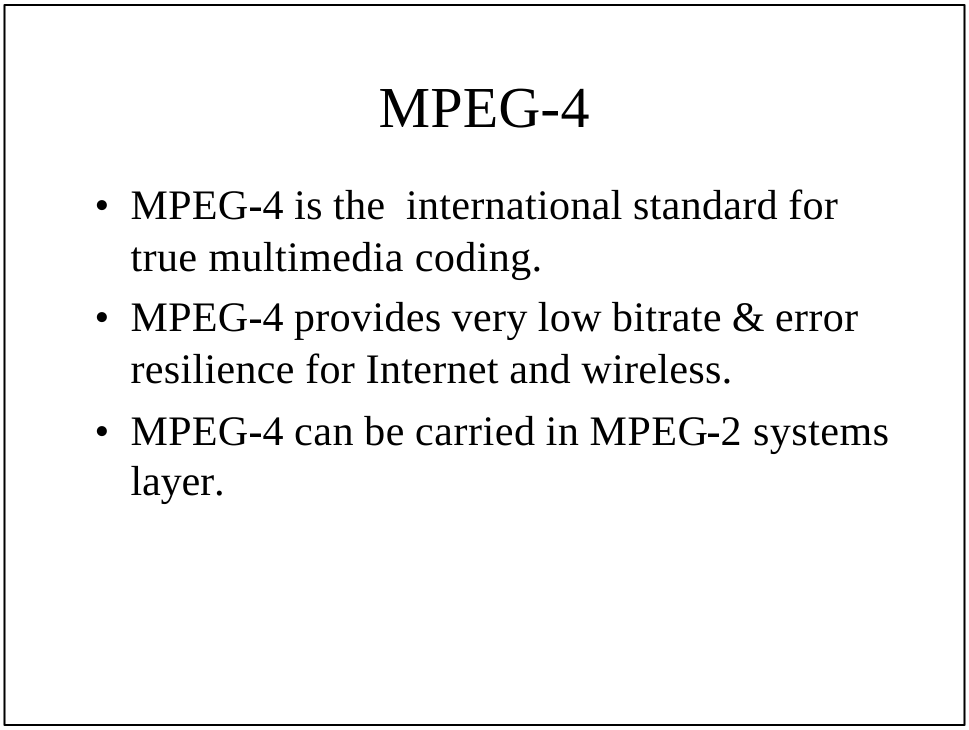
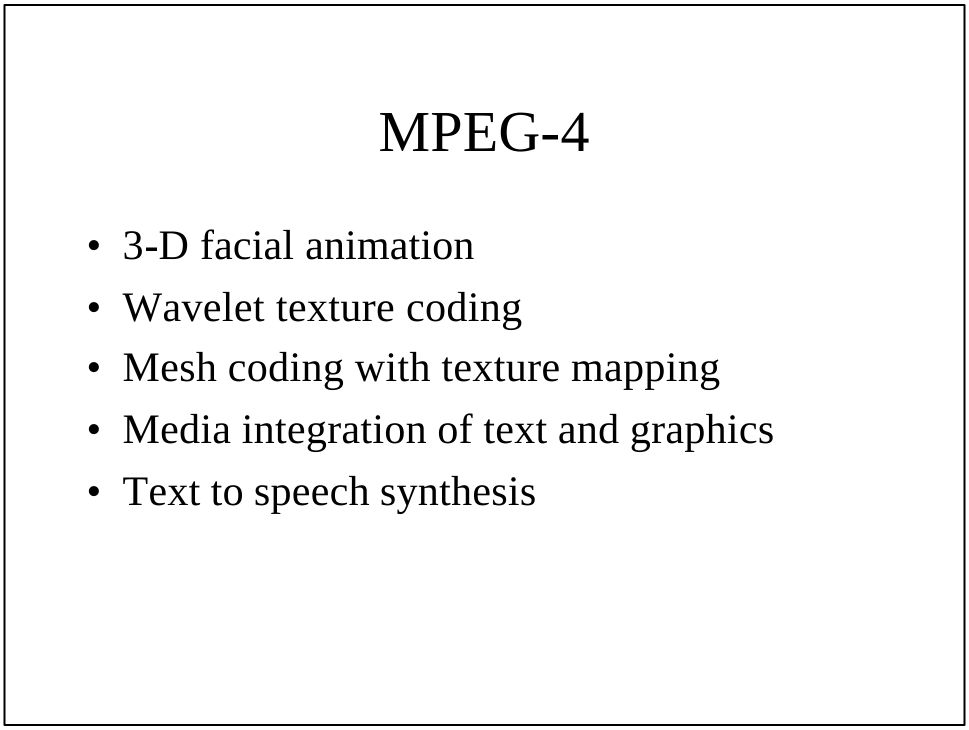
Lecture-20

MPEG-4



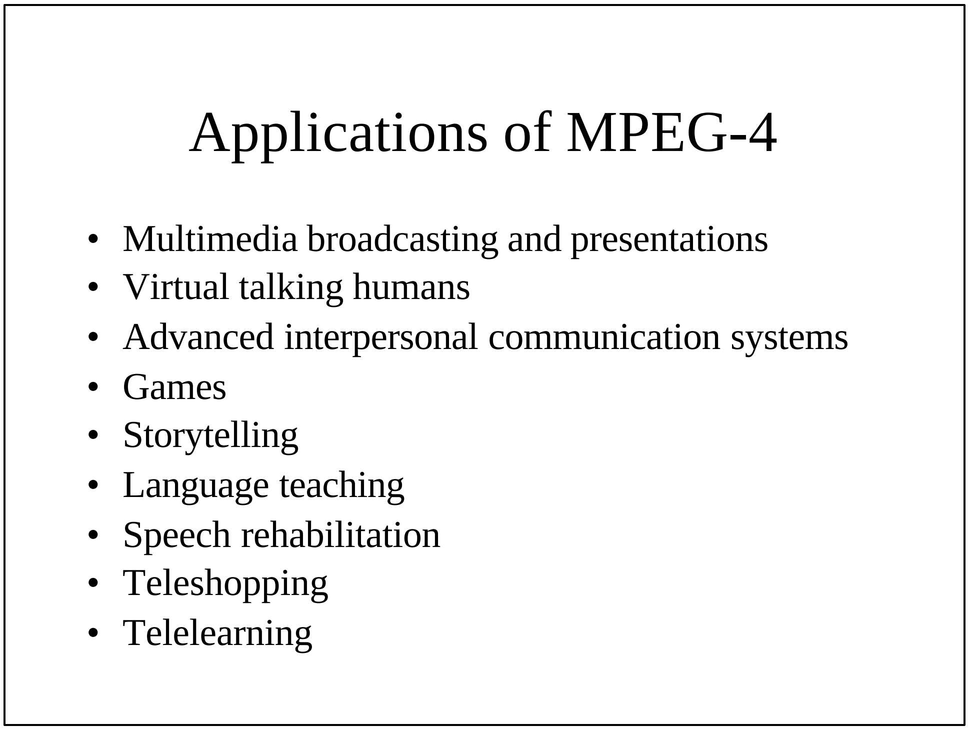
MPEG-4

- 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.



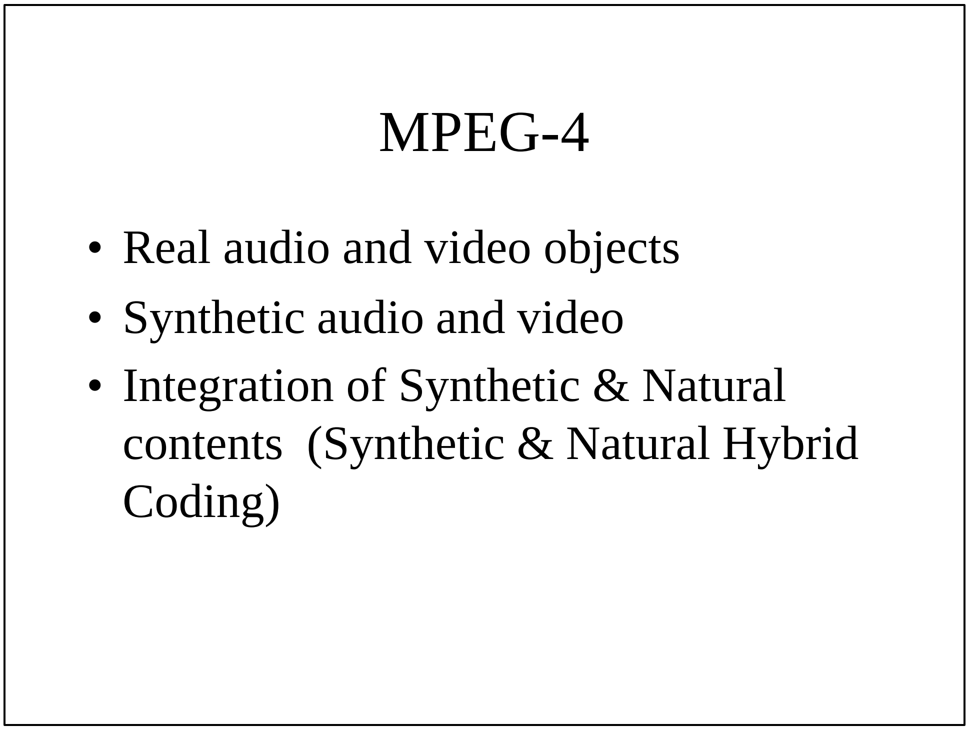
MPEG-4

- 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.

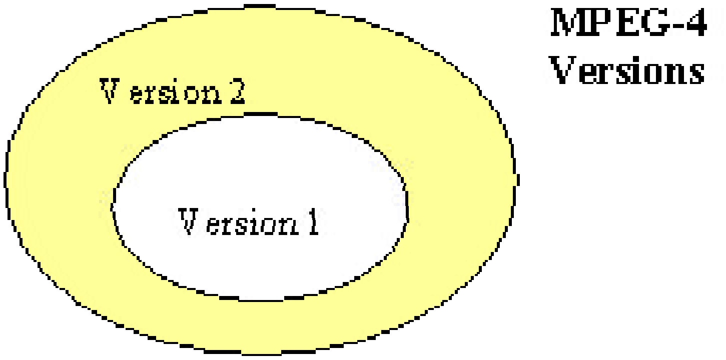
Scope & Features of MPEG-4

- Authors
  - reusability
  - flexibility
  - content owner rights
- Network providers
- End users

Media Objects

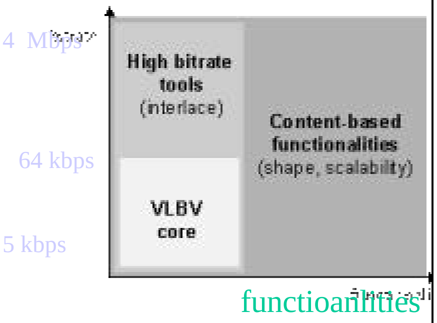
- Primitive Media Objects
- Compound Media Objects
- Examples
  - Still Images (e.g. fixed background)
  - Video objects (e.g., a talking person-without background)
  - Audio objects (e.g., the voice associated with that person)
  - etc

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


4 Mbps
64 kbps
5 kbps
High bitrate tools (interlace)
VLBV core
Content-based functionalities (shape, scalabilty)
functioanlities

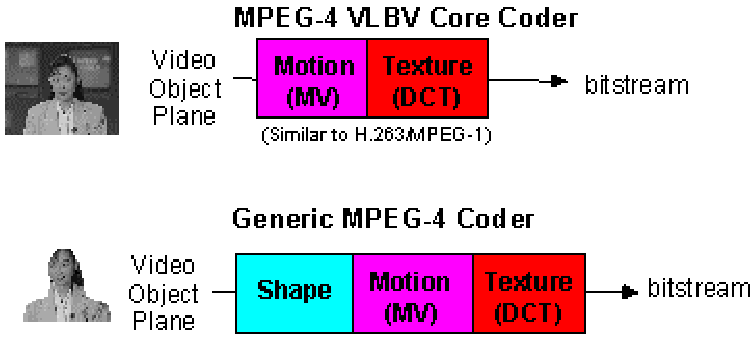
High Bitrate
1. Higher resolution
2. Higher frame rate
3. Interlaced video

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



MPEG-4 VLBV Core Coder

Video Object Plane

Motion (MV)

Texture (DCT)

bitstream

(Similar to H.263/MPEG-1)

Generic MPEG-4 Coder

Video Object Plane

Shape

Motion (MV)

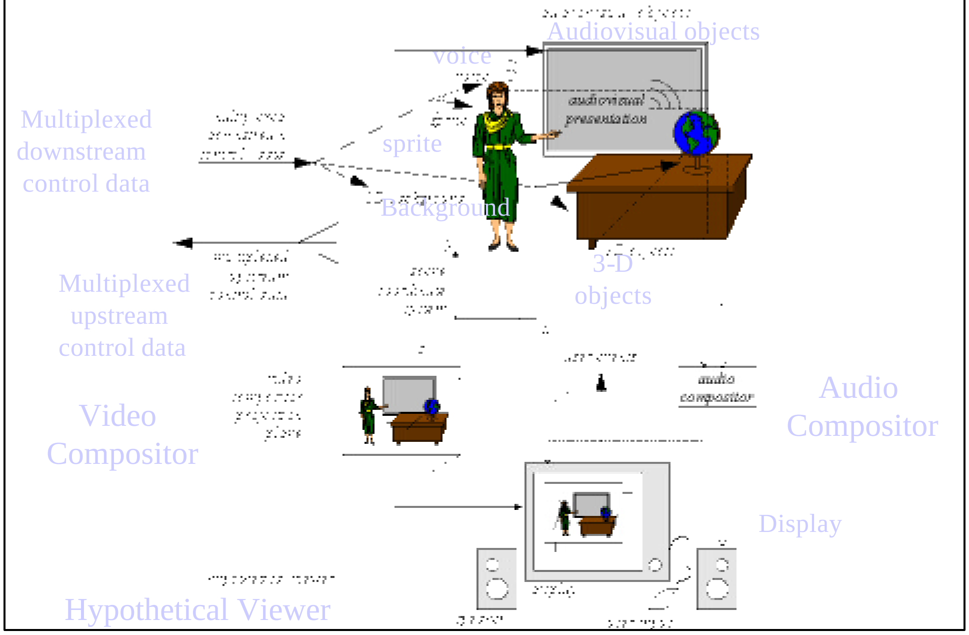
Texture (DCT)

bitstream

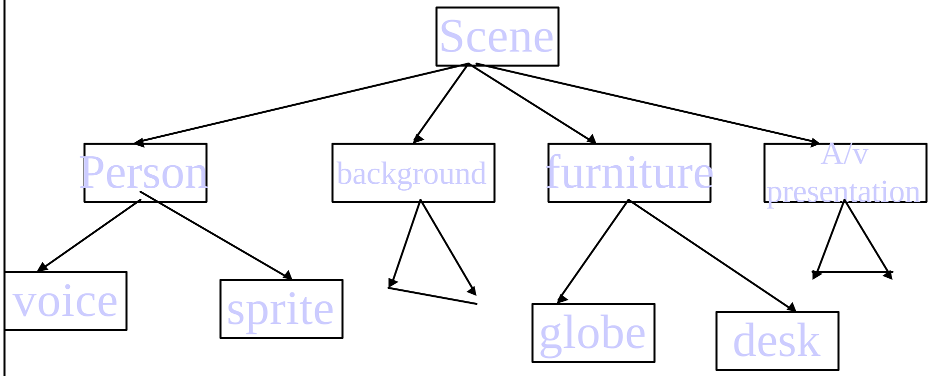
Object Composition

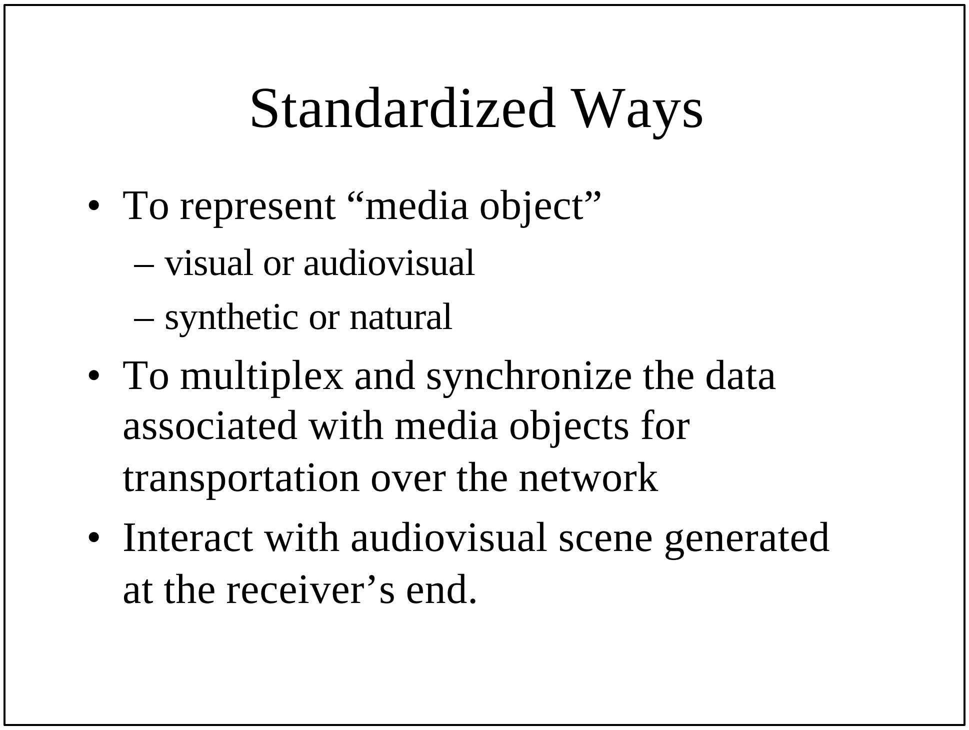
- Objects are organized in a scene graph.
- VRML based binary format BIF is used to specify scene graph.
- 2-D and 3-D objects, transforms and properties are specified.
- MPEG-4 allows objects to be transmitted once, and displayed repeatedly in the scene after transformations.

MPEG-4 Scene



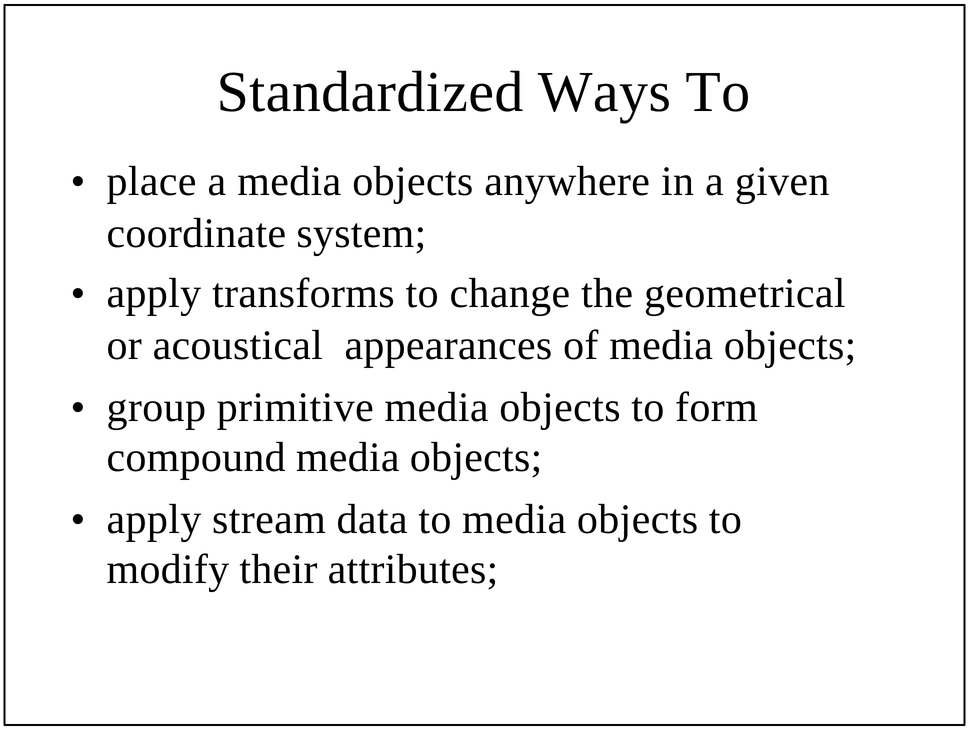
Scene Graph





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.



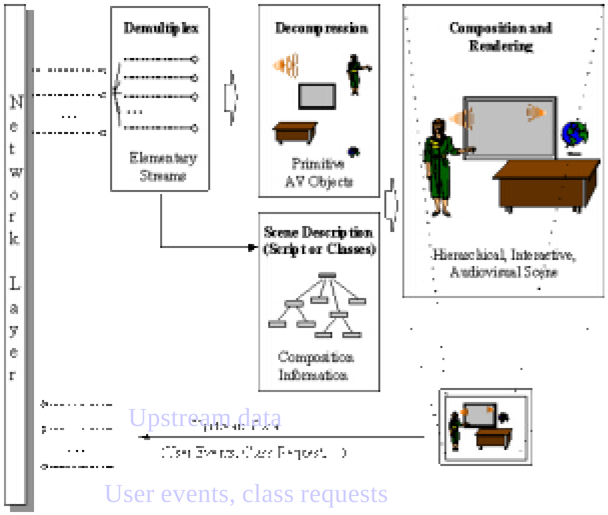
Standardized Ways To

- 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)

MPEG-4 Terminal


N e t w o r k L a y e r
Demultiplex
Elementary Streams
Decompression
Primitive AV Objects
Composition and Rendering
Hierarchical, Interactive, Audiovisual Scene
Scene Description (Script or Classes)
Composition Information
Upstream data
Upstream Data
(User Events, Class Requests)
User events, class requests

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

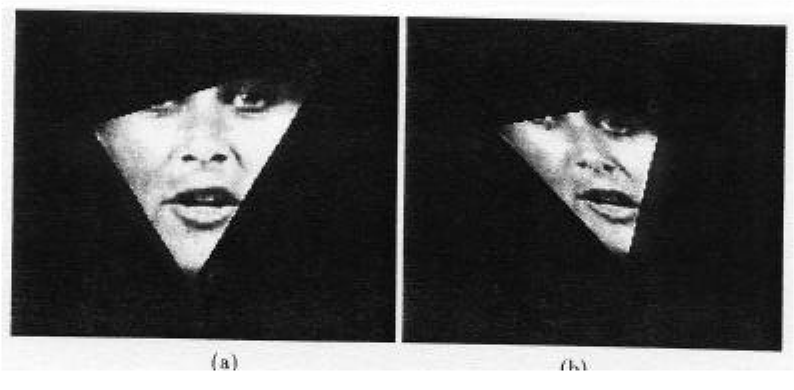
Textures, Images and Video

- 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

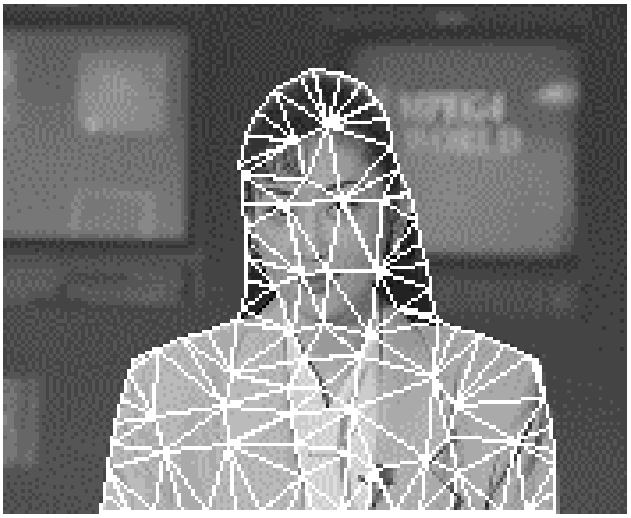
2-D Animated Meshes

- 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



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

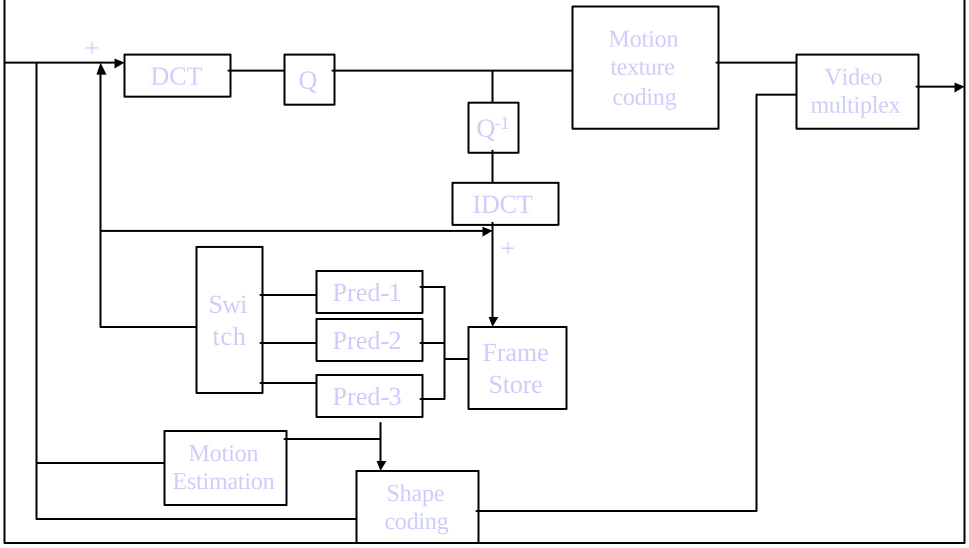
2-D Mesh Representation of Video Object

- 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

MPEG-4 Video Coder

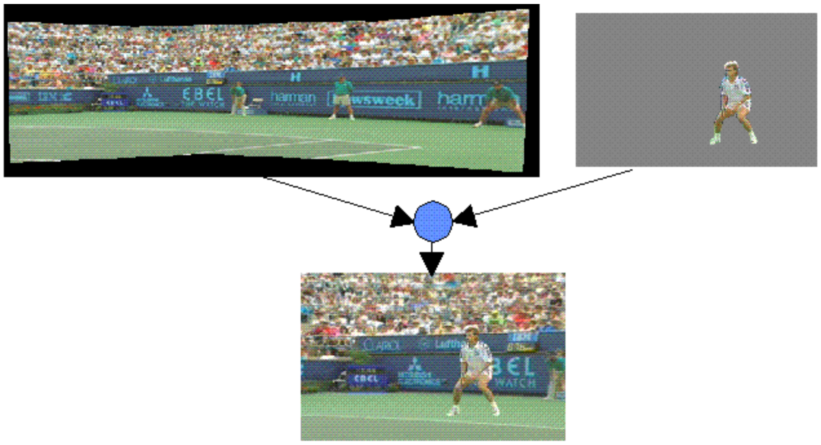


Sprite Panorama

- First compute static “sprite” or “mosaic”
- Then transmit 8 or 6 global motion (camera) parameters for each frame to reconstruct the fame 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

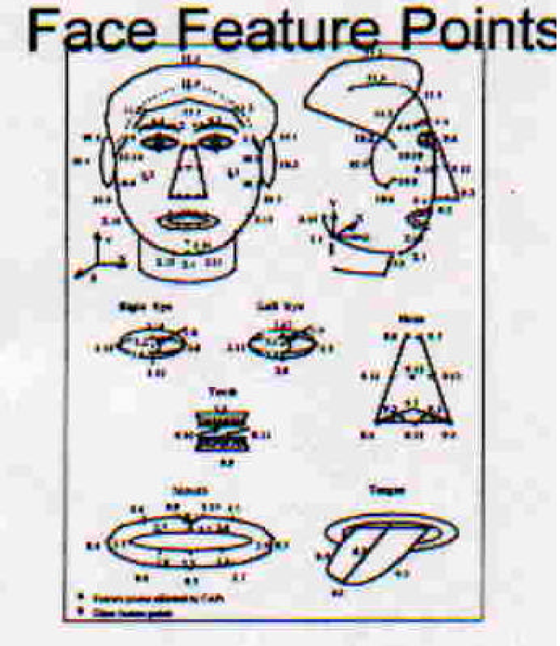
- Face animation is in MPEG-4 version 1.
- Body animation is in MPEG-4 version 2.
- Face animation parameters displace feature points from neutral position.
- Body animation parameters are joint angles.
- Face and body animation parameter sequences are compressed to low bit rate.
- Facial expressions: 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 Feature Points

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

Group	FAPS
Visemes & expressions	2
jaw, chin, inner lower-lip, corner lip, mid-lip	16
eyeballs, pupils, eyelids	12
eyebrow	8
cheeks	4
tongue	5
head rotation	3
outer lip position	10
nose	4
ears	4

FAPS

- 31: raise_l_I_eyebrow (vertical displacement of left inner eyebrow)
- 32: raise_r_I_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

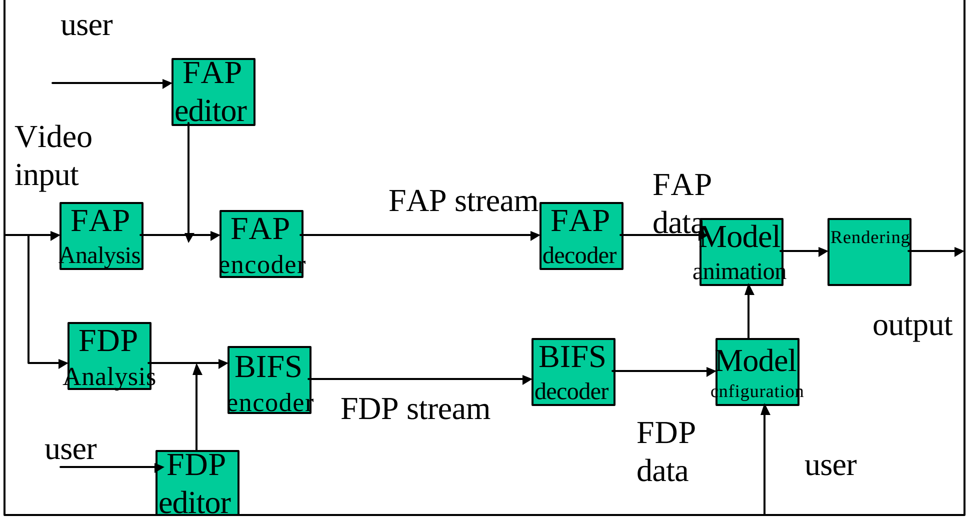
- No FAPs are coded for the corresponding group
- A mask is given indicating which FAPs in the corresponding group are coded. FAPs not coded, retain their previous values
- A mask is given indicating which FAPs in the corresponding group are coded. The decoder should interpolate FAPs not selected by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


user

FAP editor

Video input

FAP Analysis

FAP encoder

FAP stream

FAP decoder

FAP data

Model animation

Rendering

output

FDP Analysis

BIFS encoder

FDP stream

BIFS decoder

Model cnfiguration

user

FDP editor

FDP data

user

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

Phone	Example	Phone	Example	Phone	Example	Phone	Example
aa	cot	ow	boat	g	gag	q	glottal stop
ac	bat	oy	boy	gcl	g-closure	r	red
ah	butt	oy	boy	hh	hay	s	sis
ao	about	uh	book	hv	Leheigh	sh	shoe
aw	bough	uw	boot	jh	judge	t	tot
ax	the	ux	beauty	k	kick	tcl	t-closure
axr	diner	b	bob	kcl	k-closur	th	thief
ay	bite	bcl	b-closure	l	led	v	very
eh	bet	ch	church	m	mom	w	wet
er	birrd	d	dad	n	non	y	yet
ey	bait	dcl	d-closure	ng	sing	z	zoo
ih	bit	dh	they	nx	flapped-n	zh	measure
ix	roses	dx	butter	p	pop	epi	epithetic closure
iy	beat	en	button	pcl	p-closur	h#	silence
		f	fief				

Phone to Viseme Mapping

Vowel/Diphthongs

aa
ah
aw
axr
fr
ix
oy
uw

ae, eh
ao
ax,ih,iy
ay
ey
ow
uh
ux

Consonants

b,p
dh,epi
en
jh
s,sh,z
y
d,dcl,g,gc,k,kcl,l,n,t,tcl

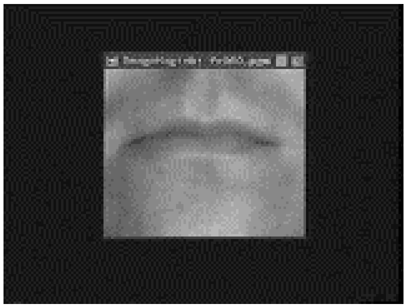
bcl,m,pcl
dx,nx,q
hh
ng
th
zh

ch
f,v
hv
r
w
h#

MPEG-4 Visems

Viseme_select	phonemes	example
0	none	na
1	p, b, m	put, bed, mill
2	f, v	far, voice
3	T, D	think, that
4	t, d	tip, doll
5	k, g	call, gas
6	tS, dZ, S	chair, join, she
7	s, z	sir, zeal
8	n, l	lot, not
9	r	red
10	A:	car
11	e	bed
12	I	tip
13	O	top
14	U	book

Visual Lipreading



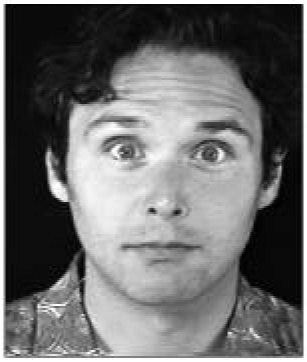
Facial Expressions

- 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.

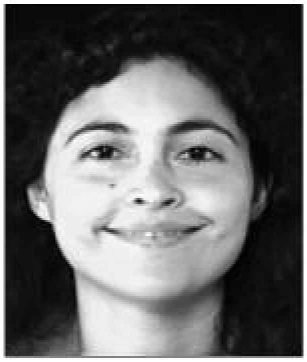
Facial Expressions

- Fear
  - The eyebrows are raised and pulled together. The inner eyebrows are bent upward. The eyes are tense and alert.
- Disgust
  - The eyebrows and eyelids are relaxed. The upper lip is raised and curled, often asymmetrically.
- Surprise
  - The eyebrows are raised. The upper eyelids are wide open, the lower relaxed. The jaw is open.

FACIAL EXPRESSIONS

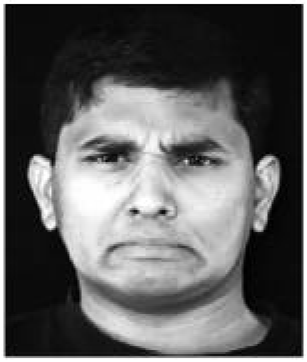


RAISE EYE BROWS

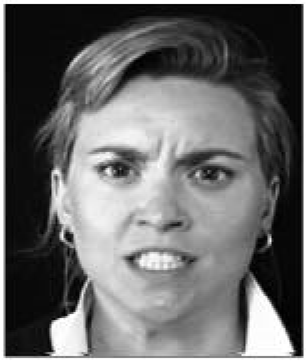


SMILE

FACIAL EXPRESSIONS

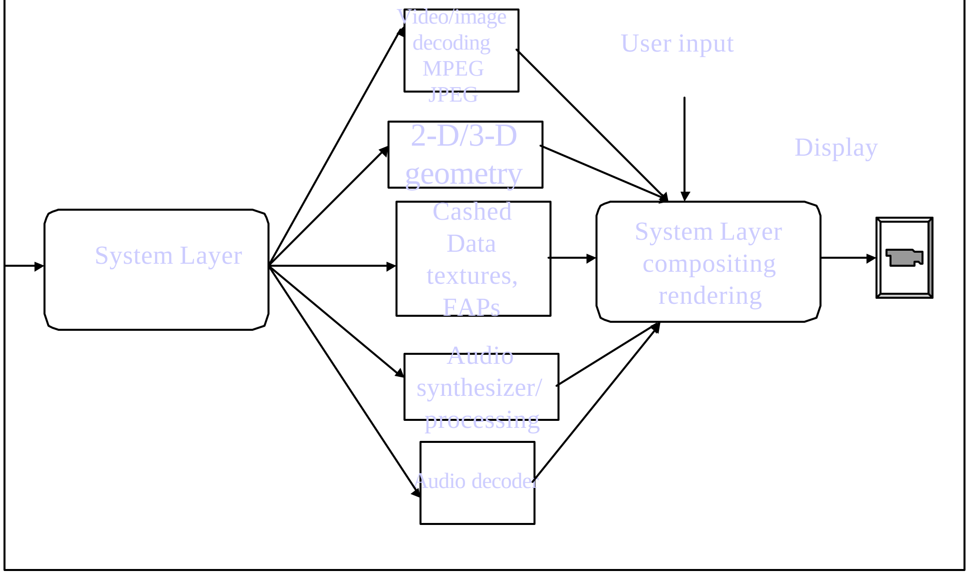


DISGUST



ANGER

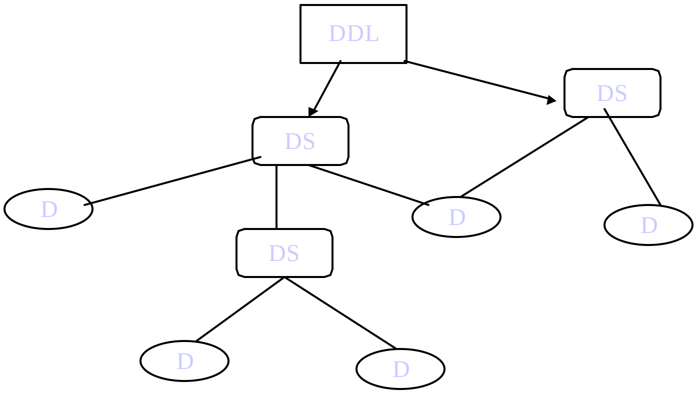
MPEG-4 Decoder

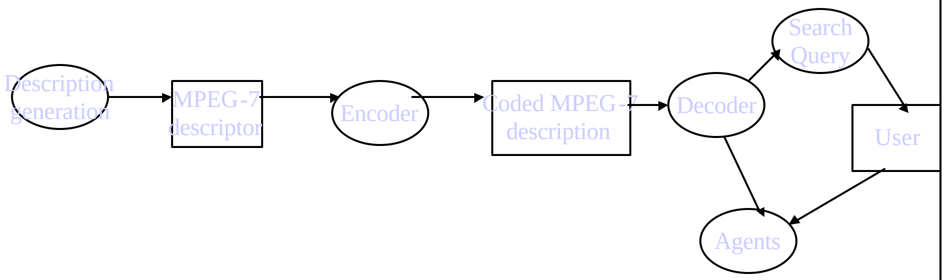


MPEG-7

- 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

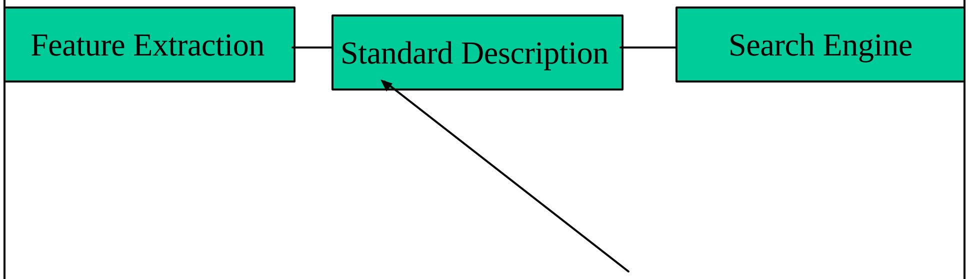
Coded MPEG-7 description

Decoder

Search Query

User

Agents



Feature Extraction

Standard Description

Search Engine

Scope of MPEG-7

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

Different Types of Features

- 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

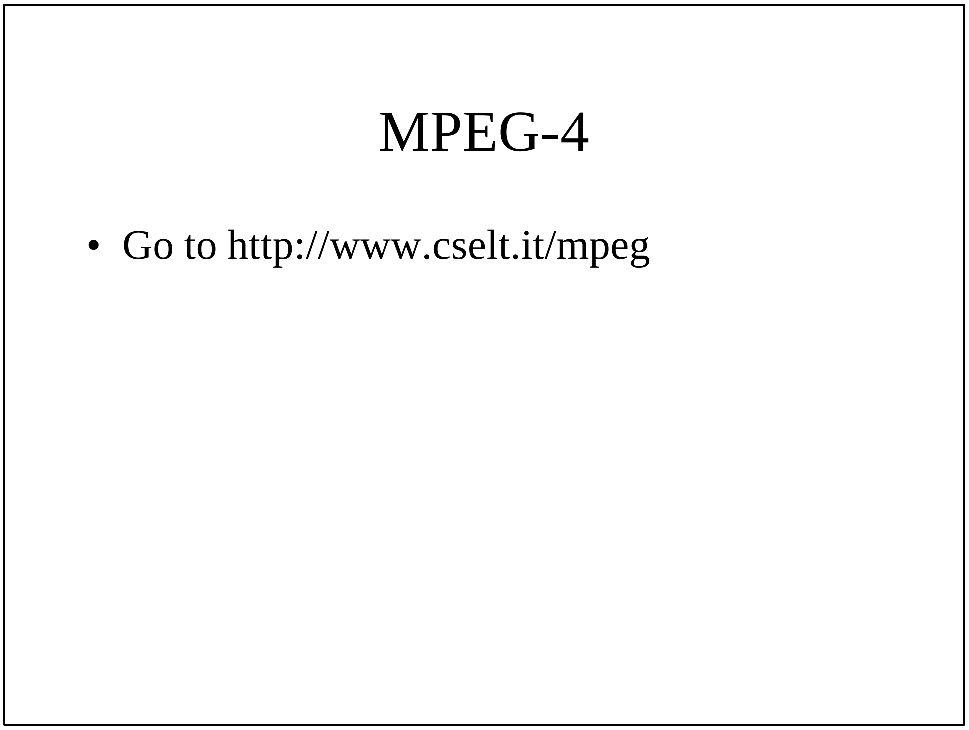
- Graphics
  - – Draw a few lines on a screen and get in return a set of images containing similar graphics, logos, ideograms,..
- Image
  - – Define objects, including color patches or textures and get in return examples among which you select the interesting objects to compose your image.

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.cselt.it/mpeg