MPEG-4 标准及相关进展

刘占平 董士海
(北京大学计算机科学技术系图形研究室,北京 100871)

摘 要 MPEG-4 标准面向众多而非特定的应用。它灵活、可扩展,满足未来音频视频应用的广泛需求,实现基于内容的交互性、可重用性和可伸缩性。本文介绍 MPEG-4 系统框架、接收端模型、系统层模型、视频对象平面、基于模型编码和基于语义编码。

关键词 MPEG-4 系统框架 接收端模型 视频对象平面 基于模型编码 基于语义编码

0 引 言

关于图象视频压缩已有:$n \times 384$kbps 电视会议 H.120 建议、$p \times 64$kbps($p \leq 30$)视频编码标准 H.261、连续色调静止图象压缩标准 JPEG、PSTN 等低比特率活动图象压缩标准 H.263。1988 年,ISO 与 CCITT 成立活动图象专家组 MPEG(Moving Picture Experts Group),研究数字存储媒体上的活动图象及其伴音的编码表示,1992 年通过 1.5Mbps 编码标准 MPEG-1,1994 年通过 2Mbps-30Mbps 高质量视频音频编码标准 MPEG-2。以上标准均偏重于某(几)个应用领域,交互性较差,至多允许视频序列可变速度的双向播放,可重用性只限于矩形视频区域及其相关音频的线性片段,无法在同一场景集成自然信息与合成信息,不能提供对各种网络的一致性访问,容错性、灵活性和可扩展性比较差。

1 MPEG-4 标准概述

MPEG 于 1991 年 5 月提出关于视频音频编码的 MPEG-4 项目,设系统、音频、视频、需求、实现研究、测试及自然合成混合编码(SNHC)子组,1998 年 11 月将成为国际标准。

1.1 MPEG-4 标准的特点[1]

MPEG-4 提供更强的交互能力。场景中的每个对象独立编码,用户可以选择性地与其中某(几)个对象交互,具有良好的重用性。重新组合音视对象 AVO(Audio Visual Object)构造新场景,可以集成各种对象,无缝地集成自然信息与合成信息,实时信息与存储信息,AVO 可以是单/双/多声道音频信息、单/双/多镜头 2D/3D 视频信息。可以透明地访问信息,通过各种网络传输的信息最终映射为本地信息,整个过程给用户的感觉就如同访问本地信息。而且允许基于内容(比特率、分辨率、帧率、防错保护和解码优先级)的可伸缩性和服务质量(QOS)参数,更加灵活,可扩展,充分考虑未来技术的发展及应用需求,将解码器可编程能力分为:不可编程的标准工具集合(Flex_0);由标准化工具及其接口灵活配置的算法(Flex_1);多种工具构成可能算法的标准化可扩展机制(Flex_2)。

1.2 MPEG-4 技术动态

ACTS 是欧洲的一个研究与技术开发组织,其 MCM 子组的工作与 MPEG-4 系统、算法和工具、脸部特征跟踪及动画等密切相关。目前 SCALAR[2] 项目研究和开发一族 PSTN 等低比特率可视电话视频编码算法,允许异质网间视频会议的比特流可伸缩性。VIDAS[3] 项目为时间相关表现、编码和 AV 流操纵设计一个合适算法,在可视电话场景(编码器)分析、(解码器)合成时充分利用语音和脸动的相关性,在真实帧之间插入合成帧提高帧率,平滑显示与语音同步的唇部运动,DVP[4] 面向分布式创作和分布式虚拟现实。蓝室(Blue Room)、计算机、合成三者可

收稿日期:1998-07-15;收到修改稿日期:1998-10-21

分布在不同地点，蓝室视频信号、控制信号和跟踪信号由远地传来，而背景和动画则在本地实时绘制，它们与蓝室视频信号合成后生成最终信号。分布式虚拟现实系统中一次性传送基本 3D 模型，然后只传输动态交互引起的模型变化量，人们通过高速网络可以在虚拟世界里交互。最近东芝公司推出首款基于 MPEG-4 视频流系统 MobileMotion 套件，包括 MPEG-4 产品、服务器和播放器软件，用于 Internet/Intranet 上构建视频和多媒体应用，支持 CIF 和亚四分之一 CIF，6kbps—384kbps 时每秒可处理 30 帧视频数据。

2　MPEG-4 系统部分[5]

2.1　MPEG-4 系统框架

场景采用层次化树型结构(如图 1 所示)。叶子节点是原子 AVO，多个原子 AVO 经过组合构成复合 AVO，多个 AVO 按照时空关系组合生成场景。AVO 在发送端经过组织、压缩后被复用(Multiplexing)，即把压缩后的 AVO 基本流 ES(Elementary Stream)、相关控制信息和系统控制信息打包到一个或多个比特流中，它们作为下行流通过信道传输给接收端。

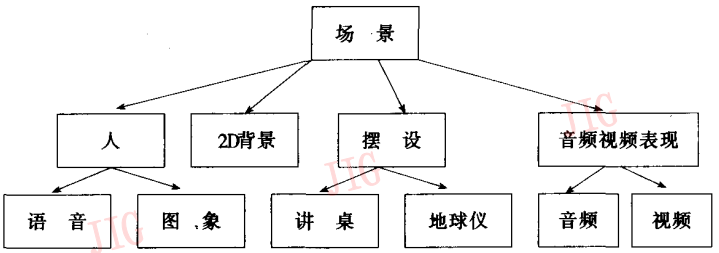

图 1　MPEG-4 场景的层次化结构

接收端[1](如图 2 所示)按照场景描述信息实现场景。AVO 数据被去复用(Demultiplexing)，交给相应解码器处理，解码器从 AVO 编码形式中恢复其数据，进行必要操作以重建原始 AVO，然后进行组合(Composition)，即把 AVO 从其自身局部坐标系映射到场景全局坐标系，按照树型结构组织起来，最后在接收端绘制。终端用户可以在组合、去复用、复用、编码四个阶段与 AVO 交互，有的交互信息需要作为上行流返馈给发送端。

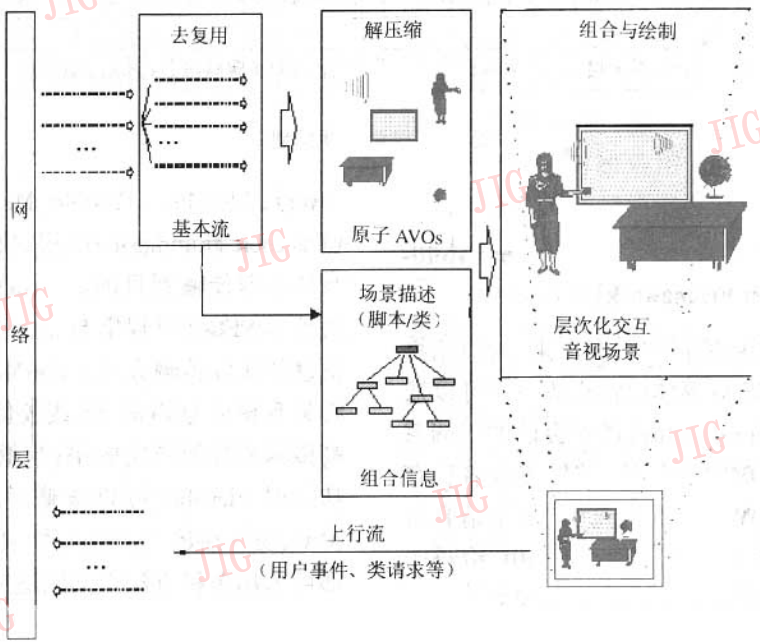

图 2　MPEG-4 接收端

2.2 MPEG-4 系统接收端模型 SRM(System Receiver Model)及音视组合

发送端通过 SRM 预测会话(Session)过程中接收端的行为。SRM 包括通信模型和系统解码器模型,解码器模型包括缓冲区模型和时间模型。

(1) 通信模型

发送端将 AVO 信息流(包括控制流)编码传输,接收端负责解码信息流并表现出来。不仅有发送端到接收端的下行流,还有接收端到发送端的上行流,接收端借此与发送端会话、能力协商、交换控制信息,实现点到点、多点到点、点到多点通信。

(2) 系统解码器模型

组成包括去复用器(Demultiplexer)、访问单元层 AL(Access unit Layer)、基本流缓冲区 EB(Elementary stream Buffer)、音视对象解码器(AVO Decoder)、组合缓冲区 CB(Composition Buffer)和组合器(Compositor)。编码数据首先被分割成访问单元 AU(Access Unit)存于 AL,AU 有显式/隐式解码时间和表现时间,解码时每个 AU 即时地从 EB 中移出、解码,作为表现单元放入 CB 中,它可以被组合器多次访问,一直到碰见过期时间戳 ETS 或者后继表现单元的表现时间已到,才从 CB 中被移出。如图 3 所示。

缓冲区模型 发送端通过缓冲区模型提前通知接收端在整个会话期间所需之最小缓冲区资源,接收端据此判定能否进行该次会话,会话过程中发送端一直监控缓冲区资源,调度数据传输次序,在接收端有充足空间的条件下预先发送非实时数据,以便给那些实时数据腾让信道资源。

时间模型 保证从信息流进入编码器直到从解码器输出的端-端延迟为恒定值,系统时间基 STB 调度解码器的一切行为,对象时间基 OTB 定义 AVO 编码器的时间机制;对象时钟参考 OCR 向解码器通报 OTB 的速度;时间戳 TS 保证音频视频信息的同步表现。

(3) 音视组合

组合 AVO 需要时空关系及可能的行为描述,AVO 局部坐标系是对它进行时空操纵的把柄。发送端把场景二值格式(BFFS)描述信息送到一个独立的 ES,接收端据此分析场景,按本地表现方法建立场景结构,把 AVO 定位在全局坐标系中,恢复对象间可能的关联(如脸可能需活动参数),保证比特流可编辑性,用户不需解码就可任意组合 AVO。

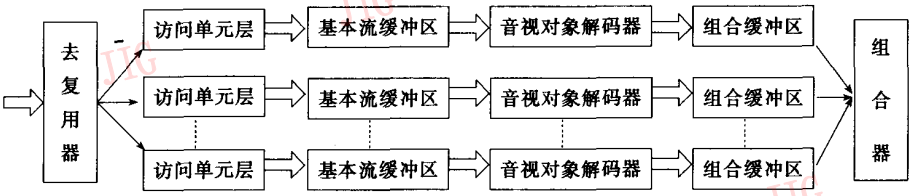

图 3 系统解码器模型

2.3 传输媒体集成框架 DMIF[5] (Delivery Multimedia Integration Framework)

DMIF 把应用从底层通信中分离出来,为应用提供一个通用界面,DMIF 对等体(Peer)是一个可以通过网络与目标实体(Target Entity)建立会话的终端系统,目标实体可以是 DMIF 对等体、传输流或者已存储的文件,应用通过 DMIF 界面为每个 ES 申请具有特定 QOS 和带宽的通道以建立会话,DMIF 给每个会话标识唯一的地址并负责会话的适时建立。

2.4 系统层模型(System Hierarchy Model)及 AVO 的复用与同步

系统层模型自上而下包括 AL(Access unit Layer)、FlexMux(Flexible Multiplexing)和 TransMux(Transport Multiplexing)3 层,决定不同 QOS 流如何同步地从源传输到目的。TransMux 位于 MPEG-4 标准之外,MPEG-4 只规定与 TransMux 的接口,并不规定信息的实际传输方式。FlexMux 提供一些复用工具,将具有相近 QOS 的 ES 或较低比特率的 ES 组合,提高传输资源的使用效率;当低层 TransMux 提供同样功能时 FlexMux 可以省略。AL 在 ES 中标识 AU、AVO、场景描述时间基的恢复和对象间同步,AU 头部可采用多种方法定义以适应不同的系统。

AVO 以一个/多个 ES 传输,ES 包含 QOS 和解码器资源等参数信息,由系统层模型标识 AU,传输时间戳和时钟参考信息,选择性地将不同 ES 数据插入到 FlexMux 流中,并进而映射到 TransMux;同时需传

输控制信息以便指示ES和FlexMux流的QOS参数。对象描述符OD和流映射表SMT把AVO和ES联系起来，OD描述每个AVO的ES数目和特征，SMT负责把每个流与通道联结标签(CAT，传输流通道的把柄)联系起来，再由DMIF把CAT映射到实际的传输通道。

2.5 语法描述语言SDL和系统描述语言MSDL

SDL(Syntax Description Language)把多媒体信息的比特流语法和解码工具、表现工具相分离，语法说明可使用正式或非正式技术，一旦给定语法说明，可使用有关算法的不同实现来解码；反之，也可以使用同一组工具修改比特流语法。MSDL(MPEG-4 System Description Language)是一个软件编程环境、可扩充类库，它描述面向对象的数据结构、媒体对象类及其编解码工具、组合工具等。MSDL与SDL的关系类似于MS Visual C++ MFC与C++的关系，SDL负责描述类库的语法及语义。MPEG-4系统、音频和视频都是可分档次(Profile)，现已提出三个档次：实时通信、基于内容的存储与检索、多媒体广播应用。

3 MPEG-4视频部分[6]

MPEG-4视频对象VO(Video Object)具有颜色、形状、运动和纹理等属性，视频部分提供一个工具箱，支持VO随机存取、容错性、高压缩比以及时空/质量/纹理/视野可伸缩性。基本编码结构有任意形状编码、标准8×8/自适应DCT纹理编码、运动预测与补偿。

3.1 网格(Meshes)

MPEG-4运用静态/动态、2D/3D网格来实现纹理编码。2D动态网格包含一个时间段内的网格几何信息、所有节点运动信息，在基于2D网格的纹理映射中，当前帧的三角碎片随着节点的运动变形为相邻帧的三角碎片，因此适于表示连续运动的变化区域。一般(generic)3D网格支持颜色、阴影，而且支持自然纹理、图象和视频到网格的映射。

3.2 视频对象平面

视频序列的每一帧可分割为一些任意形状的图象区域，即**视频对象平面**VOP(Video Object Plane)，每个VOP的颜色、形状、运动及纹理信息独立地编码、存储和传输，VOP标识以及多个VOP在接收端如何重新组合为原始序列等相关信息也要传给解码器。除了基本的基于块的运动补偿预测模式外，增加两种新模式：高级模式下增加了8×8子块的运动估值和重叠(Overlapped)运动补偿；非限制模式先在16×16宏块内作全搜索得到一个预测因子，以该预测因子为中心在一定搜索范围内作16×16和8×8的全搜索。在形状编码中，对VOP边界采用多边形匹配法代替基于块的运动预测，使用基于宏块的反复填充技术减少活动对象边缘的运动预测误差。采用零树(Zero-Tree)小波处理静止图象，提高纹理信息编码效率和更精细的可伸缩性；比特流中按照大致固定的距离插入再同步标志。

3.3 基于模型编码和基于语义编码

基于模型编码MBC(Model-Based Coding)首先在发送端和接收端按照事先约定分别建立两个相同的三维模型，发送端分析、提取特征(如人脸模型的形状参数、运动参数、表情参数等)并编码传输，接收端则利用接收到的特征参数根据建立的模型进行图象综合。

MBC利用先验知识确定对象一般线框模型，然后进行调和(adaptation)。全局调和采用仿射变换确定模型大小、位置和方向，改善一般模型与对象的匹配；局部调和建立模型各点和轮廓的对应关系，采用动态网格将模型各点以“弹簧”相连，调和时“力”通过“弹簧”向周围点传递，变形能均匀分布到各碎片表面，以此动态确定节点最佳位置。荷兰Delft技术大学的视频电话系统[7]采用变形模板提取脸形特征，描述嘴、眼睛、鼻子和下巴轮廓，将特征的形状几何参数化(例如采用三条抛物线定义上下唇边界)，根据先验知识给模板附加约束以保证参数取值合理，同时注意提取脸的边缘、凸点、凹点和纹理等。

基于语义编码(Semantic-based Coding)是针对脸部动作和表情的MBC，定义一些基本动作单元并对其变化(如嘴由张变合)编码。MPEG-4脸动及表情处理方案是先定义一个中性脸，然后从比特流接收脸定义参数FDPs(Face Definition Parameters)和脸动参数FAPs(Face Animation Parameters)，根据FDPs将一般脸调和为具有特定形状及纹理的具体脸，根据FAPs生成语音、表情并茂的脸部动作。德国Erlangen-Nuremberg大学定义一套MPEG-4脸部表情FAPs[8]，采用连续帧FAPs线性预测法生成层次化光流，用3D激光扫描生成的3D三角形B-样条模型描述脸形和纹理，使用FAPs建立表情模型并控制其局

部变形,FAPs参数值对脸动进行约束,例如手势、头转动或倾斜不许遮挡眼、嘴。香港城市大学根据MPEG-4 SNHC构造脸动模型,描述脸表情、头部全方位运动以及与语音同步的唇读视觉效果[9],提取头部前后上下左右6组视图,FAPs表情参数描述某一表情如何动作、应该移动哪些节点,例如"高兴"应该是嘴张开、眉毛舒展。

4 总 结

MPEG-4标准刻画了现在乃至未来数字音频视频通讯的灵活框架,实现各种传输媒体的通用访问和表现、基于内容的交互性、可重用性、可伸缩性,尤其提出一些新的编码压缩思想,这是一个新的挑战,也必将给我们带来新的机遇。

参 考 文 献

1 Leonardo Chiariglione-Convenor. The MPEG-4 Standard. CSELT-Italy, 1998.

2 Nyman H. SCALAR——Scaleable Architectures with Hardware Extensions for Low Bitrate Variable Bandwith Real-time Videocommunication (AC077). AccenTS Nr.2 (Special on MPEG-4), May 1996, Vol.1.

3 Fabio Lavagetto. VIDAS——Video Assisted with Speech Coding and Representation(AC057). AccenTS Nr.2, (Special on MPEG-4), May 1996, Vol.1.

4 Markus Wasserschaff. DVP——Distributed Video Production(AC089). AccenTS Nr.2, May 1996, Vol.1.

5 Rob Koenen. MPEG-4 Overview-(Tokyo Version). ISO/IEC JTC1/SC29/WG11 N2196, MPEG98.

6 MPEG Vedio Group. MPEG-4 Video Verification Model Version 7.0. ISO/IEC JTC1/SC29/WG11/N1642, Apr 1997.

7 Reinders M J T *et al*. Facial feature localization and adaptation of a generic face model for model-based coding. Signal Processing: Image Communication, 1995, 7:57~74.

8 Eisert P, Girod B. Facial expression analysis for model-based coding of video sequences. In: Proc Intern Picture Coding Symposium, Berlin, Germany, September 1997.

9 Cheung C H *et al*. Text-driven automatic frame generation using MPEG-4 SNHC for 2D head-and-shoulder scene. In: Proceeding of IEEE Conference on Image Processing, Oct 1997, Ⅱ:69~72.

刘占平 北京大学计算机系博士生。研究兴趣为多媒体,超文本,图象压缩和基于内容的检索。

董士海 北京大学计算机系教授,博士生导师。主要研究方向为图形学,人机交互和多媒体技术。

MPEG-4 and Its Related Development

Liu Zhanping, Dong Shihai

(*Computer Science Department, Peking University, Beijing* 100871)

Abstract MPEG-4 addresses a wealth of applications instead of a specific one, providing flexibility and extensibility for the various requirements of upcoming audiovisual applications, realizing content-based interactivity, reusability and scalability. The paper gives an introduction to MPEG-4 system framework, system receiver model(SRM), system hierarchy model, video object plane(VOP), model-based coding and semantic-based coding.

Keywords MPEG-4, System framework, SRM, VOP, Model-based coding, Semantics-based coding