MPEG-4

By: Kevin Brunner, Evan Spillane, Jack Cobb, Harjas Ahuja, and David Shapiro

Department of Computer Science, Virginia Tech, Blacksburg, VA
CS 4624, Multimedia, Hypertext, and Information Access
February 19, 2015

Introduction

- Increasing number of mobile users and increasing need for mobile communication
- The need for data, images, video, and other multimedia to be sent to our phones continues to expand
- Multimedia, especially video, is highly bandwidth intensive
- In order to make mobile multimedia realistically possible, we must efficiently compress it
    - Difficult since its goal is to reduce redundancy of signals, while robust delivery requirements require redundancy

Quick MPEG History

- This need to compress multimedia videos introduces the need for MPEG
- MPEG is a popular audio and video compression method
- MPEG-4 is the third MPEG standard, coming after MPEG-1 and MPEG-2
- MPEG standards are decoding standards which specify the semantics of the decoding process and the bitstream representation

Quick MPEG History

- The original purpose of the creation of the third standard (MPEG-4) was:
  - High efficiency
  - Limited complexity audio-visual videophone scenes
  - Very low bit-rates
- The scope was expanded in 1994 since there was only a moderate likely increase in compression and a need for greater functionality within certain bit ranges not supported by past standards
- This was due to three important trends:
  - Wireless communications
  - Interactive computer applications
  - Integration of audio-visual data into a number of applications
- Working draft (November 1996) to international standard (January 1999)

Improvements of MPEG-4

- In comparison, MPEG-4 video standard provides universal accessibility, including robustness in error prone environments
- Provides solutions for coding of natural or synthetic video and video
- Provides system for multiplex/demultiplex and description of scenes in a adjustable manner
- Multiple versions for ongoing specification revisions and work
- Designed to be a true encompassing multimedia standard
- Mobile indoor applications vs mobile outdoor applications
  - Indoor - lower mobility and higher bandwidth (1 megabit/second or more)
  - Outdoor - higher mobility and lower bandwidth (1-10 kilobits/second)
  - MPEG-4 is optimized for both, with bit-rates ranging from about 10 kilobits/second to around 1.5 megabits/second or higher

ITU-T Standards

- ITU-T developed standards for video coding, audio coding, and multiplex
- The related ITU-T standard for MPEG-4 is H.263
- H.263 uses motion compensated discrete cosine transform (DCT) framework, with accuracy of half-pixel
- Partitions each picture into macroblocks, consisting of 16 x 16 luminance block and the corresponding 8 x 8 chrominance blocks
- DCT coding reduces spatial redundancy and motion compensation reduces temporal redundancy, allowing for compression

Overview

- MPEG-4 covers the intersection of telecommunications, computer, and TV/film
- These types of media had been previously completely separated
- Geared towards:
  - Internet and intranet video
  - Wireless video
  - Video databases
  - Interactive home shopping
  - Video e-mail
  - Home movies
  - Virtual reality games
  - Simulation and training

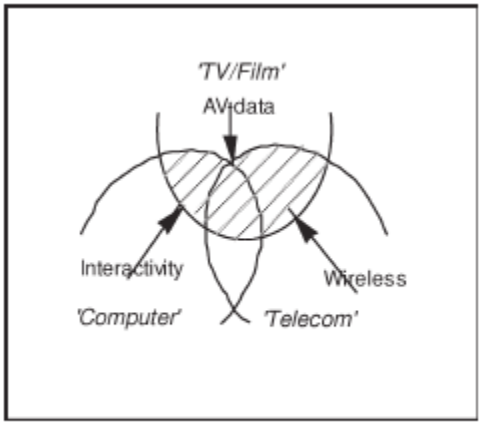


Figure 2. Applications areas addressed by MPEG-4 (shaded region).

Requirements

Table 2
Functionalities expected to be supported by MPEG-4 Version 1.

Content-based interactivity
Hybrid natural and synthetic data coding: The ability to code and manipulate natural and synthetic objects in a scene including decoder controllable methods of compositing of synthetic data with ordinary video and audio, allowing for interactivity.
Improved temporal random access: The ability to efficiently access randomly in a limited time and with fine resolution parts (frames or objects) within an audio-visual sequence. This also includes the requirement for conventional random access.
Content-based manipulation and bitstream editing: The ability to provide manipulation of contents and editing of audio-visual bitstreams without the requirement for transcoding.
Universal access
Robustness in error prone environments: The capability to allow robust access to applications over a variety of wireless and wired networks and storage media. Sufficient robustness is required, especially, for low bit-rate applications under severe error conditions.
Content-based scalability: The ability to achieve scalability with fine granularity in spatial, temporal or amplitude resolution, quality or complexity. Content based scaling of audio-visual information requires these scalabilities.
Compression
Improved coding efficiency: The ability to provide subjectively better audio-visual quality at bit-rates compared to existing or emerging video coding standards.

Video Tests

Table 3
List of MPEG-4 first evaluation formal tests and their explanation.

Compression
Class A sequences at 10, 24 and 48 kbit/s: Coding to achieve the highest compression efficiency. Input video resolution is CCIR-601 and although any spatial and temporal resolution can be used for coding, the display format is CIF on a windowed display. The test method employed is SS.
Class B sequences at 24, 48 and 112 kbit/s: Coding to achieve the highest compression efficiency. Input video resolution is CCIR-601 and although any combination of spatial and temporal resolutions can be used for coding, the display format is CIF on a windowed display. The test method employed is SS.
Class C sequences at 320, 512 and 1024 kbit/s: Coding to achieve the highest compression efficiency. Input video resolution is CCIR-601 and although any combination of spatial and temporal resolution can be used for coding, the display format is CCIR-601 on a full display. The test method employed is DSCQS.
Error robustness
Error resilience at 24 kbit/s for Class A, 48 kbit/s for Class B, and 512 kbit/s for Class C: Test with high random bit error rate (BER) of 10^{-3}, multiple burst errors with 3 bursts of errors with 50% BER within a burst, and a combination of high random bit errors and multiple burst errors. The display format for Class A and Class B sequences is CIF on a windowed display and for Class C sequences is CCIR-601 on full display. The test method employed for Class A and Class B is SS and that for Class C is DSCQS.
Error recovery at 24 kbit/s for Class A, 48 kbit/s for Class B and 512 kbit/s for Class C: Test with long burst errors of 50% BER within a burst and a burst length of 1 to 2 seconds. Display format for Class A and Class B is CIF on a windowed display and Class C is CCIR-601 on full display. The test method employed for Class A and Class B is SS and that for Class C is DSCQS.
Scalability
Object scalability at 48 kbit/s for Class A, 320 kbit/s for Class E, and 1024 kbit/s for Class B/C sequences: Coding to permit dropping of specified objects resulting in remaining scene at lower then total bit-rate; each object and the remaining scene is evaluated separately by experts. The display format for Class A is CIF on a windowed display and for Class B/C and Class E is CCIR-601 on a full display. The test method employed for Class A is SS, for Class B/C is DSCQS, and for Class E is DSIS.
Spatial scalability at 48 kbit/s for Class A, and 1024 kbit/s for Class B/C/E sequences: Coding of a scene as two spatial layers with each layer using half of the total bit-rate, however, full flexibility in choice of spatial resolution of objects in each layer is allowed. The display format for Class A is CIF on a windowed display and that for Class B/C/E is CCIR-601 on a full display. The test method employed for Class A is SS, and that for Class B/C/E is DSCQS.
Temporal scalability at 48 kbit/s for Class A, and 1024 kbit/s for Class B/C/E sequences: Coding of a scene as two temporal layers with each layer using half of the total bit-rate, however, full flexibility in choice of temporal resolution of objects in each layer is allowed. The display format for Class A is CIF on a windowed display and that for Class B/C/E is CCIR-601 on a full display.

Audio Tests

Three classes of audio test sequences, Class A, B and C were identified:
- Class A: Single source sequences consisting of a clean recording of a solo instrument.
- Class B: Single source with background sequences consisting of a person speaking with background noise.
- Class C: Complex sequences consisting of an orchestral recording.

A number of bit-rates such as 2, 6, 16, 24, 40 and 64 kbit/s were selected for testing of audio/speech.

For specific bit-rates, some candidates outperformed the reference coding schemes, although for all combinations tested, no single scheme was the clear winner.

SNHC Tests

- The SNHC group started its work much later than the video group.

- Its focus was primarily on coding for storage and communication of 2D and 3D scenes involving synthetic images, sounds, and animated geometry and its integration into scenes that contain coded natural images/video and sound.

Video Development

- Video development was started by identifying a number of needed tools and the options available for each tool as well as a reference framework.
- 40 experiments were defined and categorized into the following groups:
  - Coding efficiency
  - Shape and object texture coding
  - Robust coding
  - Multifunctional coding
- A reference coding framework known as the first Verification Model (VM1) was released

SNHC Development

- There have been a total of four iterations of SNHC VM, from VM1 to VM4.
- More mature tools of SNHC VM3 have been accepted for the visual part of the MPEG-4 Version 1 standard.
- The remaining tools have been left in VM4 for consideration for the next version of MPEG-4.

Systems Development

- The Systems layer in MPEG has been traditionally responsible for integrating media components into a single system, providing multiplexing and synchronization services for audio and video streams.
- A key requirement from the System part is the capability to combine individual audiovisual objects in scenes.
- This was accomplished by using Java but performance and compliance issues soon arose.
- A three-step approach was adopted:
  - In level 0, no programmability was allowed.
  - In level 1, facilities were provided to combine different tools into algorithms.
  - In level 2 even individual tools were considered as targets for programmable behavior.

DMIF Development

- The significance of the DMIF activity has been recognized and DMIF has been given the status of a new group.
- The charter of the DMIF group is to develop standards for interfaces between Digital Storage Media (DSM), networks, servers and clients for the purpose of managing DSM resources and controlling the delivery of MPEG bitstreams and associated data.
- The ongoing work of this group is expected to result in part 6 of the MPEG-4 standard.

MPEG-4 Video Coding Basics

- Ongoing work consists of two major areas – coding of (natural) video and coding of synthetic video
- From a top-down perspective, the organization of coded MPEG-4 Video data can be described by the following class hierarchy:
  • VideoSession: Represents the highest level in the class hierarchy and simply consists of an ordered collection of Video Objects.
  • VideoObject: represents a complete scene or a portion of a scene with a semantic meaning.
  • VideoObjectLayer (VOL): represents various instantiations of a Video Object.
  • GroupOfVideoObjectPlanes (GOV): Optional access units for editing, tune-in or synchronization.
  • VideoObjectPlane (VOP): represents snapshot in time of a Video Object.

Motion Coder

- Predicted from previous frames (p frames) or bidirectionally from previous and future frames (b frames)
  - B frames are more complex
- After prediction, coder finds the residual
- Estimation and compensation is performed on 16x16 luminance block of a macroblock
  - Motion vector is specified to half pixel accuracy
  - Often not enough; individual 8x8 block for motion vectors

Scalable Video Coding

- Allows simple decoder to produce basic quality while an enhanced decoder may produce enhanced quality
- Ensures that input video data is coded into two or more layers
  - Independently coded base layer
  - One or more enhancement layers
- Supports both Temporal and Spatial scalabilities
  - Temporal: offers decoders a means to increase temporal resolution
  - Spatial: offers decoders a means to display base or enhancement layer output

Scalable Video Coding

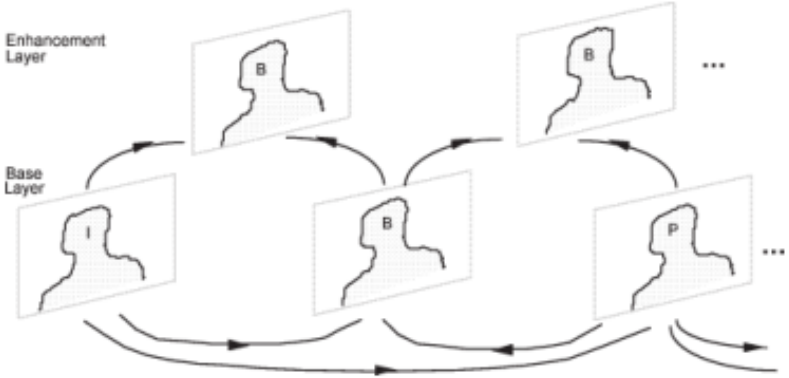


Figure 15. Temporal scalability.

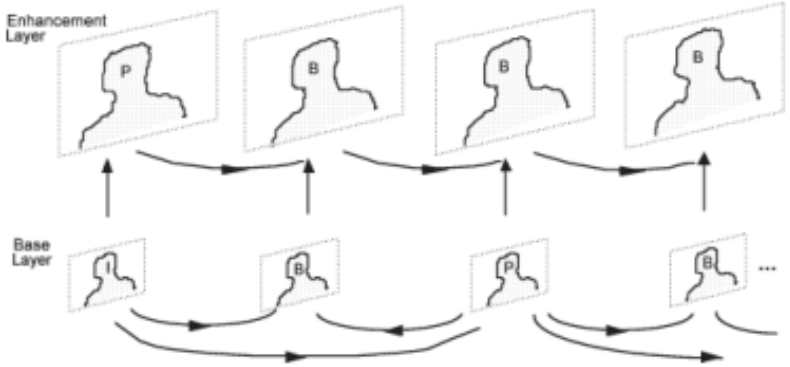


Figure 16. Spatial scalability.

- Temporal: ½ temporal resolution at base layer
- Spatial: ¼ resolution at base layer
- In reality, some flexibility is allowed in the prediction structures, so they do not always result like the ones above

Robust Video Coding

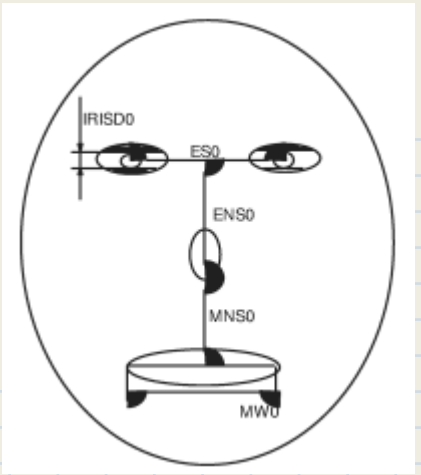
- Helps to increase error robustness
- Variety of tools available to encoder in error resilient mode
  - Resynchronization
  - Data Partitioning
  - Reversible VLCs

Data Partitioning

- Increases error resilience
  - Separates normal motion and texture blocks of a macroblock
  - Sends motion data followed by motion marker then texture data
    - Marker is unique 17 bit block not emulated by codewords
- Motion data sent for each macroblock, then texture data
- Texture data for each macro block is divided into 2 parts
  - Coded block information for luminance
  - Coded discrete cosine transform (DCT) coefficients

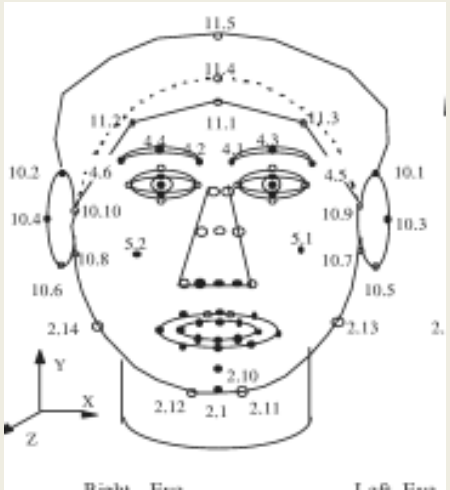
Facial Animation Coding

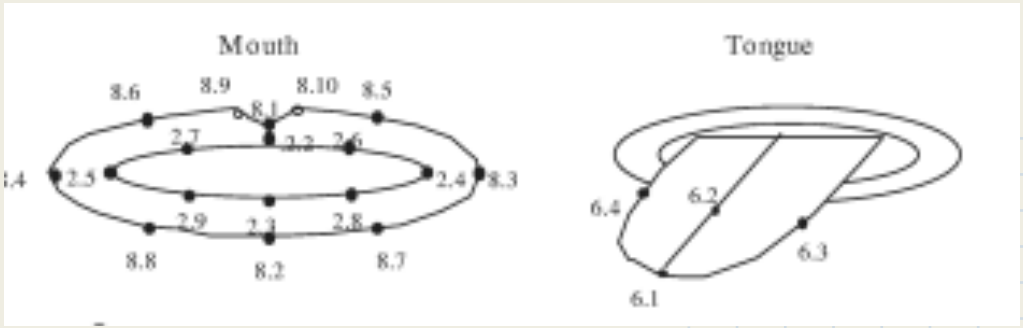
- FAPs and FDPs are set of parameters that allow animation of faces to reproduce expressions, emotions and definition of facial shape.
- The FAP set contains two high level parameters:
  Visemes: The facial expression while making any sound
  Expressions: Described by textual definitions such as joy, sadness, anger etc.
- *FAPU*: Units defined in order to interpret facial models and produce accurate results in terms of expressions and speech pronunciation
  - Eg. IRISD0, ES0, ENS0, MNS0 etc.



Facial Animation Coding

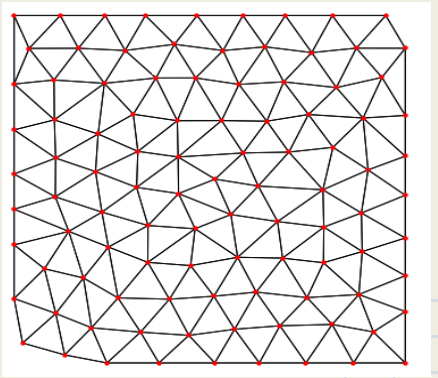
- FDPs customize the proprietary face model to a particular face. The FDP set is specified using the FDP node and supports two options:
  - Calibration info so the face can be configured using facial feature points
  - A face model is downloaded with the animation definition of the FAPs





Object Mesh Coding

- Mesh based representation is useful for natural or synthetic visual objects MPEG-4 includes a tool for 2D triangular mesh based representation.



- The vertices of the triangular patches are called nodes
- A uniform mesh can be specified by horizontal and vertical node points
- Delaunay mesh utilizes the boundary node points and then the interior node points of the mesh
- Generally, the total number of nodes and boundary nodes is encoded. The top left node is encoded, then the next clockwise boundary node is found and the difference between them is encoded. All the boundary and interior nodes are encoded in a similar fashion

Still Texture Coding

- Discrete Wavelet Transform (DWT) is used to code still image data for texture mapping as it ensures coding efficiency and continuous scalability.
- Basic modules of a zero-tree wavelet based coding scheme:
  - Decomposition of texture using DWT
  - Quantization of wavelength coefficients
  - Coding of lowest frequency subband using a predictive scheme
  - Zero-tree scanning of higher order subband wavelet coefficients
- Lowest band coefficients are quantized using a uniform mid rise quantizer, higher bands use multilevel quantization having different step sizes for each level of scalability.
- After quantization, each wavelet coefficient is zero or non-zero, the coefficients are scanned by zero-tree scanning.

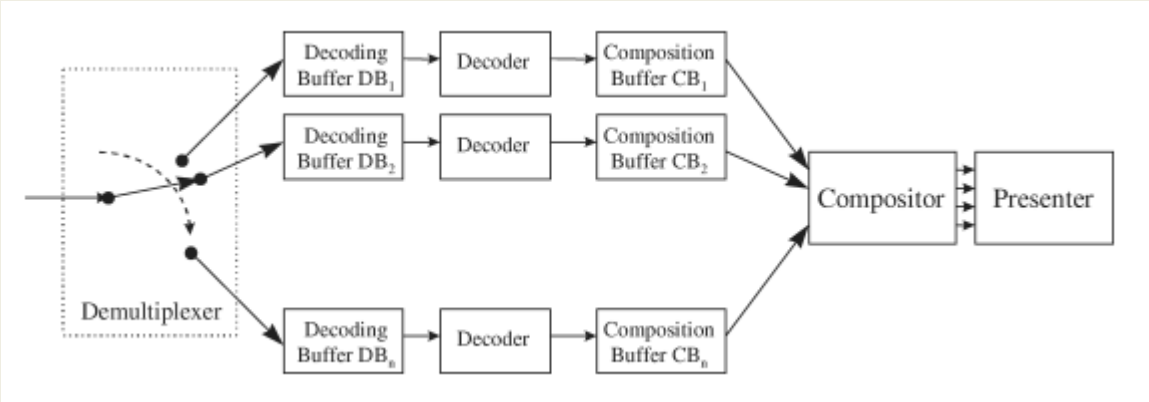
MPEG-4 Audio

- Natural Audio
  - Parametric coder: Lowest bit range between 2 and 6 kbit/s
  - CELP coder: Medium bit rates between 6 and 24 kbit/s
  - Time/frequency coder: Higher bit rates starting at about 16 kbit/s

- Text to Speech
  - Changes text into a string of phonetic symbols
  - Applications: Artificial storyteller, voice newspaper, voice internet etc.
  - Can be used for many languages, adopts concept of language code.

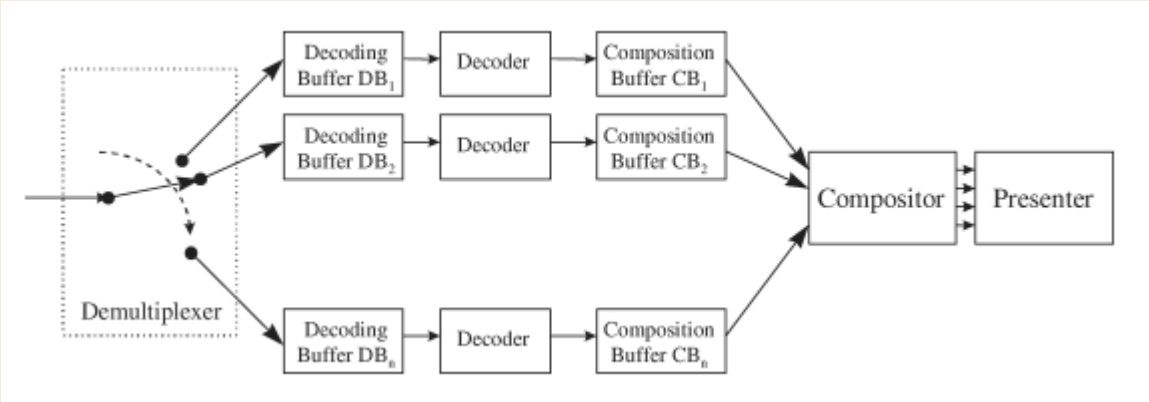
- Structured Audio
  - Use ultra low bit rate algorithmic sound models in the range 0.01-10 kbit/s
  - SAOL: signal processing language effects post production in MPEG-4

MPEG-4 System Decoder Model



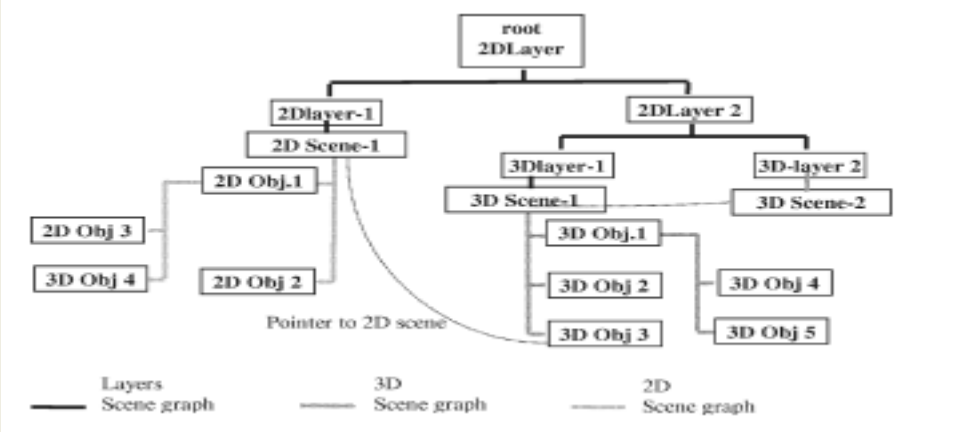
- Composed of a set of decoders for various audio and video types
- Two types of buffers
  - Decoding
  - Composition

MPEG-4 System Decoder Model



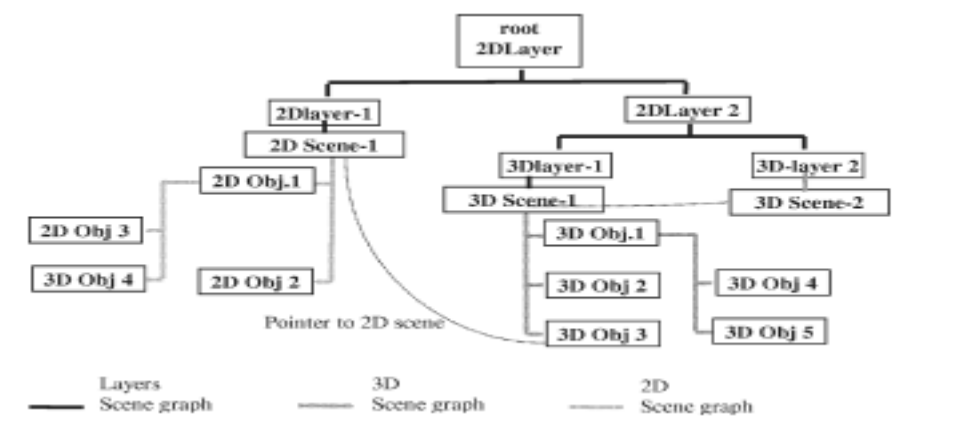
- Keeping time is very important
  - Serves to ensure that “Events” happen when content creator intended
  - So sender can control behavior of receiver
    - So receivers resources are not exhausted
    - Sender timestamps data, receiver uses timestamps to adjust clock speed

Scene Description



- New with MPEG-4
- Refers to the spatial positioning and behavior of individual objects
- Transmitted in a different scene from the audiovisual objects

Scene Description

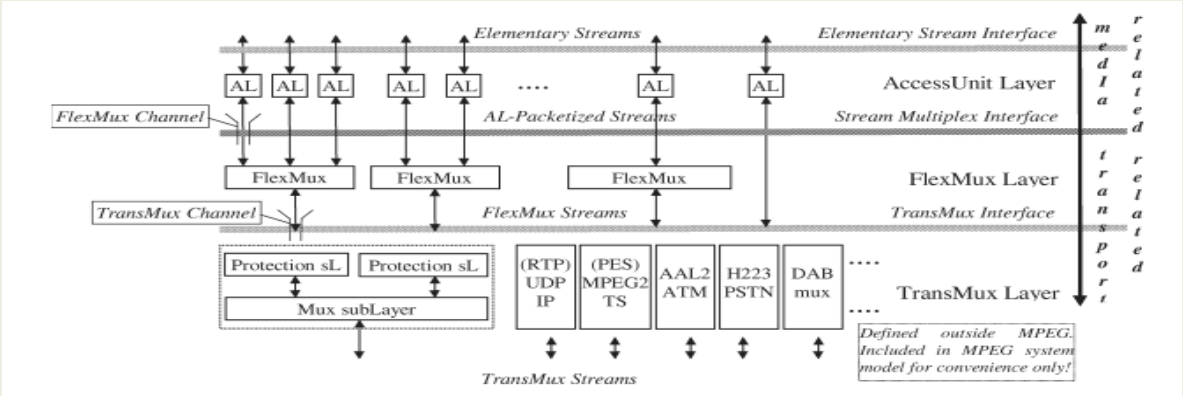


- Architecture is based on Virtual Reality Modeling Language (VRML)
- Scenes are described as a hierarchy of nodes, forming a tree
- Leafs correspond to media objects (audio or visual components)
- Intermediate nodes perform operations on children (grouping / transformations)

Multiplexing

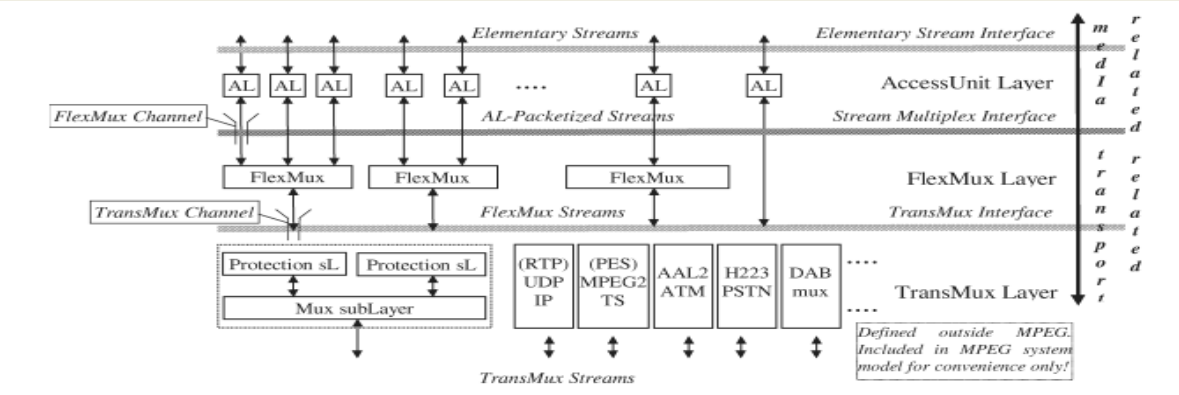
MPEG-4 can be delivered across a variety of channels

- Slow wireless streams or fast dvds



- Top most level is the Access Level Unit or AL
  - This is where the timestamping happens for receiver control
  - can be configured to include frame information, sequence numbering (useful for error prone environments)

Multiplexing cont.



- Next is the “FlexMux” layer
  - Use is optional (as seen in right side of picture)
  - Can be used as a data channel for information about objects on screen
- Finally the “TransMux” layer
  - Not specified by MPEG 4
  - Can be used as a transport layer facility

Beyond Current MPEG-4 Work

- MPEG committee has begun work on the next MPEG standard.
- It will be called MPEG 7
- Will specify a set of descriptors that will be associated with the content itself.
- Allows fast searching for material
- All content will be indexed and searchable

Questions?

References

- Atul Puri and Alexandros Eleftheriadis. 1998. MPEG-4: an object-based multimedia coding standard supporting mobile applications. Mob. Netw. Appl. 3, 1 (June 1998), 5-32. DOI=10.1023/A:1019160312366 http://dx.doi.org/10.1023/A:1019160312366