.
- One method of receiver based rate adaptivity is to combine a layered source with a layered transmission system.
- Each bit stream layer belongs to a different multicast group.
- In this way, a receiver can control the rate by subscribing to multicast groups and thus layers of the video bit stream.
Multicast groups are not transmitted on networks that have no subscribers.