Chapter 11
MPEG Video Coding — MPEG-1, 2, 4 and 7

11.1 Overview
11.2 MPEG-1
11.3 MPEG-2
11.4 MPEG-4
11.5 MPEG-7
11.1 Overview

• **MPEG**: *Moving Pictures Experts Group*, established in 1988 for the development of digital video.

• It is appropriately recognized that proprietary interests need to be maintained within the family of MPEG standards:
  - Accomplished by defining only an MPEG-compliant compressed bitstream that implicitly defines the decoder.
  - The continuous improvement and optimization of the compression algorithms on the encoder side are completely up to the manufacturers.

• As in H.261 and H.263, MPEG adopts *Hybrid Coding*, i.e., a combination of motion compensation and transform coding on prediction residual errors.
11.2 MPEG-1

- MPEG-1 adopts the CCIR601 digital TV format also known as SIF (Source Input Format).

- MPEG-1 supports only non-interlaced video. Normally, its picture resolution is:
  - $352 \times 240$ for NTSC video at 30 fps
  - $352 \times 288$ for PAL video at 25 fps
  - It uses 4:2:0 chroma subsampling

- The MPEG-1 standard is also referred to as ISO/IEC 11172. It has five parts: 11172-1 Systems, 11172-2 Video, 11172-3 Audio, 11172-4 Conformance, and 11172-5 Software.
Motion Compensation in MPEG-1

- Motion Compensation (MC) based video encoding in H.261 works as follows:

  - In Motion Estimation (ME), each macroblock (MB) of the Target P-frame is assigned a best matching MB from the previously coded I or P frame - **prediction**.

  - **prediction error**: The difference between the MB and its matching MB, sent to DCT and its subsequent encoding steps.

  - The prediction is from a previous frame — **forward prediction**.
The MB containing part of a ball in the Target frame cannot find a good matching MB in the previous frame because half of the ball was occluded by another object. A match however can readily be obtained from the next frame.

**Fig 11.1:** The Need for Bidirectional Search.
Motion Compensation in MPEG-1 (Cont’d)

- MPEG introduces a third frame type — \textit{B-frames}, and its accompanying bi-directional motion compensation.

- The MC-based B-frame coding idea is illustrated in Fig. 11.2:
  - Each MB from a B-frame will have up to two motion vectors (MVs) (one from the forward and one from the backward prediction).
  - If matching in both directions is successful, then two MVs will be sent and the two corresponding matching MBs are averaged (indicated by ‘\%’ in the figure) before comparing to the Target MB for generating the prediction error.
  - If an acceptable match can be found in only one of the reference frames, then only one MV and its corresponding MB will be used from either the forward or backward prediction.
**Fig 11.2: B-frame Coding Based on Bidirectional Motion Compensation.**
• MPEG uses $M$ to indicate the interval between a P-frame and its preceding I- or P frame, and $N$ to indicate the interval between two consecutive I-frames. In Fig. 11.3, $M = 3$, $N = 9$.
• Since the MPEG encoder and decoder cannot work for any macroblock from a B frame without its succeeding P- or I-frame, the actual coding and transmission order is different from the display order of the video.
Typical Sizes of MPEG-1 Frames

- The typical size of compressed P-frames is significantly smaller than that of I-frames — because temporal redundancy is exploited in inter-frame compression.

- B-frames are even smaller than P-frames — because of (a) the advantage of bi-directional prediction and (b) the lowest priority given to B-frames.

Table 11.4: Typical Compression Performance of MPEG-1 Frames

<table>
<thead>
<tr>
<th>Type</th>
<th>Size</th>
<th>Compression</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>18kB</td>
<td>7:1</td>
</tr>
<tr>
<td>P</td>
<td>6kB</td>
<td>20:1</td>
</tr>
<tr>
<td>B</td>
<td>2.5kB</td>
<td>50:1</td>
</tr>
<tr>
<td>Avg</td>
<td>4.8kB</td>
<td>27:1</td>
</tr>
</tbody>
</table>
Fig 11.5: Layers of MPEG-1 Video Bitstream.
11.3 MPEG-2

- **MPEG-2**: For higher quality video at a bit-rate of more than 4 Mbps.
- Defined seven **profiles** aimed at different applications
  - Simple, Main, SNR scalable, Spatially scalable, High, 4:2:2, Multiview.
  - Within each profile, up to four **levels** are defined (Table 11.5).
  - The DVD video specification allows only four display resolutions: 720×480, 704×480, 352×480, and 352×240.
  - a restricted form of the MPEG-2 Main profile at the Main and Low levels.
### Table 11.5: Profiles and Levels in MPEG-2

<table>
<thead>
<tr>
<th>Level</th>
<th>Simple profile</th>
<th>Main profile</th>
<th>SNR Scalable profile</th>
<th>Spatially Scalable profile</th>
<th>High Profile</th>
<th>4:2:2 Profile</th>
<th>Multiview Profile</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>*</td>
<td></td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>High 1440</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Main</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 11.6: Four Levels in the Main Profile of MPEG-2

<table>
<thead>
<tr>
<th>Level</th>
<th>Max. Resolution</th>
<th>Max fps</th>
<th>Max pixels/sec</th>
<th>Max coded Data Rate (Mbps)</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>1920 × 1152</td>
<td>60</td>
<td>62.7 × 10^6</td>
<td>80</td>
<td>film production</td>
</tr>
<tr>
<td>High 1440</td>
<td>1440 × 1152</td>
<td>60</td>
<td>47.0 × 10^6</td>
<td>60</td>
<td>consumer HDTV</td>
</tr>
<tr>
<td>Main</td>
<td>720 × 576</td>
<td>30</td>
<td>10.4 × 10^6</td>
<td>15</td>
<td>Studio TV</td>
</tr>
<tr>
<td>Low</td>
<td>352 × 288</td>
<td>30</td>
<td>3.0 × 10^6</td>
<td>4</td>
<td>consumer tape equiv.</td>
</tr>
</tbody>
</table>
Supporting Interlaced Video

- MPEG-2 must support interlaced video as well since this is one of the options for digital broadcast TV and HDTV.

- In interlaced video each frame consists of two fields, referred to as the top-field and the bottom-field.

  - In a Frame-picture, all scanlines from both fields are interleaved to form a single frame, then divided into 16×16 macroblocks and coded using MC.

  - If each field is treated as a separate picture, then it is called Field-picture.
Fig. 11.6: Field pictures and Field-prediction for Field-pictures in MPEG-2.
(a) Frame-picture vs. Field-pictures
(b) Field Prediction for Field-pictures
Other Major Differences from MPEG-1

- Better resilience to bit-errors: In addition to Program Stream, a Transport Stream is added to MPEG-2 bit streams.

- Support of 4:2:2 and 4:4:4 chroma subsampling.

- More restricted slice structure: MPEG-2 slices must start and end in the same macroblock row. In other words, the left edge of a picture always starts a new slice and the longest slice in MPEG-2 can have only one row of macroblocks.

- More flexible video formats: It supports various picture resolutions as defined by DVD, ATV and HDTV.
11.4.1 Overview of MPEG-4

- **MPEG-4**: a standard with different emphasis — besides compression, pays great attention to issues about user interactivities.

- **MPEG-4** departs from its predecessors in adopting a new **object-based coding**:
  - Offering higher compression ratio, also beneficial for digital video composition, manipulation, indexing, and retrieval.
  - Figure 11.11 illustrates how MPEG-4 videos can be composed and manipulated by simple operations on the visual objects.

- The bitrate for **MPEG-4** video covers a large range between 5 kbps to 10 Mbps.
Fig. 11.11: Composition and Manipulation of MPEG-4 Videos.
(VOP = Video object plane)
Overview of MPEG-4 (Cont’d)

• MPEG-4 (Fig. 11.12(b)) is a standard that emphasizes:
  
  (a) Composing media objects to create desirable audiovisual scenes.

  (b) Multiplexing and synchronizing the bitstreams for these media data entities so that they can be transmitted with guaranteed Quality of Service (QoS).

  (c) Interacting with the audiovisual scene at the receiving end — provides a toolbox of advanced coding modules and algorithms for audio and video compressions.
**Fig. 11.12:** Comparison of interactivities in MPEG standards:
(a) reference models in MPEG-1 and 2 (interaction in dashed lines supported only by MPEG-2);
(b) MPEG-4 reference model.
Overview of MPEG-4 (Cont’d)

- The hierarchical structure of MPEG-4 visual bitstreams is very different from that of MPEG-1 and -2, it is very much video object-oriented.

<table>
<thead>
<tr>
<th>Video-object Sequence (VS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Video Object (VO)</td>
</tr>
<tr>
<td>Video Object Layer (VOL)</td>
</tr>
<tr>
<td>Group of VOPs (GOV)</td>
</tr>
<tr>
<td>Video Object Plane (VOP)</td>
</tr>
</tbody>
</table>

**Fig. 11.13:** Video Object Oriented Hierarchical Description of a Scene in MPEG-4 Visual Bitstreams.
Overview of MPEG-4 (Cont’d)

1. **Video-object Sequence** (VS)—delivers the complete MPEG-4 visual scene, which may contain 2-D or 3-D natural or synthetic objects.

2. **Video Object** (VO) — a particular object in the scene, which can be of arbitrary (non-rectangular) shape corresponding to an object or background of the scene.

3. **Video Object Layer** (VOL) — facilitates a way to support (multi-layered) scalable coding. A VO can have multiple VOLs under scalable coding, or have a single VOL under non-scalable coding.

4. **Group of Video Object Planes** (GOV) — groups Video Object Planes together (optional level).

5. **Video Object Plane** (VOP) — a snapshot of a VO at a particular moment.
11.4.2 Video Object-based Coding in MPEG-4

**VOP-based vs. Frame-based Coding**

- MPEG-1 and -2 do not support the VOP concept, and hence their coding method is referred to as frame-based (also known as Block-based) coding.

- Fig. 11.14 (c) illustrates a possible example in which both potential matches yield small prediction errors for block-based coding.

- Fig. 11.14 (d) shows that each VOP is of arbitrary shape and ideally will obtain a unique motion vector consistent with the actual object motion.
Fig. 11.14: Comparison between Block-based Coding and Object-based Coding.
VOP-based Coding

• MPEG-4 VOP-based coding also employs the Motion Compensation technique:
  - An Intra-frame coded VOP is called an I-VOP.
  - The Inter-frame coded VOPs are called P-VOPs if only forward prediction is employed, or B-VOPs if bi-directional predictions are employed.

• The new difficulty for VOPs: may have arbitrary shapes, shape information must be coded in addition to the texture of the VOP.

Note: *texture* here actually refers to the visual content, that is the gray-level (or chroma) values of the pixels in the VOP.
VOP-based Motion Compensation (MC)

- MC-based VOP coding in MPEG-4 again involves three steps:
  
  (a) Motion Estimation.
  
  (b) MC-based Prediction.
  
  (c) Coding of the prediction error.

- Only pixels within the VOP of the current (Target) VOP are considered for matching in MC.

- To facilitate MC, each VOP is divided into many macroblocks (MBs). MBs are by default 16×16 in luminance images and 8 × 8 in chrominance images.
• MPEG-4 defines a rectangular bounding box for each VOP (see Fig. 11.15 for details).

• The macroblocks that are entirely within the VOP are referred to as Interior Macroblocks.

• The macroblocks that straddle the boundary of the VOP are called Boundary Macroblocks.

To help matching every pixel in the target VOP and meet the mandatory requirement of rectangular blocks in transform coding (e.g., DCT), a pre-processing step of padding is applied to the Reference VOPs prior to motion estimation.

Note: Padding only takes place in the Reference VOPs.
Fig. 11.15: Bounding Box and Boundary Macroblocks of VOP.
Padding

- For all Boundary MBs in the Reference VOP, *Horizontal Repetitive Padding* is invoked first, followed by *Vertical Repetitive Padding*.

![Diagram showing sequence of paddings](image)

| Horizontal Repetitive Padding | Vertical Repetitive Padding | Extended Padding |

**Fig. 11.16:** A Sequence of Paddings for Reference VOPs in MPEG-4.

- Afterwards, for all *Exterior Macroblocks* that are outside of the VOP but adjacent to one or more Boundary MBs, *extended padding* will be applied.
Algorithm 11.1 Horizontal Repetitive Padding:

BEGIN

FOR all rows in Boundary MBs in the Reference VOP

IF ∃ (boundary pixel) in the row

FOR all interval outside of VOP

IF interval is bounded by only one boundary pixel b

assign the value of b to all pixels in interval

ELSE // interval is bounded by two boundary pixels $b_1$ and $b_2$

assign the value of $(b_1 + b_2)/2$ to all pixels in interval

END

• The subsequent Vertical Repetitive Padding algorithm works in a similar manner.
Example 11.1: Repetitive Paddings

Fig. 11.17: An example of Repetitive Padding in a boundary macroblock of a Reference VOP:
(a) Original pixels within the VOP
(b) After Horizontal Repetitive Padding
(c) Followed by Vertical Repetitive Padding.
Motion Vector Coding

- Let $C(x + k, y + l)$ be pixels of the MB in Target VOP, and $R(x+i+k, y+j+l)$ be pixels of the MB in Reference VOP.

- **A Sum of Absolute Difference (SAD)** for measuring the difference between the two MBs can be defined as:

  $$SAD(i, j) = \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} C(x+k, y+l) \times R(x+i+k, y+j+l) \times Map(x+k, y+l)$$

  $N$ — the size of the MB. $Map(p, q) = 1$ when $C(p, q)$ is a pixel within the target VOP, otherwise $Map(p, q) = 0$.

- The vector $(i, j)$ that yields the minimum SAD is adopted as the motion vector $MV(u, v)$:

  $$MV(u, v) = \{(i, j) | SAD(i, j) \text{is minimum, } i \in [-p, p], j \in [-p, p]\}$$

  $p$ — the maximal allowable magnitude for $u$ and $v$. 
Texture Coding

• Texture coding in MPEG-4 can be based on:
  - DCT or
  - Shape Adaptive DCT (SA-DCT).

I. Texture coding based on DCT

• In I-VOP, the gray values of the pixels in each MB of the VOP are directly coded using the DCT followed by VLC, similar to what is done in JPEG.

• In P-VOP or B-VOP, MC-based coding is employed — it is the prediction error that is sent to DCT and VLC.
• Coding for the Interior MBs:
  - Each MB is 16 × 16 in the luminance VOP and 8 × 8 in the chrominance VOP.
  - Prediction errors from the six 8×8 blocks of each MB are obtained after the conventional motion estimation step.

• Coding for Boundary MBs:
  - For portions of the Boundary MBs in the Target VOP outside of the VOP, zeros are padded to the block sent to DCT since ideally prediction errors would be near zero inside the VOP.
  - After MC, texture prediction errors within the Target VOP are obtained.
II. SA-DCT based coding for Boundary MBs

- Shape Adaptive DCT (SA-DCT) is another texture coding method for boundary MBs.
- Due to its effectiveness, SA-DCT has been adopted for coding boundary MBs in MPEG-4 Version 2.
- It uses the 1D DCT-N transform and its inverse, IDCT-N:

- **1D DCT-N:**

\[
F(u) = \sqrt{\frac{2}{N}} C(u) \sum_{i=0}^{N-1} \cos \left( \frac{(2i+1)u\pi}{2N} \right) f(i)
\]  

(11.4)

- **1D IDCT-N:**

\[
\hat{f}(i) = \sum_{u=0}^{N-1} \sqrt{\frac{2}{N}} C(u) \cos \left( \frac{(2i+1)u\pi}{2N} \right) F(u)
\]  

(11.5)
where $i = 0, 1, \ldots, N - 1; u = 0, 1, \ldots, N - 1$ and

$$C(u) = \begin{cases} \frac{\sqrt{2}}{2} & \text{if } u = 0, \\ 1 & \text{otherwise.} \end{cases}$$

- SA-DCT is a 2D DCT and it is computed as a separable 2D transform in two iterations of 1D DCT-N.
- Fig. 11.18 illustrates the process of texture coding for boundary MBs using the Shape Adaptive DCT (SA-DCT).
Fig. 11.18: Texture Coding for Boundary MBs Using the Shape Adaptive DCT (SA-DCT).
11.4.3 Synthetic Object Coding in MPEG-4

2D Mesh Object Coding

- **2D mesh**: a tessellation (or partition) of a 2D planar region using polygonal patches:
  - The vertices of the polygons are referred to as *nodes* of the mesh.
  - The most popular meshes are *triangular meshes* where all polygons are triangles.
  - The MPEG-4 standard makes use of two types of 2D mesh: **uniform mesh** and **Delaunay mesh**
  - 2D mesh object coding is compact. All coordinate values of the mesh are coded in half-pixel precision.
  - Each 2D mesh is treated as a *mesh object plane (MOP)*.
**Fig. 11.21: 2D Mesh Object Plane (MOP) Encoding Process**
2D Mesh Geometry Coding

- MPEG-4 allows four types of uniform meshes with different triangulation structures.

Fig. 11.22: Four Types of Uniform Meshes.
• **Definition:** If $D$ is a Delaunay triangulation, then any of its triangles $t_n = (P_i, P_j, P_k) \in D$ satisfies the property that the circumcircle of $t_n$ does not contain in its interior any other node point $P_l$.

• A Delaunay mesh for a video object can be obtained in the following steps:

  1. **Select boundary nodes of the mesh:** A polygon is used to approximate the boundary of the object.

  2. **Choose interior nodes:** Feature points, e.g., edge points or corners, within the object boundary can be chosen as interior nodes for the mesh.

  3. **Perform Delaunay triangulation:** A *constrained Delaunay triangulation* is performed on the boundary and interior nodes with the polygonal boundary used as a constraint.
11.5 MPEG-7

• The main objective of MPEG-7 is to serve the need of audio-visual content-based retrieval (or audiovisual object retrieval) in applications such as digital libraries.

• Nevertheless, it is also applicable to any multimedia applications involving the generation (content creation) and usage (content consumption) of multimedia data.

• MPEG-7 became an International Standard in September 2001 — with the formal name Multimedia Content Description Interface.
Applications Supported by MPEG-7

- MPEG-7 supports a variety of multimedia applications. Its data may include still pictures, graphics, 3D models, audio, speech, video, and composition information (how to combine these elements).

- These MPEG-7 data elements can be represented in textual format, or binary format, or both.

- Fig. 11.27 illustrates some possible applications that will benefit from the MPEG-7 standard.
Fig. 11.27: Possible Applications using MPEG-7.
MPEG-7 and Multimedia Content Description

• MPEG-7 has developed Descriptors (D), Description Schemes (DS) and Description Definition Language (DDL). The following are some of the important terms:
  - **Feature** — characteristic of the data.
  - **Description** — a set of instantiated Ds and DSs that describes the structural and conceptual information of the content, the storage and usage of the content, etc.
  - **D** — definition (syntax and semantics) of the feature.
  - **DS** — specification of the structure and relationship between Ds and between DSs.
  - **DDL** — syntactic rules to express and combine DSs and Ds.

• The scope of MPEG-7 is to standardize the Ds, DSs and DDL for descriptions. The mechanism and process of producing and consuming the descriptions are beyond the scope of MPEG-7.
Descriptor (D)

- The descriptors are chosen based on a comparison of their performance, efficiency, and size. Low-level visual descriptors for basic visual features include:
  - **Color**
    * Color space. (a) RGB, (b) YCbCr, (c) HSV (hue, saturation, value), (d) HMMD (HueMaxMinDiff), (e) 3D color space derivable by a $3 \times 3$ matrix from RGB, (f) monochrome.
    * Color quantization. (a) Linear, (b) nonlinear, (c) lookup tables.
    * Dominant colors.
    * Scalable color.
    * Color layout.
    * Color structure.
    * Group of Frames/Group of Pictures (GoF/GoP) color.
- **Texture**
  * Homogeneous texture.
  * Texture browsing.
  * Edge histogram.

- **Shape**
  * Region-based shape.
  * Contour-based shape.
  * 3D shape.
- **Motion**
  * Camera motion (see Fig. 11.28).
  * Object motion trajectory.
  * Parametric object motion.
  * Motion activity.

- **Localization**
  * Region locator.
  * Spatiotemporal locator.

- **Others**
  * Face recognition.
Fig. 11.28: Camera motions: pan, tilt, roll, dolly, track, and boom.
Description Scheme (DS)

• Basic elements
  - Datatypes and mathematical structures.
  - Constructs.
  - Schema tools.

• Content Management
  - Media Description.
  - Creation and Production Description.
  - Content Usage Description.

• Content Description
  - Structural Description.
A Segment DS, for example, can be implemented as a class object. It can have five subclasses: Audiovisual segment DS, Audio segment DS, Still region DS, Moving region DS, and Video segment DS. The subclass DSs can recursively have their own subclasses.

- Conceptual Description.

- **Navigation and access**
  - Summaries.
  - Partitions and Decompositions.
  - Variations of the Content.

- **Content Organization**
  - Collections.
  - Models.

- **User Interaction**
  - User Preference.
Fig. 11.29: MPEG-7 video segment.
Fig. 11.30: A video summary.
Description Definition Language (DDL)

- MPEG-7 adopted the XML Schema Language initially developed by the WWW Consortium (W3C) as its Description Definition Language (DDL). Since XML Schema Language was not designed specifically for audiovisual contents, some extensions are made to it:
  - Array and matrix data types.
  - Multiple media types, including audio, video, and audiovisual presentations.
  - Intellectual Property Management and Protection (IPMP) for Ds and DSs.