MPEG-4: Multimedia Coding Standard Supporting Mobile Multimedia System

Lian Mo, Alan Jiang, Junhua Ding

School of Computer Science
Florida International University, Miami, FL 33199

Abstract: Mobile Multimedia System is a new computing paradigm of computing with the advances in wireless networking technology and development of semiconductor technology. There are many technological challenges to establishing the computing paradigm. Supporting the mobile multimedia computing is one of the important motivations of the MPEG-4 development. In this survey, firstly, we will briefly describe the mobile multimedia system concept, current state, architecture, and its challenge techniques. The discussion of MPEG-4 is focused the technical description. Its technique is described from three folds. 1. MPEG-4 DMIF and system, which describes the multimedia content delivery integration framework, object-based representation. 2. Introducing the MPEG-4 visual core technologies allowing efficient storage, transmission and manipulation of textures, images and video data for multimedia environments. 3. Describing coding of audio objects and synthesizing sounds based on structured descriptions. Finally, based on the discussion of the mobile multimedia system and MPEG-4, an application example on MPEG-4--An MPEG-4 Based Mobile Conferencing System, is given.

1 Introduction

Recent advances in wireless networking technology and the exponential development of semiconductor technology have engendered a new paradigm of computing, called Mobile Multimedia System. Users carrying portable devices access to multimedia information from a shared infrastructure independent of their physical location. There are many technological challenges to establishing this paradigm of computing, and one of the critical issues is how to represent and exchange the audio and video information in mobile multimedia system. MPEG-4 is a new audiovisual standard, that views multimedia content as a set of audiovisual objects that are presented, manipulated and transported individually. So these objects can be flexibly and interactively used and reused. MPEG-4 is more and more used in mobile multimedia system.

Mobile multimedia system will play an important role in driving technology in the next decade. In this paradigm, the basic personal computing and communication device will be an integrated, and the information accessed or exchanged should be multimedia data. It will incorporate various functions like a pager, cellular phone, laptop computer, diary, digital camera, video game, calculator and remote control. Wireless networking is necessary and it provides mobile users with versatile communication, and permits continuous access to services and resources of the land-based network. A wireless infrastructure, mobile terminals and information providers will consist of the mobile multimedia system. However, the technological challenges to establishing this paradigm of personal mobile computing are non-trivial. The challenge is to maintain a high perceived end-to-end quality without limiting applications to the point where they are no longer useful. Multimedia networking requires at least a certain minimum bandwidth allocation for satisfactory application performance. How to provide efficient solution for video and audio compression and coding is also an important issue for mobile multimedia systems. MPEG-4 can give a solution to this kind of issues.

After setting the MPEG-1 and MPEG-2 standards, MPEG (moving Pictures Experts Group) is now working on a new audiovisual standard, called MPEG-4. The MPEG-4 version 1 and version 2 are already set, and the extended working is developing. The purpose of the MPEG-4 is to address the new demands that arise in a world in which more and more audiovisual material is exchanged in digital form, and it tries to achieve much more compression and even lower bitrates. MPEG-1 and MPEG-2 deal with ‘frame-based’ video and audio, and it provides a large improvement in randomly accessing content. But MPEG-4 is ‘object-based’ audiovisual representation, and it not only aims to achieve efficient storage and transmission, but also to satisfy other needs of image communication users. MPEG-4 makes the move towards representing the scene as a composition of objects, rather than just frame. It defines an audiovisual scene as a coded representation of ‘audiovisual objects’ that have certain relations in space and time. When audio and video objects are associated, an audiovisual object results. The

new approach to information representation allows for much more interactivity, for versatile reuse of data, and for intelligent schemes to manage bandwidth processing resources and error protection. It uses the content based coding method to ease the integration of natural and synthetic audio and video material, as well as other data types, such as text overlays and graphics. It offers a new kind of interactivity, integration of objects of different nature for multimedia systems. And it helps to access multimedia information everywhere and provides the flexibility for a fast-changing environment, which is required in the mobile multimedia system. So MPEG-4 provides more facilities and functionality making it more and more used in mobile multimedia systems.

In this paper, firstly, we will discuss mobile multimedia system, which provides an application background for MPEG-4. We will briefly describe the mobile multimedia system concept, current state, architecture, and its challenge techniques on multimedia information representation and access. After that, we will detail describe the MPEG-4, which is an object-based audiovisual representation. We will give an overview of MPEG-4, however, we will focus the technical description of MPEG-4. And we will address its technique description from three folds. 1. MPEG-4 DMIF and system, which describes the multimedia content delivery integration framework. And the object-based audiovisual representation will be introduced in this section. 2. MPEG-4 Video. In this section, we will introduce the core technologies allowing efficient storage, transmission and manipulation of textures, images and video data for multimedia environments. 3. MPEG-4 Audio. We will describe coding of audio objects and synthesizing sounds based on structured descriptions. Finally, based on the discussion of the mobile multimedia system and MPEG-4, we will give an application example on MPEG-4--An MPEG-4 Based Mobile Conferencing System.2 Mobile Multimedia System.

2 Mobile Multimedia System

Mobile terminals such as palm PC, and a communication infrastructure especially supporting the wireless communication are necessary to enable a mobile multimedia system. In

this paper, we only focus on the terminal system. Most mobile systems actually are portable computers to be equipped with wireless interfaces, allowing networked communication even mobile. So they are part of a greater computing infrastructure. The integration of multimedia applications and mobile computing will lead to a new application domain and market in the near future.

2.1 State of Arts of Mobile System

The research community and the industry have expended considerable effort toward mobile computing and the design of portable computers and communication devices. Inexpensive tools that are small enough to fit in a pocket are joining the ranks of notebook computers, cellular phones and video games. Communication, data processing and entertainment will be supported by the same mobile system and enhanced by the worldwide Internet connectivity. We first take a brief look at the various mobile systems on the market today. Some of these systems have no built-in wireless networking capability, but rather rely on an external wireless modem for wireless connectivity. The wireless modem is in general based on a cellular phone, or on wireless LAN (WLAN) products. Current mobile systems can be classified into the following categories based on their functions.

- *Pocket Computer* – Which includes the laptop, pen tablet, and handed PC. They are only a simplified computer or just a small size personal computer. The input devices for this kind of mobile system are different, such as optical pen, some buttons, or mouse etc. They are connected to the network based on WLAN or through radio modem, wire-line modem and infrared port.

- *Virtual books* – These systems have good quality displays, and a rather conventional architecture. User input is limited to a few buttons, and a pen. But most of them are only desktop companion, which means that it has to connect the network based on desktop.

- *Personal Digital Assistants (PDA)* – the PDA is generally a monolithic device without a keyboard and fits in the user's hand. Communication abilities involve a docking port or serial port

for connecting to and synchronizing with a desktop computer, and some of them already can connect to internet through wireless modem and particular internet service provider to connect with the Internet.

- *Smart phones* – They are combination devices are essentially PC-like devices attached to a cellular phone.

- *Wireless terminal* – These systems are wireless extended input and output of a desktop machine. These systems are designed to take advantage of high-speed wireless networking to reduce the amount of computation required on the portable.

It will be clear that current mobile systems are primarily either data processing terminals or communication terminals. When these devices are used as personal communication tool, it still impossible to transform multimedia information, such as video and audio at the same time. When they are used as mobile computers, it is still very difficult to deploy them in the Internet. They are still need to connect the Internet through a desktop, WLAN or particular ISP.

2.2 Mobile Multimedia System Architecture and Requirements

The future mobile multimedia system should be a small personal portable computer and wireless communications device (Here we refer the mobile multimedia system terminals) that can replace cash, check book, passport, keys, diary, phone, pager, maps and briefcases. It is a hand-held device that is resource-poor, i.e. small amount of memory, low processing power, and connected with the environment through a wireless network with variable connectivity. But it meets several major requirements: high performance, energy efficient, Quality of Service (QoS), small size, and low design complexity. It can run some simple application systems in itself, or run some complex applications via servers. And it can exchange data with a desktop, and it could be used personal communication system and exchange multimedia information. It interacts with the environment and so is part of an open distributed system. It needs to communicate with – possibly hostile – external services under varying communication and operating conditions. It

provides the communication facilities to ubiquitously access the network, and the network access should support heterogeneity in many dimensions (transport media, protocols, data-types, etc.).

To the communication infrastructure, it seems that in the future we can have several different wireless networks. Future devices and applications can be able to connect to these different networks since most of them will offer TCP/IP based services. One scenario will be an overlay network where clients can make vertical handovers to different networks.

The approach to achieve a system (mobile terminal) as described above is to have autonomous, reconfigurable modules such as network, video and audio devices, interconnected by a switch, and to offload as much as work as possible from the CPU to programmable modules that are placed in the data streams. Thus, communication between components is delivered exactly where it is needed, work is carried out where the data passes through, bypassing the memory. The amount of buffering is minimized, and if it is required at all, it is placed right on the data path, where it is needed. To support this, the operating system must become a small, distributed system with co-operating processes occupying programmable components – like CPU, DSP, and programmable logic – among which the CPU is merely the most flexibly programmable one. The interconnection of the architecture is based on a switch, called *Octopu*s, which interconnects a general-purpose processor, multimedia devices, and a wireless network interface.

The systems that are needed for multimedia applications in a mobile environment must meet different requirements than current workstations in a desktop environment can offer. The basic characteristics that multimedia systems and applications needs to support are:

- *Continuous-media data types* – Media functions typically involve processing a continuous stream of data, which implies that temporal locality in data memory accesses no longer holds. Remarkably, data caches may well be an obstacle to high performance and energy efficiency for continuous-media data types because the processor will incur continuous cache-misses.

- *Provide Quality of Service (QoS)* – Instead of providing maximal performance, systems must provide a QoS that is sufficient for qualitative perception in applications like video. QoS control is a key feature for efficient utilization of resources in wireless networks supporting mobile multimedia.

- *Fine-grained and coarse-grained parallelism* – Typical multimedia functions like image, voice and signal processing require a fine-grained parallelism in that the same operations across sequences of data are performed. In many applications a pipeline of functions process a single stream of data to produce the end result.

- *High instruction reference locality* – The operations on the data demonstrate typically high temporal and spatial locality for instructions.

- *High memory and network bandwidth* – Many multimedia applications require huge memory bandwidth for large data sets that have limited locality. Streaming data – like video and images from external sources – requires high network and I/O bandwidth.

The challenge is to maintain a high perceived end-to-end quality without limiting applications to the point where they are no longer useful. Multimedia networking requires at least a certain minimum bandwidth allocation for satisfactory application performance. The minimum bandwidth requirement has a wide dynamic range depending on the users' quality expectations, application usage models, and applications' tolerance to degradation. In addition, some applications can gracefully adapt to sporadic network degradation while still providing acceptable performance. MPEG-4 will provide technique solution to some of these requirement, such as MPEG-4 provides efficient compression and coding for multimedia data, and provides objet-based coding method to support the interactive of multimedia information, and it is also benefit to the QoS. MPEG-4 is an important role in the Mobile Multimedia System.

3 MPEG-4 Technical description

The Moving Picture Coding Experts Group is a working group of ISO/IEC in charge of the development of international standards for compression, decompression, processing, and coded representation of moving pictures, audio and their combination. The first two standards produced by MPEG are: MPEG-1, a standard for storage and retrieval of moving pictures and audio on storage media. MPEG-2 is a standard for digital television. MPEG is has recently finalized MPEG-4 Version 2, a standard for multimedia applications. MPEG has also started work on a new standard known as MPEG-7: a content representation standard for information search, scheduled for completion in Fall 2001.

MPEG-4 is the standard for coding of audiovisual information in multimedia systems. It provides audiovisual functionality, which includes content manipulation, content scalability and content-based access, for multimedia systems. The MPEG-4 standard is necessary because the ways in which audiovisual material is produced, delivered and consumed are still evolving. Furthermore, hardware and software keep getting more powerful. So there is more and more multimedia information such as audio and video is possible to be transferred in the network. And more and more synthetic multimedia information is generated for business application, which require more flexible and reusable operation on multimedia information. MPEG-4 provides the object-based representation and some technique based on MPEG-1 and MPEG-2 to satisfy the new requirements.

3.1 Scope and Features of MPEG-4 Standard

The focus and scope of MPEG-4 is defined in the intersection of the traditionally separate industries of telecommunications, computer, and digital TV/movies. One of the examples is the mobile multimedia systems. In detail, the application of MPEG-4 standard is aimed at such as Internet and Internet video, wireless video, multimedia database, Interactive home shopping, multimedia e-mail, home movies, virtual reality games, simulation and training.

The users of the MPEG-4 can be divided into three categories: the author of the multimedia content, the network service providers and the end users of MPEG-4. For authors, MPEG-4 provides facilities for the content reusable and operation flexible. For network service providers, it offers transparent information, which can be interpreted and translated into the appropriate native signaling messages of each network with the help of relevant standards bodies. For end users, it brings higher levels of interaction with content, within the limits set by the author. MPEG-4 achieves these goals by providing standardized ways to:

1. Content-based interactivity: It provides the object-based methodology to represent units of aural, visual or audiovisual content as media objects. These objects can be natural or synthetic originals. Describe the composition of these objects to create compound media objects that form audiovisual scenes. And Interact with the audiovisual scene generated at the receiver's end.

2. Universal accessibility: The ability to access audiovisual data over a diverse range of storage and transmission media. The capability to access to applications over a variety of wireless and wired networks and storage media. And the ability to achieve scalability with fine granularity in spatial, temporal or amplitude resolution, quality or complexity. In order to provide this service, efficient compression is provided in MPEG-4.

3. Object composition: Multiplex and synchronize the data associated with media objects, so that they can be transported over network channels providing a QoS appropriate for the nature of the specific media objects.

3.2 Techniques on Systems and MPEG-4 DMIF

In this section, we give an overview of MPEG-4 working procedure, which describes how the coming streaming from the sender side is rendered in the receiver side. And then we will discuss the Delivery Multimedia Integration Framework (DMIF) from its composition, architecture and computational model. Based on these descriptions, the detailed technique description of object-based representation is given.

3.2.1 MPEG-4 System

MPEG-4 system not only refers to overall architecture, multiplexing, and synchronization, but also encompasses scene description, interactivity, content description, and programmability. Its mission is to *"Develop a coded, streamable representation for audiovisual objects and their associated time-variant data along with a description of how they are combined"*. In detail, all information that MPEG-4 contains is binary coded for bandwidth efficiency, and MPEG-4 is built on the concept of streams that have a temporal extension. MPEG-4 System does not deal with the encoding of audio or visual information (the coding of audiovisual is based on the previous MPEG-1 and MPEG-2 and with optimization). But it deals with the information related to the combinations of streams: combination of audiovisual objects to create an interactive audiovisual scene, synchronization of streams, multiplexing of streams for storage or transport.

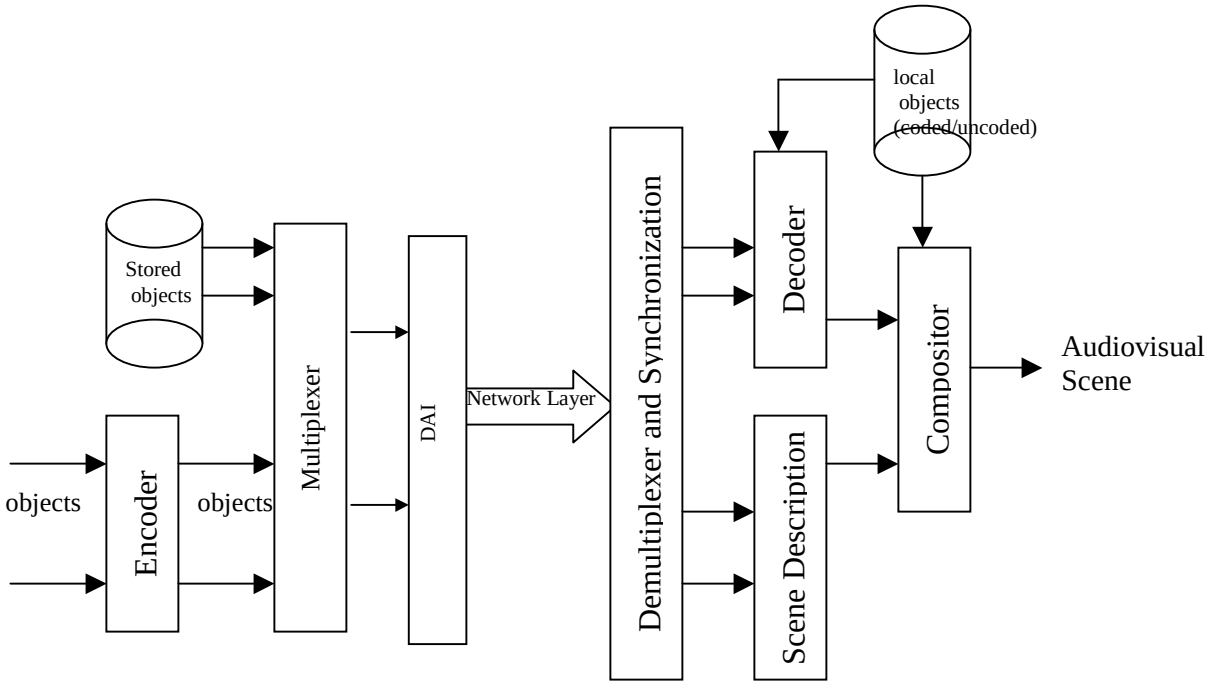


Figure 1 Architecture of an MPEG-4 System

The architecture of an MPEG-4 system is presented in Figure 1. At the sending side, different objects are encoded independently and the several elementary streams associated to the

various objects are multiplexed together. At the receiving side, the elementary streams are demultiplexed, the various media objects are decoded, and finally the scene is composed using the scene description information. Here we give a walkthrough of the system:

The transport of the MPEG-4 data can occur on a variety of delivery systems. This includes MPEG-2 Transport Streams, UDP over IP, ATM or the DAB (Digital Audio Broadcasting) multiplexer. Then the delivery layer provides to the MPEG-4 terminal a number of elementary streams. The DMIF Application Interface (DAI) interface defines the process of exchanging information between the terminal and the delivery layer in a conceptual way, using a number of primitives. The DAI defines procedures for initializing an MPEG-4 session and obtaining access to the various elementary streams that are contained in it. These streams can contain a number of different information: audio-visual object data, scene description information, control information in the form of object descriptors, as well as meta-information that describes the content or associates intellectual property rights to it. The synchronization layer provides a common mechanism for converting time and framing information. It is a flexible and configurable packetization facility that allows the inclusion of timing, fragmentation, and continuity information on associated data packets. Such information is attached to data units that comprise complete presentation units. From the SL information we can recover a time base as well elementary streams. The streams are sent to their respective decoders that process the data and produce composition units. In order for the receiver to know what type of information is contained in each stream, control information in the form of *object descriptors* is used. These descriptors associate sets of elementary streams to one audio or visual object, define a scene description stream, or even point to an object descriptor stream.

One of the streams must be the scene description information associated with the content. The scene description information defines the spatial and temporal position of the various objects, their dynamic behavior, as well as any interactivity features made available to the user. This scene description information is at the heart of the MPEG-4 vision and thus at the basis of most of

the new functionality that MPEG-4 can provide. The scene description information is the glue that structures a scene - *which players are in the stage, where and partly how they should look like* - and after controls their spatial and temporal evolution - *where the players move and when.* The architecture requires MPEG-4 not only to address the coding of the raw audiovisual data, and of facial animation data, but also the coding of the scene description information. The scene description coding format specified by MPEG-4 is known as BIFS (BInary Format for Scene description), and represents a pre-defined set of scene objects, e.g. video, audio, 3D faces, and corresponding behaviors along with their spatio-temporal relationships.

3.2.3 DMIF

DMIF is a session protocol for the management of multimedia streaming over generic delivery technologies, especially to address the delivery integration of three major technologies: the broadcast technology, the interactive network technology and the disk technology. It is similar to FTP in principle, but it returns pointers to where to get streamed data, but the FTP returns data.

The functionality provided by DMIF is expressed by an interface called DMIF-Application Interface (DAI), and applications access information from the underlie network or storage through the provided primitives, which will be translated into protocol messages. And these messages maybe different according to the different networks it operates. So the DAI provide a generic interface for applications to access multimedia content regardless the underlie network difference. And the DAI allow the DMIF users to specify the requirements for the desired stream. The DAI is also used for accessing broadcast material and local files. So the integration framework of DMIF covers three major technologies, interactive network technology, broadcast technology and the disk technology. An application accesses data through the DAI, irrespective of whether such data comes from a broadcast source, from local storage or from a remote server. In all scenarios the local application only interacts through a uniform interface DAI. Different DMIF instances will then translate the local application requests into specific messages to be

delivered to the remote application, taking care of the peculiarities of the involved delivery technology. Similarly, data entering the terminal (from remote servers, broadcast networks or local files) is uniformly delivered to the local application through the DAI.

How the DMIF provide a generic mechanism for the three technologies. From concept, a remote application accessed through a network is no different than an emulated remote producer application getting content from a broadcast source or from a disk. In the former case, however, the messages exchanged between the two entities have to be normatively defined to ensure interoperability. In the latter case, on the other hand, the interfaces between the two DMIF peers and the emulated remote application are internal to a single implementation and need not be considered in this specification. When considering the broadcast and local storage scenarios, it is assumed that the emulated remote application has knowledge on how the data is delivered/stored. When considering the remote interactive scenario instead, the DMIF layer is totally application-unaware. An additional interface—the DMIF – Network Interface (DNI) – is introduced to emphasize what kind of information DMIF peers need to exchange; an additional module – signal mapping – take care of mapping the DNI primitives into signaling messages used on the specific network.

Now, we describe the DMIF computational model. The high level walk-through of DMIF service consists of four steps; it is illustrated in the figure 2:

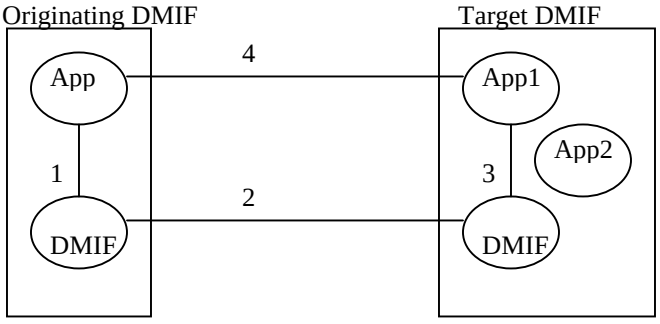


Figure 2 — DMIF Computational Model

1. The originating application request the activation of a service to its local DMIF layer -- a communication path between the originating application and its local DMIF peer is established in the control plane.

2. The originating DMIF peer establishes a network session with the target DMIF peer -- a communication path between the originating DMIF peer and the target DMIF peer is established in the control plane.

3. The target DMIF peer identifies the target application and forwards the service activation request -- a communication path between the target DMIF peer and the target application is established in the control plane.

4. The peer applications create channels (requests flowing through communication paths 1, 2 and 3). The resulting channels in the user plane (4) will carry the actual data exchanged by the Applications.

The DMIF Layer automatically determines whether a particular service is supposed to be provided by a remote server on a particular network e.g., IP based, or ATM based, by a broadcast network, or resides in a local storage device. The selection is based on the peer address information provided by the Application as part of a URL passed to the DAI.

3.2.3 Object-based Representation and BIFS

In order to improve the content reusability and the operation flexibility, in MPEG-4, the audio or video coding data model based a concepts which need to be deeply associated to the media content structure: the objects. Objects have typically a semantic associated and are user meaningful entities in the context of the relevant application. Since human beings do not want to interact with abstract entities, such as pixels, but rather with meaningful entities that are part of the scene, the concept of content is central to MPEG-4. MPEG-4 understands an audiovisual scene as a composition of audiovisual objects with specific characteristics and behavior, notably in space and time. MPEG-4 is the first truly digital audiovisual representation standard. The

object composition approach allows supporting new functionality, such as content-based interaction and manipulation, as well as to improve already available functionality, such as coding efficiency.

MPEG-4 defines a syntactic description language to describe the exact binary syntax for bitstreams carrying media objects and for bitstreams with scene description information. In addition to providing support for coding individual objects, MPEG-4 also provides facilities to compose a set of such objects into a scene. The necessary composition information forms the scene description, which is coded and transmitted together with the media objects. Starting from VRML (the Virtual reality Modeling Language), MPEG has developed a binary language for scene description called BIFS.

In order to group objects together, an MPEG-4 scene follows a hierarchical structure, which can be represented as a directed acyclic graph. Each node of the graph is a media object, the leaf nodes are primitive objects, and the middle nodes correspond to the compound nodes. The tree structure are dynamic, and the BIFS provides commands to support the operations such as add, replace and remove nodes, and also the node attributes could be updated. In order to associate objects to space and time, audiovisual objects have both a spatial and a temporal extent. Each media object has a local coordinate system. A local coordinate system for an object is one where the object has a fixed spatio-temporal location and scale. The local coordinate system serves as a handle for manipulating the media object in space and time. Media objects are positioned in a scene by specifying a coordinate transformation from the object's local coordinate system into a global coordinate system defined by one more parent scene description nodes in the tree. To providing the attribute value selection, individual media objects and scene description nodes expose a set of parameters to the composition layer through which part of their behavior can be controlled. Examples include the pitch of a sound, the color for a synthetic object, activation or deactivation of enhancement information for scaleable coding, etc.

3.3 Techniques on MPEG-4 Video

MPEG-4 video offers technology that covers a large range of existing applications as well as new ones. It is used on reliable communication over limited rate wireless channels. It can be used for surveillance data compression. And it can provide high quality video and audio for entertainment business application. MPEG-4 visual standard provides standardized core technologies allowing efficient storage, transmission and manipulation of textures, images and video data for multimedia environments. It allows the decoding and representation of atomic units of image and video content, called video objects (VOs).

3.3.1 Overview of *MPEG-4 Video Coding*

An input video sequence is a sequence of related snapshots or pictures, separated in time. In MPEG-4, each picture is considered as consisting of temporal instances of objects that undergo a variety of changes, and new objets enter a scene or existing objects depart leading to the presence of temporal instances of certain objects only in certain pictures. MPEG-4 supports only the access of whole of sequence of picture, but also temporal instance of the picture or a particular object.

Video Object Planes (VOPs) is the temporal instance of VOs. It is described by texture variations and shape representation. And VOPs are obtained by semi-automatic or automatic segmentation, and its result shape information is binary shape mask or gray scale shape. Such as a scene includes a book and a cup on the background, so the book and a cup are segmented as VOP1 and VOP2, and the background is segmented as VOP0. Thus, a segmented sequence contains a set of VOP0, a set of VOP1, and a set of VOP2. Each VO is encoded separately and multiplexed to form a bitstream, and it is sent out with the composition scene information.

From top to bottom, MPEG-4 coded data is a tree structure; the video session is on the highest level and consisted of an ordered collection of VO. And the VO represents a complete scene or a portion of scene with semantic meaning. And the VO is consisted of video object layer,

which represents various instantiations of a VO. And the video object layer is consisted of group of object planes. And the bottom is the VOP, which represents a snap shot in term of a video object.

An MPEG-4 video encoder includes three components: Motion Coder, Texture Coder and Shape Coder. The Motion Coder uses macroblock and block motion estimation and compensation similar to MPEG-1and MPEG-2 but with arbitrary shapes. The Texture Coder uses block DCT coding based on MPEG-1 and MPEG-2 but with arbitrary shapes. And the Shape Coder is used to coding the arbitrary shape video. The data of VO is sent to the System Multiplexer and transfer out.

3.3.2 Binary Shape Coding

For each VO given as a sequence of VOPs of arbitrary shapes, the corresponding sequence of binary alpha planes is assumed to be known. For the binary alpha plane, a rectangular bounding box enclosing the shape to be coded is formed such that its horizontal and vertical dimensions are multiples of 16 pixels (macroblock size). For efficient coding, it is important to minimize the number of macroblocks contained in the bounding box. The pixels on the boundaries or inside the object are assigned a value of 255 and are considered opaque while the pixels outside the object but inside the bounding box are considered transparent and are assigned a value of 0. If a 16 * 16 block structure is overlaid on the bounding box, three types of binary alpha blocks exist: completely transparent, completely opaque, and partially transparent (or partially opaque). Coding of each 16 * 16 binary alpha block representing shape can be performed either loss or without loss. The degree of loss in coding the shape of a video object is controlled by a threshold which can take values of 0, 16, 32, 64, . . . 256. Each binary alpha block can be coded in intra mode or in inter mode. In intra mode, no explicit prediction is performed. In inter mode, shape information is differenced with respect to the prediction obtained using a motion vector, the resulting binary shape prediction error may or may not be coded. The motion

vector of a binary alpha block is estimated at the encoder by first finding a suitable initial candidate from among the motion vectors of 3 previously decoded surrounding texture macroblocks as well as the 3 previously decoded surrounding shape binary alpha blocks. Next, the initial candidate is either accepted as the shape motion vector, or is used as the starting basis for a new motion vector search, depending on if the resulting prediction errors of the initial motion vector is below a threshold. The motion vector is coded differentially and included in the bitstream.

3.3.3 Motion Coding

The motion coding tool in MPEG-4 includes a Motion Estimator, Motion Compensator, Previous/Next VOPs store and Motion Vector (MV) Predictor and Coder. In case of P-VOPs, Motion Estimator computes motion vectors using the current VOP and temporally previous reconstructed VOP available from the previous reconstructed VOPs store. In case of B-VOPs, Motion Estimator computes motion vectors using the current VOP and temporally previous reconstructed VOP from the previous reconstructed VOP Store, as well as, the current VOP and temporally next VOP from the next reconstructed VOP store. The Motion Compensator uses these motion vectors to compute motion compensated prediction signal using the temporally previous reconstructed version of the same VOP (reference VOP). The MV Predictor and Coder generates prediction for the MV to be coded.

3.3.4 Video Texture Coding

The texture coding tool in MPEG-4 is to code the luminance and chrominance variations of blocks forming macroblocks with a VOP. The blocks that lie inside the VOP are coded using DCT coding, and the blocks that lie on the VOP boundary are first padded and then coded similar to the block that lie inside the VOP. The remaining blocks are transparent and are not coded.

The texture coding tool uses block DCT coding and codes blocks of size 8 * 8, and the blocks on the VOP boundary require padding prior to texture coding. The general operations in the texture encoding are: DCT on original or prediction error blocks of size 8 * 8, quantization of 8 * 8 block DCT coefficients, scanning of quantized coefficients and variable length coding of quantized coefficients.

3.3.5 Sprite Coding

A sprite refers to synthetic object that undergoes some form of transformation, and in MPEG-4, a static sprite is a large image built by integration of many frames of sequence spatially and many frames of a sequence temporally. One of the main components in coding using natural sprites is generation of the sprite itself. For generating a static sprite, the entire video object is assumed to be available. For each VOP in the VO, the global motion is estimated according to a transformation model using which a VOP is then registered with the sprite by warping the VOP to sprite coordinate system. Finally, the warped VOP is blended with the sprite which is used to for estimation of motion of subsequent VOP.

3.3.6 Scalable Video Coding

Scalability of video allows a simple video decoding tool to decode and produce basic quality video while an enhanced decoding tool may decode and produce enhanced quality video, all of them from same coded video bitstream. Scalable video encoding ensures that input video data is coded as two or more layers, an independently coded base layer and one or more enhancement layers coded dependently, thus producing scalable video bitstreams. The first enhancement layer is coded with respect to the base layer, the second enhancement layer with respect to the first enhancement layer and so forth.

MPEG-4 video offers a generalized scalability framework supporting both the Temporal and the Spatial scalabilities, the primary type of scalabilities. Temporally scalable encoding offers

decoders a means to increase temporal resolution of decoded video using decoded enhancement layer VOPs in conjunction with decoded base layer VOPs. Spatial scalability encoding on the other hand offers decoders a means to decode and display either the base layer or the enhancement layer output; typically, since base layer uses one-quarter resolution of the enhancement layer, the enhancement layer output provides the better quality, albeit requiring increased decoding complexity.

3.3.7 Facial Animation Coding

The facial animation object can be used to render an animated face. It is defined using a set of Facial Definition Parameters (FDPs) and Facial Animation Parameters (FAPs). These parameters define the face shape, texture, emotion and expressions of the face. The FAPs apply to different facial models result in reasonably similar expressions and speech pronunciation without the need to initialize or calibrate the model. And the FDPs allow the definition of a precise facial shape and texture in the setup phase.

Upon construction, the face object contains a generic face with a neutral expression. This face can already be rendered or it can also immediately receive the animation parameters from the bitstream to produce animation of the face: expressions, speech etc. Meanwhile, definition parameters can be sent to update the appearance of the face from something generic to a particular face with its own shape and texture. So a complete face model can be downloaded via the FDP set, and the animation can be got through the FAP set.

The FAP contains two parameters, one is viseme, and another is expression. A viseme is a visual correlate to a phoneme, and it allows viseme rendering and enhances the result of other parameters, ensuring the correct rendering of visemes. And the expression parameter allows the definition of high-level facial expression, which is defined by textural descriptions such as joy, sadness, anger, fear, disgust and surprise. The FDP set is specified using the FDP node, which defines the face model to be used at receiver. It supports two options: 1. Calibration information

is downloaded so that the proprietary of the receiver can be configured using facial feature and optionally 3-D mesh or texture. 2. A face model is downloaded with the animation definition of the FAPs. And this face model replaces the proprietary face model in the receiver.

3.3.8 Body animation Coding

The body animation coding is very similar to the facial animation coding. The body is an object capable of producing virtual body models and animations in form of a set of 3-D polygonal meshes ready for rendering. There are also two sets of parameters, which are defined for the body: Body Definition Parameter (BDP) set, and Body Animation Parameter (BAP) set. The BDP set defines the set of parameters to transform the default body to a customized body with its body surface, body dimensions, and (optionally) texture. The Body Animation Parameters (BAPs) produces reasonably similar high-level results in terms of body posture and animation on different body models.

3.3.9 2-D animated meshes

For general nature or synthetic visual objects, mesh based representation can be useful for enabling a number of functions such temporal rate conversion, content manipulation, animation, augmentation, transfiguration and others. MPEG-4 provides a tool for triangular mesh based representation of general purpose objects.

A 2D triangular mesh or a mesh object plane is a planar graph that tessellates or partitions a video object plane or its bounding box into triangular patches. The vertices of each patch are called node points. A 2D mesh object, which consists of a sequence of mesh object planes (MOPs), is compactly represented by mesh geometry at some key intra MOPs and mesh motion vectors at all other inter MOPs. The mesh geometry refers to the location of the node points on the key mesh object planes. 2D mesh animation is accomplished by propagating the 2D mesh defined on key MOPs using one motion vector per node point per object plane until the next key

MOP. Both mesh geometry and motion information is predictively coded for an efficient binary representation. The mesh based representation of an object and the traversal of nodes for mesh geometry coding is illustrated in Figure 3:

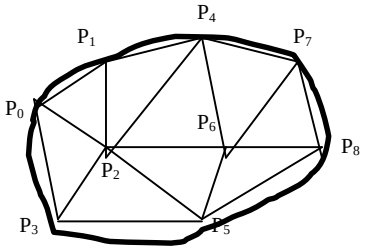


Figure 3 2D mesh representation of an object, and coding of mesh geometry

First, the total number of nodes and the number of boundary nodes is encoded. The tope-left node p_0 is coded without prediction. Then the next clockwise boundary node p_1 is found and the difference between p_0 and p_1 is encoded; then all other boundary nodes are encoded in similar way. Then, the no previously encoded interior node that is nearest to the last boundary node is found and the difference between these two nodes is encoded; this process is repeated until all the interior nodes are covered. The mesh initialized with respect to a particular VOP of the corresponding visual object; it consists of the initial positions of the mesh nodes.

To mesh-based motion, triangular patches in the current frame are mapped onto triangular patches in the reference frame. And the texture inside each patch in the reference frame is warped onto the current frame using a parametric mapping, such as affine mapping, as a function of the node point motion vectors. This process is called texture mapping, which is an integral part of mesh animation. The affine mapping between coordinates (x',y') at time t' and (x, y) at time t is given by :

$$x = a_1 x' + a_2 y' + a_3$$
$$y = a_4 x' + a_5 y' + a_6$$

where a_i are the affine motion parameters. The six degrees of freedom in the affine mapping matches that of warping a triangle by the motion vectors of its three vertices (with two degrees of freedom in each). Furthermore, if proper constraints are imposed in parameter (node

motion vector) estimation, an affine transform can guarantee the continuity of the mapping across the boundaries of adjacent triangles. Thus, 2D mesh modeling corresponds to non-uniform sampling of the motion field at a number of salient feature points (node points), from which a continuous, piece-wise affine motion field can be reconstructed.

2-D object-based mesh representation is able to model the shape and motion of a VOP in a unified framework, which is also extensible to the 3-D object modeling when data to construct such models is available. In particular, the 2-D mesh representation of video objects enables the following functionalities:

1. Video Object Manipulation

• Augmented reality: Merging virtual images with real video to create enhanced display information. The computer-generated images must remain in perfect registration with the moving real images.

• Synthetic-object-transfiguration/animation: Replacing a natural video object in a video clip by another video object. The replacement video object may be extracted from another natural video clip or may be transfigured from a still image object using the motion information of the object to be replaced.

• Spatio-temporal interpolation: Mesh motion modeling provides more robust motion-compensated temporal interpolation (frame rate up-conversion).

2. Video Object Compression

2-D mesh modeling may be used for compression if one chooses to transmit texture maps only at selected key frames and animate these texture maps (without sending any prediction error image) for the intermediate frames. This is also known as self-transfiguration of selected key frames using 2-D mesh information.

3. Content-Based Video Indexing

Mesh representation enables animated key snapshots for a moving visual synopsis of objects. It provides accurate object trajectory information that can be used to retrieve visual

objects with specific motion. And its representation provides vertex-based object shape representation which is more efficient than the bitmap representation for shape-based object retrieval.

3.4 Techniques on MPEG-4 Audio

MPEG-4 audio coding effort occurred in parallel with the MPEG-2 AAC coding effort. It supports the concept of audio object. MPEG-4 audio objects provides both representing natural sounds (such as speech and music) and for synthesizing sounds based on structured descriptions. The representation for synthesized sound can be derived from text data or instrument descriptions and by coding parameters to provide effects. The representations provide compression, scalability, effects processing and other functionality.

The MPEG-4 Audio coding covering 6kbit/s to 24kbit/s have the same quality as the AM digital audio broadcasting application in collaboration with the NADIB. It was found that higher quality could be achieved in the same bandwidth with digital techniques and that scalable coding configurations offered performance superior to a simulcast alternative. Since MPEG-4 Audio is compatible to MPEG-2 AAC coding, it supports all tools defined in MPEG-2.

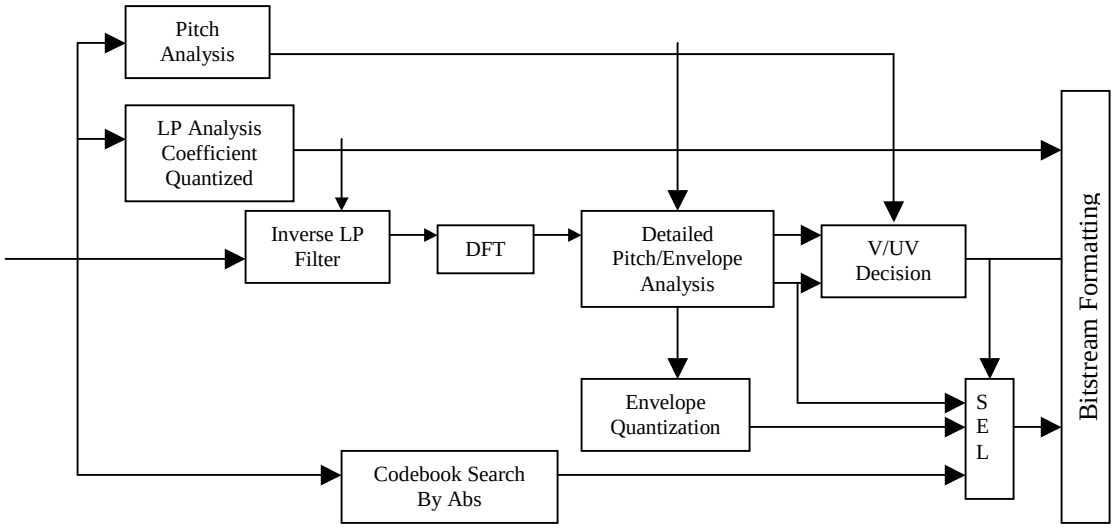
3.4.2 Nature Sound

MPEG-4 standardizes natural audio coding at bitrates ranging from 2 kbit/s up to and above 64 kbit/s. When variable rate coding is allowed, coding at less than 2 kbit/s, such as an average bitrate of 1.2 kbit/s, is also supported. Speech coding at bitrates between 2 and 24 kbit/s is supported by using Harmonic Vector eXcitation Coding (HVXC) for a recommended operating bitrate of 2 - 4 kbit/s, and Code Excited Linear Predictive (CELP) coding for an operating bitrate of 4 - 24 kbit/s. For general audio coding at bitrates at and above 6 kbit/s, transform coding techniques, namely TwinVQ and AAC, are applied. The audio signals in this region typically have sampling frequencies starting at 8 kHz.

• Parametric Audio Coding: Parametric Audio Coding uses the Harmonic and Individual Lines plus Noise (HILN) technique to code general audio signals at bit rates of 4 kbit/s and above using a parametric representation of the audio signal. In this coding method, the input signal is divided into audio objects, which are described by appropriate source models and represented using model parameters.

• HVXC

A basic block diagram of HVXC is depicted in Figure 5. HVXC first performs LP (Linear Predication) analysis to find the LP coefficients. Quantized LP coefficients are supplied to the inverse LP filter to find the prediction error. The prediction error is transformed into a frequency domain and the pitch and the envelope of the spectrum are analyzed. The envelope is quantized by weighted vector quantization in voiced sections. In unvoiced sections, closed-loop search of an excitation vector is carried out.



Abs: Analysis-by-Synthesis U/UV: Voiced/Unvoiced

Figure 4 HVXC

• CELP

The basis idea of CELP is illustrated in Figure 6. Firstly, the LP input signal coefficients are analyzed. Then quantized to be used in an LP synthesis filter driven by the output of the excitation codebooks. There are two steps to perform the encoding. First step, the Long-Term prediction coefficients are calculated. In the second step, a perceptually weighted error between

the input signal and the output of the LP synthesis filter is minimized. This minimization is achieved by searching for an appropriate code vector for the excitation codebooks. Quantized coefficients, as well as indexes to the code vectors of the excitation codebooks and the long-term prediction coefficients, form the bitstream. The LP coefficients are quantized by vector quantization and the excitation can be either MPE (multipulse excitation) or regular pulse excitation RPE (regular pulse excitation).

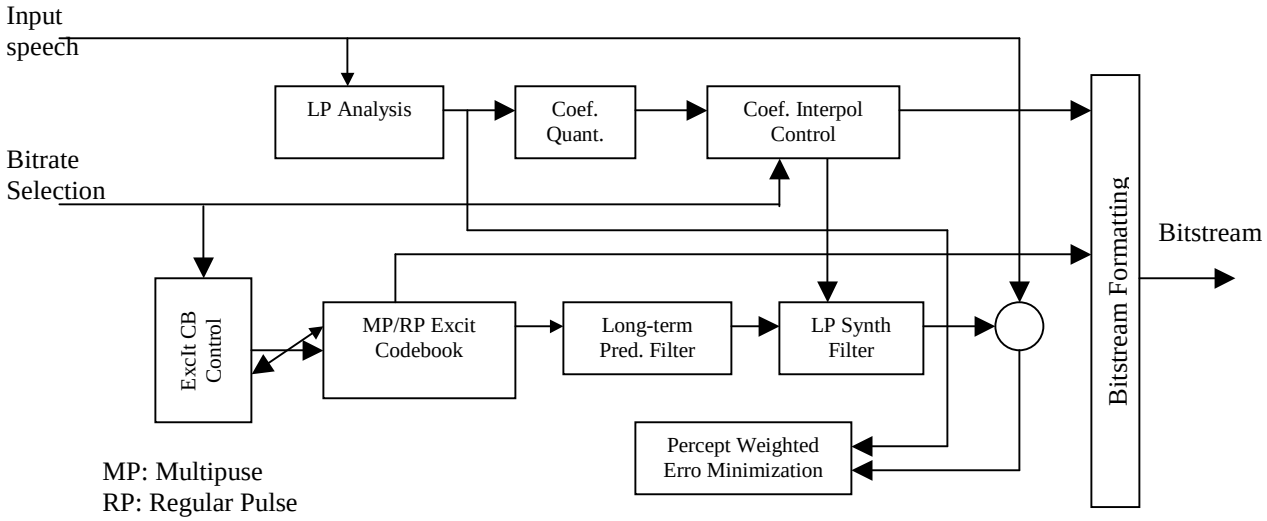


Figure 5 CELP

MPE allows more freedom on the interpulse distance than RPE which has a fixed interpulse distance. MPE achieves better coding quality than RPE. On the other hand, RPE requires less computation than MPE by trading off its coding quality.

- CELP Silence Compression: In the encoder, a voice activity detector is used to distinguish between regions with normal speech activity and those with silence or background noise. During normal speech activity, the CELP coding is used. Otherwise a Silence Insertion Descriptor (SID) is transmitted at a lower bit rate. This SID enables a Comfort Noise Generator (CNG) in the decoder. The amplitude and spectral shape of this comfort noise is specified by energy and LPC parameters similar as in a normal CELP frame.

3.4.2 Synthetic Sound

Synthetic sound of MPEG-4 can be generated from decoder based on structured description inputs. Here we introduce the basic ideas of Text to Speech generation and the Score Driven Synthesis.

- Text To Speech: It allows a text or a text with prosodic parameters (pitch contour, phoneme duration, and so on) as its inputs to generate intelligible synthetic speech. It supports the generation of parameters that can be used to allow synchronization to associated face animation, international languages for text and international symbols for phonemes. Additional markups are used to convey control information within texts, which is forwarded to other components in synchronization with the synthesized text. The application of TTS includes such as artificial story teller, synthesized speech output synchronized with Facial Animation, voice Internet etc.

- Score Driven Synthesis: The Structured Audio tools decode input data and produce output sounds. This decoding is driven by a special synthesis language called SAOL (Structured Audio Orchestra Language), which is standardized as a part of MPEG-4. This language is used to define an "orchestra" made up of "instruments" (downloaded in the bitstream, not fixed in the terminal) which create and process control data. An instrument is a small network of signal processing primitives that might emulate some specific sounds such as those of a natural acoustic instrument. The signal-processing network may be implemented in hardware or software and include both generation and processing of sounds and manipulation of pre-stored sounds.

Control of the synthesis is accomplished by downloading scores or scripts in the bitstream. A score is a time-sequenced set of commands that invokes various instruments at specific times to contribute their output to an overall music performance or generation of sound effects. The score description, downloaded in a language called SASL (Structured Audio Score Language), can be used to create new sounds, and also include additional control information for modifying existing sound. This allows the composer finer control over the final synthesized sound. For synthesis processes that do not require such fine control, the established MIDI protocol may also be used to control the orchestra.

4 An MPEG-4 Based Mobile Conferencing System

Based on the discussion of MPEG-4 technology, here we introduce a mobile conference system. The purpose of the system is to provide a virtual conference room for participants, who are not seat in the real conference room. And the virtual conference can show the participants and multimedia information. Participants can present information, discuss each other like in the real conferences room and even more.

The room is described as a BIFS scene. The scene description contains references to Audiovisual Objects (AVOs) representing the voice and the picture of the participants. The scene description itself and the AVOs are transmitted by elementary stream. The audio/video elementary streams are transmitted using UDP/IP since it is not sensitive. However, the sensitive scene description information is transmitted using TCP/IP.

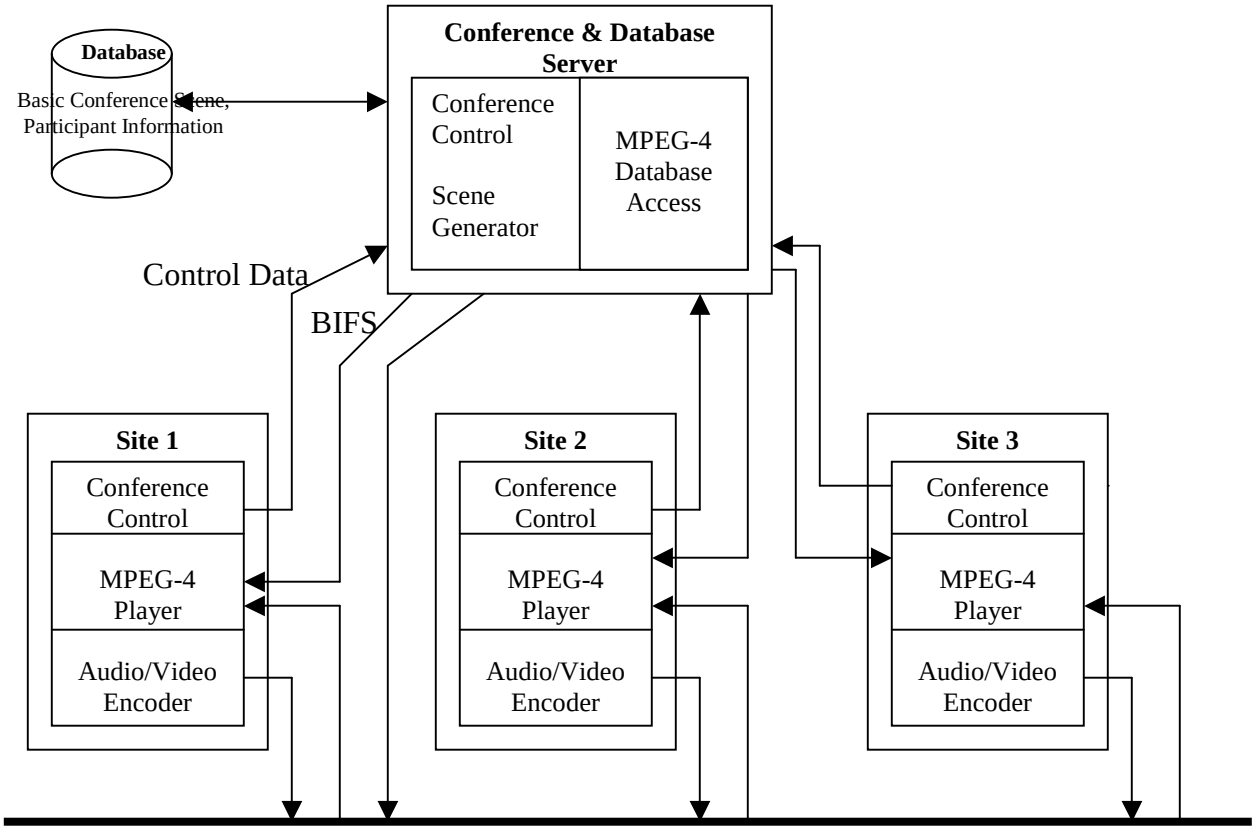


Figure 6 Client-Server Architecture for Mobile Conference System

The system is a client/server structure, and each participant stay with the client such as a PC including video and audio input and output devices. The client encodes the audio and video information of the participant, and then multicast the stream to all other clients, or the stream are sent to server which redirects the streams to other clients. And the server manages the virtual meeting room, the participants and their actions. It composes the scene description of the conference room tailored for a dedicated client and sends it to this client. A set of conference room description templates is stored in a multimedia database, such that a participant can easily select one. The database also provides MPEG-4 pre-encoded material used in the conference.

5 Summary

Mobile Multimedia System is a new computing paradigm of computing with the advances in wireless networking technology and development of semiconductor technology. There are many technological challenges to establishing the computing paradigm. Supporting the mobile multimedia computing is one of the important motivations of the MPEG-4 development. In this survey, firstly, we briefly describe the mobile multimedia system concept, current state, architecture, and its challenge techniques. After discussion of the application, scope and feature of MPEG-4, our works are focused on the technical description of MPEG-4. Its technique is described from three folds. 1. MPEG-4 DMIF and system, which describes the multimedia content delivery integration framework, and the object-based representation and BIFS was described in this section. 2. Introducing the MPEG-4 visual core technologies allowing efficient storage, transmission and manipulation of textures, images and video data for multimedia environments. 3. Describing coding of audio objects and synthesizing sounds based on structured descriptions. Finally, based on the discussion of the mobile multimedia system and MPEG-4, an application example on MPEG-4--An MPEG-4 Based Mobile Conferencing System, is given.

Reference:

1. A. Pearmain, V. Typpi, et al., "The MoMuSys MPEG-4 Mobile Multimedia Terminal and Field Trials", ACTS Mobile Summit 1999, pages 741-746, June 1999.
2. Agrawal, P., Hyden, E., Krzyzanowski, P., Mishra, P., Srivastava, M. B., and Trotter, J. A. “SWAN: A Mobile Multimedia Wireless Network”. IEEE Personal Communications 3, 2 (April 1997), 18--33.
3. F. Pereira et al., "MPEG-4 video subjective tests procedures and results", Special Issue on MPEG-4 of IEEE Trans. on CSVT, vol.7, February 1997.
4. Meggers, J., Bautz, G., and Park, A. S.B. “Providing Video Conferencing for the Mobile User”. In Proceedings of the 21st Conference on Local Computer Networks (Minneapolis, Minnesota, October 1996), pp. 526--534.
5. N. Brady, F. Bosse, N. Murphy, "Context-based arithmetic encoding for compressing 2D shape sequences", IEEE International Conference on Image Processing, Vol. 1, pg. 29-32, Santa Barbara, US, October 1997.
6. R. Koenen, F. Pereira, L. Chiariglione, "MPEG-4: Context and Objectives", Image Communication Journal: MPEG-4 Special Issue, vol. 9, May 1997.
7. ISO/IEC JTC1/SC29/WG11 Doc.N3747, “MPEG-4 Overview”, October 2000
8. ISO/IEC JTC1/SC29/WG11 Doc. N2725, “MPEG-4 Overview”, March 1999.
9. ACTS MOMUSYS Project Information. http://www.tnt.unihannover.de/project/eu/momusys
10. MPEG-4 Video Web Site: http://wwwam.hhi.de/mpeg-video
11. Tutorial Issue on the MPEG-4 Standard, http://www.cselt.it/leonardo/icjfiles/mpeg-4_si/

Recent MPEG Standards for Future Media Ecosystems. White papers. Ad hoc groups.Â Complex, interactive multimedia computing is a very broad subject: support for numerous coding formats for natural audio and video, fundamental differences in user interface (2D internet portals, 3D gaming), in delivery networks (cable/satellite broadcasts, broadband internet or mobile networks) and in terminals (PC, set-top boxes, PDAs/Mobile Phones) have increased the multimedia market segmentation.Â MPEG-4 data is carried by elementary streams, or logical transportation channels, and a stream can only carry a given type of data (scene data, visual data, etcâ€¦). MPEG-4 Visual, Systems, and Advanced Video Coding licensing is managed by MPEG LA LLC (http://www.mpegla.com/). These licenses cover the manufacture and sale of devices or software and, for some content disseminators, levy fees according to number of endusers or the extent of content delivered.Â The standard defines a set of tools that provide binary coded representation of individual audiovisual objects, text, graphics, and synthetic objects.Â The delivery of MPEG-4 content is supported by the Delivery Multimedia Framework or DMIF and its application interface. MPEG-J is described in Part 1 of the standard (ISO/IEC 14496-1:2004). This API for the interoperation of MPEG-4 media players with Java code is contrasted with a conventional parametric system. media processor Processor with features specic to multimedia coding and processing. motion. Prediction of a video frame with modelling of motion.Â A multimedia coding standard. NAL. Network Abstraction Layer.Â MPEG-4 (a multi-part standard covering audio coding, systems issues and related aspects of audio/visual communication) was rst conceived in 1993 and Part 2 was standardised in 1999. The H.264 standardisation effort was initiated by the Video Coding Experts Group (VCEG), a working group of the International Telecommunication Union (ITU-T) that operates in a similar way to MPEG and has been responsible for a series of visual telecommunication stan-dards. Port details. mpeg4ip Standards-based system to encode, stream, and play MPEG-4 audio/video. 1.6.1_46 multimedia =18 1.6.1_46Version of this port present on the latest quarterly branch. DEPRECATED: old, unmaintained version; superseded by multimedia/ffmpeg This port expired on: 2019-06-15.Â libhttp.so:multimedia/mpeg4ip. No installation instructions: this port has been deleted. The package name of this deleted port wasÂ The port also depends on other outdated, unmaintained code, most notably ffmpeg0 which is also scheduled for removal on 2019-06-15. The functionality provided by this port is available through multimedia/ffmpeg, which is actively maintained. PR: 238093 Reported by: tobik.