MPEG-4

- MPEG-4 is the international standard for true multimedia coding.
- MPEG-4 provides very low bitrate & error resilience for Internet and wireless.
- MPEG-4 can be carried in MPEG-2 systems layer.
MPEG-4

- 3-D facial animation
- Wavelet texture coding
- Mesh coding with texture mapping
- Media integration of text and graphics
- Text to speech synthesis

Applications of MPEG-4

- Multimedia broadcasting and presentations
- Virtual talking humans
- Advanced interpersonal communication systems
- Games
- Storytelling
- Language teaching
- Speech rehabilitation
- Teleshopping
- Telelearning
MPEG-4

- Real audio and video objects
- Synthetic audio and video
- Integration of Synthetic & Natural contents (Synthetic & Natural Hybrid Coding)

MPEG-4

- Traditional video coding is block-based.
- MPEG-4 provides object-based representation for better compression and functionalities.
- Objects are rendered after decoding object descriptions.
- Display of content layers can be selected at MPEG-4 terminal.
MPEG-4

• User can search or store objects for later use.
• Content does not depend on the display resolution.
• Network providers can re-purpose content for different networks and users.

Scope & Features of MPEG-4

• Authors
  – reusability
  – flexibility
  – content owner rights
• Network providers
• End users
Media Objects

- Primitive Media Objects
- Compound Media Objects
- Examples
  - Still Images (e.g. fixed background)
  - Video objects (e.g., a talking person-without background)
  - Audio objects (e.g., the voice associated with that person)
  - etc

MPEG-4 Versions
MPEG-4

VLB Core
1. Low resolution CIF (360X288)
2. Low frame rate 15fps
3. High coding efficiency
4. Low complexity, low error
5. Random access
6. Fast forward/reverse

High Bitrate
1. Higher resolution
2. Higher frame rate
3. Interlaced video

Content-based functionalities
1. Interactivity
2. Flexible representation and Manipulation in the compressed Domain
3. Hybrid coding

User Interactions

• Client Side
  – content manipulation done at client terminal
    • changing position of an object
    • making it visible or invisible
    • changing the font size of text

• Server Side
  – requires back channel
• Efficient representation of visual objects of arbitrary shape to support content-based functionalities
• Supports most functionalities of MPEG-1 and MPEG-2
  – rectangular sized images
  – several input formats
  – frame rates
  – bit rates
  – spatial, temporal and quality scalability
Object Composition

- Objects are organized in a scene graph.
- VRML based binary format BIF is used to specify scene graph.
- 2-D and 3-D objects, transforms and properties are specified.
- MPEG-4 allows objects to be transmitted once, and displayed repeatedly in the scene after transformations.
Standardized Ways

- To represent “media object”
  - visual or audiovisual
  - synthetic or natural
- To multiplex and synchronize the data associated with media objects for transportation over the network
- Interact with audiovisual scene generated at the receiver’s end.
Standardized Ways To

- place a media objects anywhere in a given coordinate system;
- apply transforms to change the geometrical or acoustical appearances of media objects;
- group primitive media objects to form compound media objects;
- apply stream data to media objects to modify their attributes;

Interaction with media objects

- change the viewing/listening point of the scene, e.g., by navigating through a scene;
- drag objects in the scene to a different position;
- trigger a cascade of events by clicking on specific objects, e.g., starting or sopping a video stream;
- select the desired language when multiple language tracks are available;
- more complex behavior (e.g., virtual phone rings, user answers and communication link is established)
Textures, Images and Video

- Efficient compression of
  - images and video
  - textures for texture mapping on 2D and 3D meshes
  - implicit 2D meshes
  - time-varying geometry streams that animate meshes
Textures, Images and Video

- Efficient random access to all types of visual objects
- Extended manipulation functionalities for images and video sequences
- Content-based coding of images and video
- Content-based scalability of textures, images and video
- Spatial, temporal and quality scalability
- Error robustness and resilience

2-D Animated Meshes

- A 2-D mesh is tessellation of a 2-D planar region into triangles.
- Dynamic meshes contain mesh geometry and motion.
- 2-D meshes can be used for texture mapping. Three nodes of triangle defines affine motion.
Texture Mapping

(a) (b)

2-D Mesh Modeling
2-D Mesh Representation of Video Object

• Video Object Manipulation
  – Augmented Reality
  – Synthetic-object-transfiguration/animation
  – Spatio-temporal interpolation (e.g., frame rate up-conversion)

• Video Object Compression
  – transmit texture maps only at key frames
  – animate texture maps for the intermediate frames

2-D Mesh Representation of Video Object

• Content-Based Indexing
  – Provides vertex-based object shape representation which is more efficient than the bitmap representation of shape-based object retrieval
  – Provides accurate object trajectory information that can be used to retrieve visual objects with specific motion
  – Animated key snapshots as visual synopsis of objects
MPEG-4 Video and Image Coding Scheme

- Shape coding and motion compensation
- DCT-based texture coding
  - standard 8x8 and shape adapted DCT
- Motion compensation
  - local block based (8x8 or 16x16)
  - global (affine) for sprites
Sprite Panorama

• First compute static “sprite” or “mosaic”
• Then transmit 8 or 6 global motion (camera) parameters for each frame to reconstruct the frame from the “sprite”
• Moving foreground is transmitted separately as an arbitrary-shape video object.

Steps in Sprite Construction

• Incremental mosaic construction
• Incremental residual estimation
• Computation of significance measures on the residuals
• Spatial coding and decoding
• Visit http://www.wisdom.weizmann.ac.il/~irani/abstracts/mosaics.html
Other Objects

- Text and graphics
- Talking synthetic head and associated text
- Synthetic sound
Face and Body Animation

- Face animation is in MPEG-4 version 1.
- Body animation is in MPEG-4 version 2.
- Face animation parameters displace feature points from neutral position.
- Body animation parameters are joint angles.
- Face and body animation parameter sequences are compressed to low bit rate.
- Facial expressions: joy, sadness, anger, fear, disgust and surprise.
- Visemes

Face Model

- Face model (3D) specified in VRML, can be downloaded to the terminal with MPEG-4
Neutral Face

- Face is gazing in the Z direction
- Face axes parallel to the world axes
- Pupil is 1/3 of iris in diameter
- Eyelids are tangent to the iris
- Upper and lower teeth are touching and mouth is closed
- Tongue is flat, and the tip of tongue is touching the boundary between upper and lower teeth
Face Node

- **FAP (Facial Animation Parameters)**
  - FAPs allow to animate 3-D facial node at the receiver.
  - Animation of key feature points and reproduction of visemes & expressions

- **Face Definition Parameters (FDP)**
  - FDP allow to configure facial model to be used at the receiver, either by sending a new model, or by adapting a previously available model. Sent only once.

- **Face Interpolation Table (FIT)**
  - FIT allow to define interpolation rules for FAPs that have to be interpolated at the receiver. The 3-D model is animated using FAPs sent and FAPs interpolated.

- **Face Animation Table (FAT)**
  - It specifies for each selected FAP the set of vertices to be affected in a new downloaded model, as well as the way they are affected. E.g. FAP ‘open jaw’, then table defines what that means in terms of moving the feature points.

---

Facial Animation Parameters (FAPS)

- 2 eyeball and 3 head rotations are represented using Euler angles
- Each FAP is expressed as a fraction of neutral face mouth width, mouth-nose distance, eye separation, or iris diameter.
FAP Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>FAPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visemes &amp; expressions</td>
<td>2</td>
</tr>
<tr>
<td>jaw, chin, inner lower-lip, corner lip, mid-lip</td>
<td>16</td>
</tr>
<tr>
<td>eyeballs, pupils, eyelids</td>
<td>12</td>
</tr>
<tr>
<td>eyebrow</td>
<td>8</td>
</tr>
<tr>
<td>cheeks</td>
<td>4</td>
</tr>
<tr>
<td>tongue</td>
<td>5</td>
</tr>
<tr>
<td>head rotation</td>
<td>3</td>
</tr>
<tr>
<td>outer lip position</td>
<td>10</td>
</tr>
<tr>
<td>nose</td>
<td>4</td>
</tr>
<tr>
<td>ears</td>
<td>4</td>
</tr>
</tbody>
</table>

FAPS

- **31**: raise_l_l_eyebrow (vertical displacement of left inner eyebrow)
- **32**: raise_r_l_eyebrow (vertical displacement of right inner eyebrow)
- **33**: raise_l_m_eyebrow (vertical displacement of left middle eyebrow)
- **34**: raise_r_m_eyebrow (vertical displacement of right middle eyebrow)
- **35**: 
FAP Data

- Synthetically generated
- Extracted by analysis
  - Real-time (video phones)
  - Off-line (story telling)
  - Fully automatic (video phones)
  - Human-guided (teleshopping & gaming)

FAPs Masking Scheme Options

- No FAPs are coded for the corresponding group
- A mask is given indicating which FAPs in the corresponding group are coded. FAPs not coded, retain their previous values
- A mask is given indicating which FAPs in the corresponding group are coded. The decoder should interpolate FAPs not selected by the group mask.
- All FAPs in the group are coded.
Four Cases of FDP

- No FDP data is sent, residing 3-D model at the receiver is used for animation
- Feature points (calibrate the model) are sent
- Feature points and texture are sent
- Facial Animation Tables (FATs) and 3-D model are sent
  - FAT specify the FAP behavior (which and how the new model vertices should be moved for each FAP)

- It is difficult for the sender to know precisely the appearance of the synthesized result at the receiver since a large number of models may be used.
3-D Facial Animation System

FAPs

• Speech recognition can use FAPs to increase recognition rate.
• FAPs can be used to animate face models by text to speech systems
• In HCI FAPs can be used to communicate speech, emotions, etc, in particular in noisy environment.
Visemes and Expressions

- For each frame a weighted combination of two visemes and two facial expressions
- After FAPs are applied the decoder can interpret effect of visemes and expressions
- Definitions of visemes and expressions using FAPs can be downloaded

Phonemes and Visemes

- 56 phonemes
  - 37 consonants
  - 19 vowels/diphthongs
- 56 phonemes can be mapped to 35 visemes
- A triseme is made up of three visemes to capture co-articulations
## 56 Phonemes

<table>
<thead>
<tr>
<th>Phone</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>eeg</td>
</tr>
<tr>
<td>ac</td>
<td>bet</td>
</tr>
<tr>
<td>ah</td>
<td>butt</td>
</tr>
<tr>
<td>ao</td>
<td>about</td>
</tr>
<tr>
<td>aw</td>
<td>bough</td>
</tr>
<tr>
<td>ax</td>
<td>the</td>
</tr>
<tr>
<td>axr</td>
<td>dinger</td>
</tr>
<tr>
<td>ay</td>
<td>bite</td>
</tr>
<tr>
<td>eh</td>
<td>bet</td>
</tr>
<tr>
<td>er</td>
<td>bird</td>
</tr>
<tr>
<td>ey</td>
<td>bit</td>
</tr>
<tr>
<td>ih</td>
<td>bit</td>
</tr>
<tr>
<td>ix</td>
<td>roses</td>
</tr>
<tr>
<td>iy</td>
<td>beat</td>
</tr>
<tr>
<td>ow</td>
<td>boat</td>
</tr>
<tr>
<td>ox</td>
<td>joy</td>
</tr>
<tr>
<td>uh</td>
<td>ut</td>
</tr>
<tr>
<td>uw</td>
<td>ut</td>
</tr>
<tr>
<td>ux</td>
<td>ut</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>bcl</td>
<td>m</td>
</tr>
<tr>
<td>pcl</td>
<td>f</td>
</tr>
<tr>
<td>ch</td>
<td>d</td>
</tr>
<tr>
<td>dcl</td>
<td>n</td>
</tr>
<tr>
<td>dx</td>
<td>p</td>
</tr>
<tr>
<td>en</td>
<td>button</td>
</tr>
<tr>
<td>f</td>
<td>gief</td>
</tr>
<tr>
<td>g</td>
<td>gcl</td>
</tr>
<tr>
<td>gcl</td>
<td>g-closure</td>
</tr>
<tr>
<td>hh</td>
<td>hay</td>
</tr>
<tr>
<td>hv</td>
<td>Lehigh</td>
</tr>
<tr>
<td>jh</td>
<td>judge</td>
</tr>
<tr>
<td>k</td>
<td>kick</td>
</tr>
<tr>
<td>kcl</td>
<td>k-closure</td>
</tr>
<tr>
<td>l</td>
<td>led</td>
</tr>
<tr>
<td>m</td>
<td>mom</td>
</tr>
<tr>
<td>n</td>
<td>non</td>
</tr>
<tr>
<td>ng</td>
<td>sing</td>
</tr>
<tr>
<td>nx</td>
<td>flapped-n</td>
</tr>
<tr>
<td>p</td>
<td>gop</td>
</tr>
<tr>
<td>pcl</td>
<td>p-closure</td>
</tr>
<tr>
<td>q</td>
<td>glottal stop</td>
</tr>
<tr>
<td>r</td>
<td>red</td>
</tr>
<tr>
<td>s</td>
<td>sis</td>
</tr>
<tr>
<td>sh</td>
<td>shoe</td>
</tr>
<tr>
<td>t</td>
<td>tot</td>
</tr>
<tr>
<td>tcl</td>
<td>t-closure</td>
</tr>
<tr>
<td>th</td>
<td>thief</td>
</tr>
<tr>
<td>v</td>
<td>very</td>
</tr>
<tr>
<td>w</td>
<td>wet</td>
</tr>
<tr>
<td>y</td>
<td>get</td>
</tr>
<tr>
<td>z</td>
<td>zoo</td>
</tr>
<tr>
<td>zh</td>
<td>measure</td>
</tr>
<tr>
<td>ep</td>
<td>epithetic</td>
</tr>
<tr>
<td>h#</td>
<td>closure</td>
</tr>
</tbody>
</table>

## Phone to Viseme Mapping

### Vowel/Diphthongs

- aa: ae, eh
- ah: ao
- aw: ax, ih, iy
- axr: ay
- fr: ey
- ix: ow
- oy: uh
- uw: ux

### Consonants

- b, p: bcl, m, pcl
- dh, epi: dx, nx, q
- en: hh
- jh: ng
- s, sh, z: th
- y: zh
- d, dcl, g, gc, k, kcl, l, n, t, tcl
- q: glottal stop
- r: red
- s: sis
- sh: shoe
- t: tot
- tcl: t-closure
- th: thief
- v: very
- w: wet
- y: get
- z: zoo
- zh: measure
- ep: epithetic
- h#: closure
<table>
<thead>
<tr>
<th>Viseme_select</th>
<th>phonemes</th>
<th>example</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>none</td>
<td>na</td>
</tr>
<tr>
<td>1</td>
<td>p, b, m</td>
<td>put, bed, mill</td>
</tr>
<tr>
<td>2</td>
<td>f, v</td>
<td>far, voice</td>
</tr>
<tr>
<td>3</td>
<td>T, D</td>
<td>think, that</td>
</tr>
<tr>
<td>4</td>
<td>t, d</td>
<td>tip, doll</td>
</tr>
<tr>
<td>5</td>
<td>k, g</td>
<td>call, gas</td>
</tr>
<tr>
<td>6</td>
<td>tS, dZ, S</td>
<td>chair, join, she</td>
</tr>
<tr>
<td>7</td>
<td>s, z</td>
<td>sir, zeal</td>
</tr>
<tr>
<td>8</td>
<td>n, l</td>
<td>lot, not</td>
</tr>
<tr>
<td>9</td>
<td>r</td>
<td>red</td>
</tr>
<tr>
<td>10</td>
<td>A:</td>
<td>car</td>
</tr>
<tr>
<td>11</td>
<td>e</td>
<td>bed</td>
</tr>
<tr>
<td>12</td>
<td>I</td>
<td>tip</td>
</tr>
<tr>
<td>13</td>
<td>O</td>
<td>top</td>
</tr>
<tr>
<td>14</td>
<td>U</td>
<td>book</td>
</tr>
</tbody>
</table>

**Visual Lipreading**
## Facial Expressions

- **Joy**
  - The eyebrows are relaxed. The mouth is open, and mouth corners pulled back toward ears.

- **Sadness**
  - The inner eyebrows are bent upward. The eyes are slightly closed. The mouth is relaxed.

- **Anger**
  - The inner eyebrows are pulled downward and together. The eyes are wide open. The lips are pressed against each other or opened to expose teeth.

<table>
<thead>
<tr>
<th>Facial Expressions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fear</strong></td>
</tr>
<tr>
<td>- The eyebrows are raised and pulled together. The inner eyebrows are bent upward. The eyes are tense and alert.</td>
</tr>
<tr>
<td><strong>Disgust</strong></td>
</tr>
<tr>
<td>- The eyebrows and eyelids are relaxed. The upper lip is raised and curled, often asymmetrically.</td>
</tr>
<tr>
<td><strong>Surprise</strong></td>
</tr>
<tr>
<td>- The eyebrows are raised. The upper eyelids are wide open, the lower relaxed. The jaw is open.</td>
</tr>
</tbody>
</table>
FACIAL EXPRESSIONS

RAISE EYE BROWS
SMILE

DISGUST
ANGER
MPEG-4 Decoder

System Layer

- Text/Image decoding MPEG-4
- 2-D/3-D geometry
- Cashed Data textures, FAPs
- Audio synthesizer/processing
- Audio decoder

Display

User input

System Layer compositing rendering

MPEG-7
• MPEG-7 will specify a standard set of descriptors that can be used to describe various types of multimedia information.
  – Descriptors
  – Description Scheme
  – Description Definition Language (DDL)

• MPEG-7 represents information about the content, not the content itself (“the bits about the bits”)

Description generation → MPEG-7 descriptor → Encoder → Coded MPEG-7 description → Decoder → User

Search Query → Agents
Different Types of Features

- Lower abstraction level
  - shape
  - size
  - texture
  - color
  - movement
  - position (where in the scene can the object be found)
Different Types of Features

• Audio
  – key
  – mood
  – tempo
  – tempo changes
  – position in sound space

• Highest Level Abstraction (semantic)
  – “This is a scene with a barking brown dog on the left and a blue ball that falls down on the right, with the sound of passing cars in the background.”
Other Type of Information

• The form
  – coding scheme (JPEG, MPEG-2)
  – size
• Conditions for accessing the material
• Links to other relevant material
• The context (e.g. Olympic 1996)

Search

• MPEG-7 data will be used to answer user queries.
• Music
  – Play a few notes on a keyboard and get in return a list of musical pieces containing required tune or images somehow matching the notes, e.g., in terms of emotions.
Search

- Graphics
  - Draw a few lines on a screen and get in return a set of images containing similar graphics, logos, ideograms,..
- Image
  - Define objects, including color patches or textures and get in return examples among which you select the interesting objects to compose your image.

Search

- Movement
  - On a given set of objects, describe movements and relations between objects and get in return a list of animations fulfilling the described temporal and spatial relations.
- Scenario
  - On a given content, describe actions and get a list of scenarios where similar actions happen.
Search

• Voice
  – Using an excerpt of Pavarotti’s voice, and getting a list of Pavarotti’s records, video clips, where Pavarotti is singing or video clips where Pavarotti is present

MPEG-4

• Go to http://www.cseleit.it/mpeg