一个 MPEG-4兼容的人脸动画系统

王奎武 王 洵 董兰芳 陈意云
(中国科学技术大学计算机科学与技术系 合肥 230027)
(kwwang@ mail.ustc.edu.cn)

摘 要 MPEG-4是一个基于对象的多媒体压缩标准,允许将场景中的音频视频对象(自然的或合成的)独立编码.它能够将人脸动画和多媒体通信集成在一起,并且可以在低带宽的网络上控制虚拟人脸.首先介绍 MPEG-4中关于人脸动画的基本概念,然后提出一个 MPEG-4兼容的人脸动画系统.

关键词 MPEG-4,人脸动画,脸部定义参数,脸部动画参数,语音动画
中图法分类号 TP391.41

AN MPEG-4 COMPLIANT FACIAL ANIMATION SYSTEM

WANG Kui-Wu, WANG Xun, DONG Lan-Fang, and CHEN Yi-Yun
(*Department of Computer Science and Technology, University of Science and Technology of China, Hefei* 230027)

Abstract MPEG-4 is an object-based multimedia compression standard, which allows the encoding of different audio visual objects (natural or synthetic) in the scene independently. It enables integration of face animation with multimedia communications and allows the face animation over low bit rate communication channels. First given in this paper is an overview of the basic technology about facial animation in MPEG-4, and then an MPEG-4 compliant facial animation system is proposed.

Key words MPEG-4, facial animation, facial definition parameters (FDPs), facial animation parameters (FAPs), speech animation

1 引 言

传统的网络交互手段主要是基于文字,如果能够增加一个可以说话的虚拟人脸,将大大提高交互的效率和趣味性.目前国内外有很多学者在从事这方面的研究,已经取得了一定的成就[1~6]. ISO的 MPEG-4标准的提出,进一步拓宽了虚拟人脸的应用,使得即使在低带宽的网络上也可以实现高质量的人脸动画.

MPEG-4采用了基于对象的编码方法,将一个场景看成是由一些音频视频对象(audio video object, AVO)组成.一个 AVO可以是自然的,也可以是合成的. MPEG下面有一个组织专门致力于研究合成自然混合编码(synthetic and natural hybrid coding, SNHC).人脸对象(facial object)是 MPEG-4中一个非常重要的对象.在 MPEG-4第 1版中, SNHC对三维人脸动画做了标准化.但是它只是定义了一个格式,对于具体问题,并没有给出解决方案,这就给研究者提供了广阔的空间[7,8].本文就是在 MPEG-4的基础上设计了一个人脸动画系统.在接下来的几节中,第 2节介绍 MPEG-4中关于人脸动画的基本概念;第 3节介绍我们提出的与 MPEG-4兼容的人脸动画系统;最后一节给出结论和进一

原稿收到日期: 2001-03-07
本课题得到中国科学技术大学青年科学基金资助

步的工作.

2 MEPG-4中的人脸动画

在 MPEG-4中,一个三维(或二维)的人脸对象是人类脸部的一种结构化的表示,通过人脸对象能够获得说话者说话时的口型,识别出他(她)的情绪,从而达到栩栩如生的效果.

在动画阶段,人脸对象通过一个叫脸部动画参数(facial animation parameter, FAP)的流来驱动. FAP操纵人脸网格模型上的一些特征点来产生可视音素和各种表情.利用 FAP来控制远程终端上的人脸对象,即可获得实时的、高度真实感的场景序列,这样就避免了每一帧都要传输人脸图像.

但是上述情况 FAP只是控制解码端私有的人脸模型.更复杂的情况是解码端从编码端下载一些脸部定义参数(facial definition parameter, FDP)来初始化私有的人脸模型.这样编码端希望的人脸的纹理信息和几何信息都可以表示出来[2,7].

下面的几个小节分别介绍一下关于人脸对象的几个概念.

2.1 FAP集

FAP很接近于脸部的肌肉运动,表示了一个完整的基本脸部运动集合. FAP能够表示绝大部分的自然表情,通过赋给 FAP一些夸张的值,还能表示卡通人物非常夸张的表情. MPEG-4定义了 68个 FAP,分成 10组,其中包括了两个高层次的参数: 可视音素(visual phoneme, Viseme)和表情. MPEG-4对每个 FAP只给出了文字描述,表 1~表 3是其中的一部分 FAP.

表 1 可视音素

viseme-select	Phonemes	Example
0	None	Na
1	P, b, m	put, bed, mill
…	…	…

表 2 表 情

expression-select	Expression name	Textual description
0	Na	Na
1	Joy	眉毛自然放松,嘴张开,嘴角朝耳朵方向收缩
…	…	…

表 3 FAP定义、单位、方向、分组和步长

#	FAP name	FAP description	Units	Uni or Bidir	Pos motion	Grp	FDP subgrp num	Quant step size
1	Viseme	Set of values determining the mixture of two visemes for this frame (e.g. pbm, fv, th)	Na	Na	na	1	Na	1
2	Expression	A set of values determining the mixture of two facial expression	Na	Na	na	1	Na	1
3	Open-jaw	Vertical jaw displacement (does not affect mouth opening)	MNS	U	down	2	1	4
…	…	…	…	…	…	…	…	…

2.2 中性状态人脸和脸部动画参数单位

在一个视频序列的开始,人脸被假定处于中性位置.不为 0的 FAP值被解释成从中性人脸出发的某个偏移.中性人脸的定义如下:

① 眼睛注视 Z轴的正方向;② 所有的脸部肌肉松弛;③ 眼睑与虹膜相切;④ 瞳孔是虹膜直径的 1/3;⑤ 上下嘴唇互相接触,唇线是水平的,与嘴角处于同一高度;⑥ 嘴是紧闭的,上下牙齿闭合;⑦ 舌头是舒展的,舌尖在上下牙齿之间.

对于不同的人脸模型,同样的 FAP值希望得到相同的动作(主要指幅度),但是又要避免涉及到过多的模型细节. MPEG-4定义了脸部动画参数单位(facial animation parameter unit, FAPU)来达到上述效果. FAPU与中性人脸上某些关键特征之间的距离成正比.具体定义见表 4.

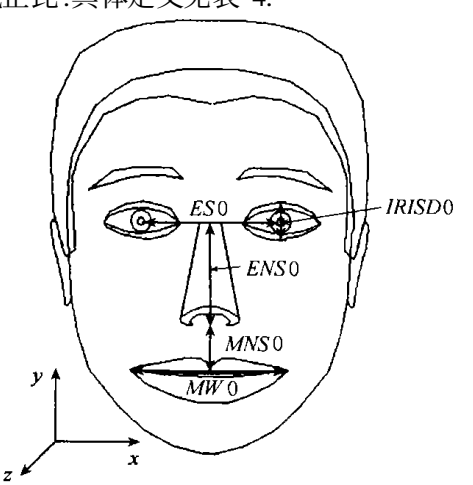


图 1 中性状态人脸

表 4　脸部动画参数单位

Description		FAPU Value
$IRISD0 = 3.1.y - 3.3.y = 3.2.y - 3.4.y$	虹膜直径	$IRISD = IRISD0/1024$
$ES0 = 3.5.x - 3.6.x$	两眼球的距离	$ES = ES0/1024$
$ENS0 = 3.5.y - 9.15.y$	眼鼻的距离	$ENS = ENS0/1024$
$MNS0 = 9.15.y - 2.2.y$	口鼻的距离	$MNS = MNS0/1024$
$MW0 = 8.3.x - 8.4.x$	嘴的宽度	$MW = MW0/1024$
AU	角度单位	10^{-5} rad

2.3　FDP集

FDP是用来将解码端的私有人脸模型定制成一个特定的人脸模型,或者从编码端传输一个人脸模型到解码端. FDP集采用 BIFS语法 (类似于 VRML)描述.典型的工作方式是: 对话的开始时,传输 FDP集合,然后跟随一个压缩的 FAP流,解码端使用 FDP校准私有模型,使用 FAP流进行动画. 在 FDP集合中定义了 84个特征点来表示人脸的几何信息.其参考位置如图 2所示.

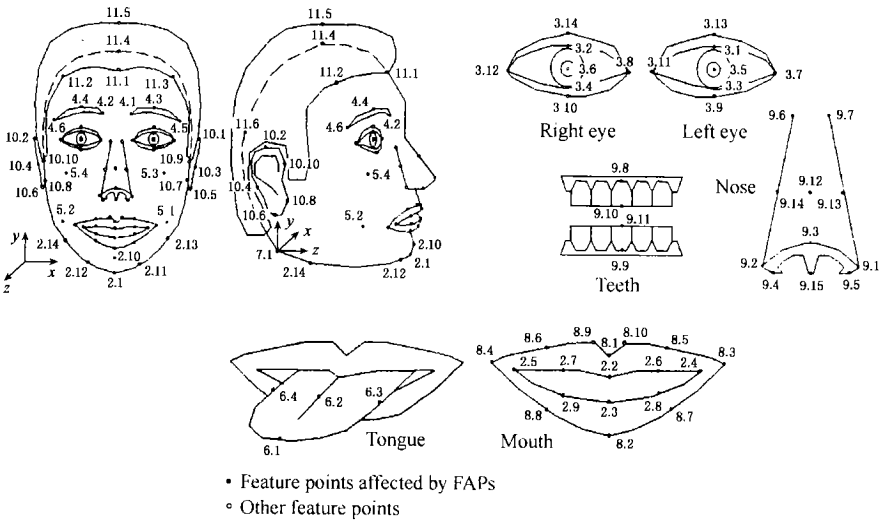


图 2　特征点

3　MPEG-4兼容的人脸动画系统

根据使用 FDP数据的方式的不同,一个人脸动画系统可以分成如下 4种工作方式:

(1) 不使用 FDP(只有 FAP信息).解码端接收到的 FAP流用于控制未经校准的私有模型.因为 FAP使用 FAPU作为单位,所以即使没有经过校准也能产生不错的结果.

(2) 只使用特征点. FDP中包含图 2所示的特征点,这些特征点用来校准解码端的私有模型,然后使用 FAP控制校准过的人脸模型.这样编码端可以部分预知动画的结果.

(3) 使用特征点和纹理信息. FDP中除了包含特征点,还有纹理图和特征点对应的纹理坐标.

(4) 使用脸部动画表 (facial animation table, FAT)和三维模型.一个新的三维模型从编码端传送到解码端,脸部动画表用来定义 FAP如何控制这个新模型.这时编码端对解码端的动画进行完全控制,能够全部预知动画的结果.

其中 (4)只需要解码端作很少量的工作即可进行动画,所以目前本文提出的系统没有考虑这种情况.这里主要考虑前面的 3种情况.系统的结构框图如图 3所示.下面结合图 3分别介绍一下我们的人脸动画系统的各个组成部分.

3.1　建　模

人脸动画的第一步是建模.到目前为止,自动建模仍然很困难. Volker Blanz和 Thomas Vetter通过扫描得到一个人脸模型库,然后对这个模型库进行主分量分析 (PCA),得到一组正交基.这些正交基的线性组合可以确定一个模型.然后将这个模型与输入图像拟合,不断修改线性组合的系数,从而达到最佳拟合[6].整个过程可以自动进行.但是比较费时,而且可以建立的模型很大程度上依赖于模型库

(比如,很难从西方人的模型库得到一个东方人的模型).我们所采用的方法是先建立一个通用的模型,然后在对这个通用模型进行校准,得到特定人的模型.整个过程需要少量的用户交互.

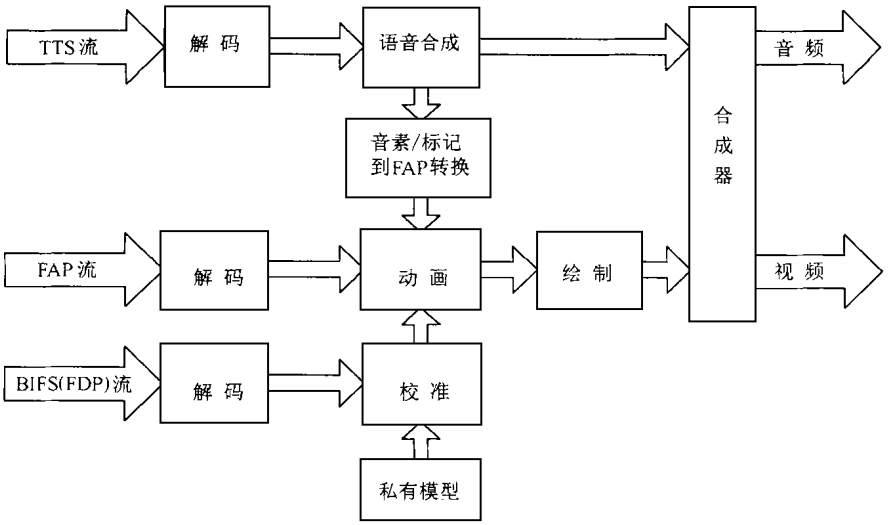
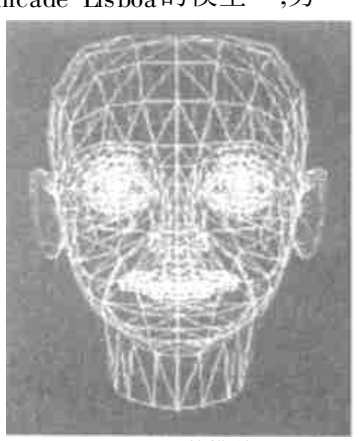


图 3 人脸动画系统框架图

3.1.1 建立私有模型

人脸模型采用是的网格模型,由点和面组成.系统可以接受以 VRML文件表示的模型数据.这些模型数据可以从真人扫描得到,也可以由艺术家手工绘制得到.我们在实验过程中采用了两个模型,一个是葡萄牙 Instituto Superior Técnico (LST), Universidade Técnicade Lisboa的模型[2],另一个是 AT&T的模型,分别如图 4(a)、(b)所示.建立人脸的网格模型以后,还需要在这个模型上定义特征点.这些特征点的位置根据图 2来确定.此外 MPEG-4还给出这些特征点位置的依赖关系.比如 3.1.x= (3.7.x+ 3.11.x)/2; 5.1.y= 8.3.y等等.图 4(c)表示定义了特征点的 IST模型.



(a) IST的模型

(b) AT&T的模型

(c) 在(a)上定义的特征点

图 4 网格模型

3.1.2 模型校准

如果接收到的 FDP数据中包含特征点的信息,就需要对客户端的私有模型作调整,使得这个模型符合编码端的要求.在我们的系统中,特征点的位置是从某个特定人的相片上提取的,因为需要三维信息,所以使用了正面和侧面的相片(图 5(a)、(b)).这种方法具有很广泛的应用,比如利用这个系统可以实现一个网络聊天室.

从图像中自动提取脸部特征是一个很有挑战性的课题,目前有很多学者正在研究.我们采用了比较简单的需要用户交互地方法.将私有模型分别投影到正面和侧面的相片上,提供给用户一个可视化的工具,交互的调整特征点的位置.

一个关键的问题是当特征点的位置改变时,网格点的位置如何随之改变.这是一个散乱数据插值的问题,有多种解决方法.其中的一种方法是在三维

空间中,建立 sibson局部坐标系,计算网格点的 sibson坐标.以 sibson坐标作为网格点相对于特征点的权值[4].

在将这种方法应用到人脸校准时,我们对文献[4]中的方法做了一些适当的改进[9].图 5(c)是 IST模型根据图 5(a)、(b)作校准后的模型.

(a) 正面相片

(b) 侧面相片

(c) 由(a)、(b)校准后的模型

图 5 模型校准

3.2 人脸动画

在接收到的 FAP数据中,如果某个 FAP不为零,模型就要改变,产生相应的动作(表情).这里主要有两个问题需要解决:一是这个 FAP影响到哪些网格点,二是这些网格点受影响的程度(权值).理想的情况是实现这个 FAP时不依赖于所使用的模型,这样能够很方便地更换模型.文献[2]使用 FAP直接控制网格点,这种做法虽然可以达到非常好的效果,但是依赖于所使用的模型,无法更换模型.文献[1]除了需要网格模型的几何信息(点和面)之外,还需要提供模型的语义信息,其中包括每一个 FAP影响的网格点,这种方法对模型的依赖性也比较大.

本文采用一种间接的方法,FAP作用在特征点上,再由特征点来影响网格点. FAP的影响区域是相应特征点的影响区域的并集.特征点的影响区域可以自动得到.在模型校准阶段,可以得到特征点的影响区域,但是这个影响区域是根据点的空间位置确定的,没有考虑点与点之间的连接关系,对于动画来说,这个影响区域有点大了.文献[3]采用边深度,进行宽度优先搜索来求影响区域.本文利用边深度求得的影响区域和校准时的影响区域求交得到一个影响区域,作为特征点动画时的影响区域.因为校准时的影响区域利用 Delaunay三角剖分求自然邻居得到,所以这种方法综合考虑了距离和边的连接关系.网格点相对于特征点的权因子使用第 3.2节中的 sibson坐标.采用这种技术实现的 FAP如图 6所示.篇幅所限,未给出全部的 FAP.利用这种技术实现的 FAP从效果上看与 IST大致相同,但是模型依赖性小很多.

(a) fap20close_t_r_eyelid
竖直移动上、右眼睑

(b) fap50head_roll
左右晃动头部

(c) fap64 bend_nose
鼻尖水平移动

图 6 FAP的实现

3.3 脸部附件

脸部附件包括眼球、牙齿、舌头、眼镜等.之所以把这些作为附件单独处理,是因为这些部分与脸部的其它特征相对独立.而且不同的人这些部分大体相同,在模型校准阶段,一般只需要对附件的大小进行缩放.它们(舌头除外)都是刚体,一般不会发生扭曲,动作局限于整体的平移、旋转.舌头的卷曲也很有规律性,可以看成舌尖绕一个轴旋转.

3.4 语音动画

本文的系统实现的语音动画是指输入一段文字,虚拟人脸模型能够模仿人在说这段文字时的口形.系统采用了两种文本到语音(text to speech, TTS)的开发包,一种是微软公司的 Speech API,另一种是中国科大讯飞公司提供的 TTS开发包.MPEG-4中的音素划分对于英语效果比较好,但应用在汉语上效果不佳.为此,我们利用中国科大讯飞公司在中文语音合成上的研究成果,对汉语发音的声韵母进行分类,将发音口形类似的声母分成一类,每一类认为口形相同.韵母分成单韵母和复合韵母两种.复合韵母可以由单韵母或单韵母和韵尾组成.其中声母 8组,单韵母 6组,韵尾 2组,复合韵母 26组.图 7表示了一些发音的口形.

(a) 汉语‘a’

(b) 汉语‘zhchsh’

(c) 汉语‘gkh’

图 7 可视音素

4 结论和进一步的工作

MPEG-4认为三维人脸动画有很好的应用前景,所以在它的标准中对人脸对象进行了专门的描述.根据这些描述,本文提出了一个人脸动画系统.这个系统稍加扩展,即可作为 MPEG-4用于人脸动画的客户端.事实上,我们的系统具有更多的功能,因而应用范围更广泛,比如说可以将这个系统应用于虚拟社区或虚拟聊天室,用户提供几幅图片,即可定制一个个性化的形象代表.

进一步的研究工作主要有:

(1) 纹理映射.纹理映射可以极大地提高人脸的真实感.纹理信息可以从 MPEG-4的编码端得到(如果编码端提供的话),也可以从本地的两幅相片中提取.究竟采用哪种方式取决于具体应用.纹理坐标可以在模型校准的时候产生.这部分的工作正在进行.

(2) 更换模型.目前试验的模型是 IST的模型,这个模型比较简单,在表示三维人脸时,显得比较粗糙.由于我们的方法跟模型的相关性很小,因此可以很容易地更换一个更好的模型.

(3) 口型和表情.由于 MPEG-4对口型和表情只作了一些文字性的描述,因此我们设计的口型和表情只是凭借自己的主观感觉,与真实的人脸还有比较大的差距.要得到更加逼真的效果,需要有美工人员精心设计 FAP的值,建立口型库和表情库.

(4) 语音驱动.系统与声音的接口采用了 TTS方式.进一步的工作可以考虑增加语音驱动,用户输入一段声音,可以产生相应的表情和动作[5].

参 考 文 献

1 Lavagetto F, Pockaj R. The facial animation engine: Toward a high-level interface for the design of MPEG-4 compliant animated faces. IEEE Trans on Circuits and Systems for Video Technology, 1999, 9(2): 277~289

2 Abrantes G, Pereira F. MPEG-4 facial animation technology: Survey, implementation, and results. IEEE Trans on Circuits

and Systems for Video Technology, 1999, 9(2): 290~305
3 Jun-yong Noh, Douglas Fidaleo, Ulrich Neumann. Animated deformations with radial basis functions. In: ACM Virtual Reality and Software Technology (VRST'1999). London: ACM Inc, 1999. 166~174
4 Moccozet L, Magnenat-Thalmann N. Dirichlet free-form deformations and their application to hand simulation. In: Proc of Computer Animation'97. IEEE Computer Society, 1997. 93~102
5 Matthew Brand. Voice puppetry. In: SIGGRAPH'99. Los Angeles: ACM Inc, 1999. 21~28
6 Volker Blanz, Thomas Vetter. A morphable model for the synthesis of 3D faces. In: SIGGRAPH'99. Los Angeles: ACM Inc, 1999. 187~194
7 ISO/IEC 14496-2. Information technology—Coding of audio-visual objects: Visual. ISO/IEC JTC1/SC29/WG11 N2202. Tokyo, 1998
8 高 文,吴 枫. MPEG-4编码的现状和研究. 计算机研究与发展, 1999, 36(6): 641~652
(Gao Wen, Wu Feng. Researches and developments of MPEG-4 coding. Journal of Computer Research and Development (in Chinese), 1999, 36(6): 641~652)

王奎武 男,1977年生,硕博连读研究生,研究方向为计算机图形学、多媒体通信.

王 洵 男,1966年生,中国科学技术大学计算机科学技术系副教授,主要研究方向为计算机图形学、并行与分布计算.

董兰芳 女,1970年生,中国科学技术大学计算机科学技术系讲师,主要研究方向为计算机图形学、软件体系结构.

陈意云 男,1946年生,中国科学技术大学计算机科学技术系教授,博士生导师,主要研究领域为多媒体通信、软件体系结构.

欢迎参加《计算机辅助几何设计》高级研讨班

为推动我国计算机辅助几何设计 (CAGD)与计算机图形学 (CG)事业的发展,经教育部批准,清华大学国家 CAD工程中心将于 2001年 6月 21日至 6月 30日,举办第一届有关计算机辅助几何设计 (CAGD)理论与方法的高级研讨班.

高研班参加人员主要为高等学校学术带头人、学科骨干和后备力量,在 CAD与图形学领域取得一定成果或有发展潜力的教学、科研人员.本次高研班由孙家广院士主持.拟聘请国家 CAD工程中心的学术委员及国内外知名专家学者作学术报告. SIAM几何设计委员会主席、CAGD杂志编委、加州大学 Davis分校 Rida Farouki教授将作特邀报告.

有意参加的同志请从高研班网页了解高研班详情.

网址: http://ncc.cs.tsinghua.edu.cn/~CAGD