MPEG-4 中脸部动画参数和序列重绘的肌肉模型

虞 露
(浙江大学信息与通信工程研究所, 杭州 310027)

摘 要 MPEG-4 中定义了"人脸对象"这样一种特殊的视频对象,并通过脸部动画参数 FAP 和脸部定义参数 FDP 来对这类对象进行编码,以实现极低码率的视频编码. 通过对 MPEG-4 中"人脸对象"这类视频码流的句法结构和参数编码方法的详细分析. 以及通过对 MEPG-4 解码器图象重绘(rendering)过程的研究,在 Waters 的以肌肉收缩强度为参数的肌肉模型基础上,提出了更适应于 MPEG-4 参数的位移控制肌肉模型(displacement-controlling muscle model),从而实现了通过利用 MEPG-4 码流中的 FAP 和 FDP 参数来重建自然表情的人脸视频序列.

关键词 MPEG-4 FAP FDP 人脸重绘 肌肉模型
中图法分类号: TN919.8 **文献标识码**: A **文章编号**: 1006-8961(2001)01-0036-06

MPEG-4 Facial Animation Parameters and Muscle-Model for Sequence Rendering

YU Lu
(*Institute of Information and Communication Engineering, Zhejiang University, Hangzhou* 310027)

Abstract Face object, is a special visual object defined in MPEG-4. Facial definition parameter (FDP) and facial animation parameter (FAP) are the sets of parameters to calibrate and animate the face object. The bitstream syntax for FDPs and FAPs in MPEG-4 are analyzed in this paper. The meaning of high level FAPs, viseme and expression, are explained with examples. All FAPs can be compressed with mask-based DPCM and/or DCT encoding. A rendering technology—— displacement-controlling muscle model is proposed to reconstruct video sequences with natural facial expression from decoded FAPs and FDPs in MPEG-4 bitstream. There are two muscle models for different muscle stryles, parallel muscle and orbicular muscle. Different from traditional muscle models, which are controlled by intensity of muscle contraction, the proposed models are controlled by displacement of key-points affected by contraction of muscle. The reason behind this technology is that the visible displacement of key points is easier to be extracted from orighnal video sequence then the invisible muscle contraction. As an experiment result, reconstructed "Mona Lisa" with different expression is illustrated.

Keywords MPEG-4, FAP, FDP, Facial image rendering, Muscle model

0 引 言

传统的图象压缩方法是将任何图象都划分成大小一定的方块,然后用预测和变换等方法进行压缩,但这种方法在低码率时,容易暴露出方块效应等许多不可克服的缺点,所以有人提出了面向对象(Object-Oriented)的编码方法[1]. 将图象合理地划分成物体、背景等与实际景物相对应的对象平面(Object Plane)[2];然后对各对象平面的形状、纹理和运动进行编码. 显然,这样要合理得多,在恢复图象上也不会出现人为的方块. 这是 MPEG-4 标准的一个核心思想.

在 MPEG-4 视频流中定义了视频序列、静止纹

基金项目:国家自然科学基金资助项目(69782004)
收稿日期:2000-01-05;改回日期:2000-05-22

理、网格结构等一些普适的对象,还定义了“人脸对象”这样一类较特殊的对象,因为图象中出现得最典型的物体就是人本身了,而最受观众关注的部分就是人脸. 对于这一类物体,人们往往借助一个专用人脸模型(比如 Linkoping 大学提出的 CANDIDE 模型[3]或浙江大学提出的模型[4,5]),通过只传输一些模型参数来达到高效压缩的目的,也就是说,对人脸对象,MPEG-4 用的是参数编码方法来实现极低码率的编码. 这些参数包括用于描述特定脸形的 FDP (Facial Definition Parameter)参数和描述脸部活动的 FAP(Facial Animation Parameter)参数.

与传统的视频解码器不同,MPEG-4 解码器在恢复了人脸对象的所有参数之后,并不能看到恢复图象的结果,而是需要有一个重绘(rendering)过程才能重新展现被压缩的图象. 该重绘过程即是在人脸模型的基础上,根据人脸肌肉运动的一些规律,将抽象的 FAP 数值转化为具体的、可供显示的图象阵列.

实际上,MPEG-4 不仅可对自然景物图象进行编码,也可对人工图象甚至自然/人工合成图象进行有效的编码,比如用一组 FAP 参数可以产生出一组实际人脸做不出的夸张表情,用于动画片、特技制作等. MPEG-4 的人脸对象编码不仅可用于多媒体会议、可视电话等双端通信系统,也适合影视制作、人机界面、读唇助听等众多单端应用.

本文将在讨论符合 MPEG-4 句法结构的人脸对象参数及其压缩编码和解码的基础上,重点解决如何运用位移控制肌肉模型,以便根据恢复的 FAP 参数来重绘真实感人脸序列.

1　人脸参数 FDP 和 FAP

MPEG-4 定义了如下的 FDP 数据结构:

```
FDP {
        FeaturePointsCoord
        TextureCoords
        UseOrthoTexture
        FaceDefTables
        FaceSceneGraph
        }
```

其内容包括脸部线框模型或对已有模型的校准(faceDefTables、featurePointsCoord)、线框上的纹理(textureCoords 、faceSceneGraph)以及纹理映射到线框上的投影方式(useOrthoTexture)是柱面投影,还是正交投影. 这些内容文献[6]中有更详细的说明.

FDP 可以用来下载一个人脸模型,以及如何使之根据后续的一系列 FAP 参数,而产生活动图象的一套规则,因此,一般来讲,FDP 只在一个场景序列中出现一次,且出现在序列的开头. 若序列没有提供模型下载,也可以用 FDP 定义的特征点坐标来调整解码端已有的通用人脸模型,使之成为特定的人脸模型,也就是说,FDP 可以携带模型信息或模型调整信息. 若解码器没有收到 FDP 信息,也还是可以根据收到的 FAP 和本地的模型直接绘制人脸动画. 这种情况往往出现在单端系统,而非通信系统中,或者出现在恢复图象与原始图象有相同的运动和表情,但表现在完全不同的脸上的情况下.

与静态的 FDP 参数相对应的是动态的 FAP 参数,而 FAP 是一个完整的脸部基本运动的集合,它是基于对人脸细微运动的研究,与脸部肌肉运动密切相关,所以用 FAP 可以描述自然的脸部表情,当然也可以创作出夸张的非自然所能达到的表情.

在 MPEG-4 中,FAP 参数分成 10 组,包括口形和表情、下巴、眼部、眉毛、脸颊、舌、头部转动、嘴唇、鼻子、耳朵等,共 68 项[2]. 其中第一组是口形参数(viseme)和表情参数两个高层参数,与其它稍有不同.

其中,口形参数是与音素(phoneme)相对应的视频参数,它代表了一定发音的嘴部形状. MPEG-4 标准中例举了 14 种对应不同音素的口形(表 1).

表 1　口形参数的可选值

viseme_select	对应的音素	实例
0	None	Na
1	p,b,m	put,bed,mill
2	f,v	far,voice
3	T,D	think,that
4	t,d	tip,doll
5	k,g	call,gas
6	tS,dZ,S	chair,join,she
7	s,z	sir,zeal
8	n,l	lot,not
9	R	Red
10	A:	Car
11	E	Bed
12	I	Tip
13	O	Top
14	U	Book

注:实例中字母下划线,指所对应的音素

另外,还可以利用一系列连续的静态口形参数,重绘出一个视频序列. MPEG-4 中口形参数的数据结构如下:

viseme(){	取值范围
Viseme_select1	0～14
Viseme_select2	0～14
Viseme_blend	0～63
Viseme_def	0～1
}	

即每一次口形参数给出了由两种口形叠加而形成一帧图象所需要的那些参数值. 叠加而成的口形为:

final viseme = (viseme 1) × (viseme_blend/63) + (viseme 2) × (1 − viseme_blend/63)

例如在发音"f-A:"这样一个过程持续了 5 帧图象,则口形参数 Viseme_select1 = 2, Viseme_select2 = 10,而各帧的 Viseme_blend 参数可以分别为 63,48,32,16,0. 由这样重绘得到的图象序列即可看到,口形从对应于音素"f"(第 1 帧)转变为对应于音素"A:"(第 5 帧)的一个过程.

当 viseme_def 被置 1 时,则当前嘴部 FAP 参数的组合就定义了当前所选的口形,并在解码器中保存这种定义. 这样当下一次选中同样的口形时,这组 FAP 就可以直接引用,而不需要再传输具体有关的 FAP 参数了. 也就是说,viseme_def 即可帮助在解码器中建立起一张口形的定义表或数据库.

另一个高层 FAP 参数——表情参数是从情绪、情感等心理角度来描述脸部视象的,其实每一个表情参数就对应了一组表情控制点的移动. 这种表情参数的数据结构如下:

Expression(){	取值范围
Expression_select1	0～6
Expression_intensity1	0～63
Expression_select2	0～6
Expression_intensity2	0～63
init_face	0～1
Expression_def	0～1
}	

一般人脸的基本表情可分成喜悦、悲伤、愤怒、恐惧、厌恶和吃惊 6 种,对应的 Expression_select 标号分别为 1～6,标号 0 表示为中性表情. 每一帧图象中的表情也同口形的形成相似,是由两种被选中的表情叠加而成,即

final Expression = Expression1 × (Expression_intensity1/63) + Expression2 × (Expression_intensity2/63)

如果 Expression_def 被置 1,则当前的 FAP 参数就定义了所选的表情,并形成一张 FAP 表,这张 FAP 表可以在以后用到同样的表情时,减少需要传输的信息. 如果 init_face 被置 1,那么在使用 3～68 的 FAP 参数之前,先要从闭嘴、张眼、视线、头朝向等方面将人脸模型调整为一个"中性的脸".

MPEG-4 中为了统一脸部运动的起始状态,对"中性的脸"作出了较为明确的定义,即:

(1) 头部轴线平行于坐标轴;

(2) 视线方向与 Z 轴方向一致;

(3) 所有脸部肌肉放松;

(4) 眼睑与虹膜相切;

(5) 瞳孔直径是虹膜直径的三分之一;

(6) 双唇接触,唇线水平,并与嘴角在同一水平线上;

(7) 嘴闭合,上下齿扣合;

(8) 舌头水平平坦,并且舌尖与上下齿缘相触.

实际上,一般一个序列的开始,都先将脸调整为"中性的脸",然后再以 FAP 参数描述脸部运动.

除了口形和表情这两个高层的 FAP 参数之外,标准中还定义了其它 9 组,共 66 个 FAP 参数,这些参数代表了脸部表情的最基本运动,例如"左上眼睑闭合"(左上眼睑的垂直位移)、"右侧眉毛挤压"(右侧眉毛的水平位移)、"左眼珠侧视"(左眼珠水平方向转角)等等. 因为参数比较多,在这里就不一一例举了,详细内容请参见文献[2]. 这些参数的值,除了转动参数用弧度作单位以外,平动参数都选模型上某一特征距离的若干分之一作为单位. 这些特征距离包括双眼间距、眼鼻间距、鼻嘴间距、嘴宽、虹膜直径等. 这样就可保证 FAP 参数是相对位移,即便用在不同的模型上,也能得到一致的结果.

2 FAP 的压缩编码

对原始图象序列而言,FDP 和 FAP 参数已经是对序列的一种压缩表示了,但为了提高压缩效率,对这些参数还要进一步作压缩编码.

2.1 FAP 掩模

实际上由于 68 个 FAP 参数并非每个序列中都用到,所以可用掩模的方法来指明哪些参数出现在码流中. 这里用一种两层掩模方式,第 1 层掩模,对每组 FAP 用一个 2bit 信息(fap_mask_type)来指

明本组参数的传递为以下 4 种情况的哪一种:

(1) 码流中不包含本组 FAP;

(2) 码流中包含本组 FAP,并用第 2 层组内掩模(fap_group_mask)来指明本组内哪些 FAP 将被传递,然后对那些没有传递的参数用以前的值取代;

(3) 码流中包含本组 FAP,并用第 2 层组内掩模(fap_group_mask)来指明本组内哪些 FAP 将被传递,然后对那些没有传递的参数则由解码器插值产生;

(4) 码流中包含本组所有 FAP.

2.2　FAP 参数编码

经 FAP 掩模标明的将出现在码流中的具体 FAP 参数值,还要经过预测编码或 DCT 编码压缩. 一般 MPEG-4 对脸部对象定义了一套句法,它既可以用单个脸部对象平面的时间序列来组织,也可以用脸部对象平面组(又称为段 segment)的时间序列来组织,其中,脸部对象平面组由 16 个脸部对象平面构成. 该两种序列分别用不同的方法进行参数编码,即前者用基于帧的预测编码;后者则用 DCT 编码. 这里,基于帧的预测编码是在对每个 FAP 参数与前一帧相应参数进行差分、量化和自适应算术编码后,即写入码流;而 DCT 编码则是将一个段内连续的 16 个相同的 FAP 参数进行 DCT 变换、量化,并按行程长度作 Huffman 编码,最后写入码流.

值得指出的是,即使是在脸部对象平面组中,其它参数都用 DCT 编码,但第一组 FAP 参数,即口形和表情参数,还是采用预测编码,而不是 DCT 编码.

3　利用 FAP 生成脸部序列

MPEG-4 作为国际标准只定义了码流结构,而对人脸对象而言,它也只规定了有哪些参数,它们分别出现在码流的什么位置,以及经过怎样的变换可以恢复,等等. 如图 1 所示,在一个完整的编解码系统中,只有参数编解码部分包括在标准范畴之内,而参数的提取和图象的重绘,则在标准范畴之外,但这

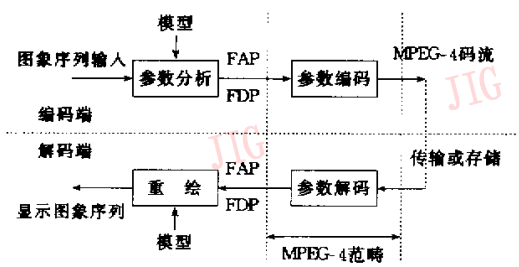


图1　MPEG-4人脸对象的编解码系统

两部分又是实际系统不可或缺的. 笔者已经在图象分析,对提取脸部形状、运动等参数进行了一些研究[7,8]. 另外,参数分析过程对动画制作等某些应用来说,虽然也可以用人工给出的方式来代替,但是重绘过程是重现图象所不能逾越的. 其实,图象的重绘对国际标准来说是一个开放的过程,也就是说,可以根据不同的应用来采用完全不同的方法,而最终的重绘效果如何,却恰恰取决于此,因此,这里主要结合通信系统中人脸表情图象序列的生成来讨论.

重绘过程需要的基础条件是一套 FDP 参数(就是一个人脸模型及其覆盖在上面的脸部纹理色彩)、一套 FAP 参数(就是某些脸部特征点,或称表情控制点的位移)和一套规则. 前面的分析已说明 FDP 和 FAP 可从码流中得以恢复,留下的是这一套规则的设计问题.

这套规则解决的是如何将控制点的运动转化为多边形模型上各有关顶点的运动,进而得到各象素的运动,然后根据一个已知的纹理(比如第一帧图象)重建出后续的视频序列,因此这就需要一个由控制点牵扯周围点运动的模型. Water 通过对脸部肌肉的研究,给出了一个较好的肌肉模型[9],但这个模型需要提供肌肉收缩强度来作为参数,而这个参数在 FAP 中是没有的,也不容易从图象分析过程中提取,所以本文对此模型进行了修改,即以控制点的位移来取代肌肉收缩强度,以作为图象重绘的参数,并将它称为位移控制肌肉模型(displacement-controlling muscle model).

3.1　位移控制肌肉模型

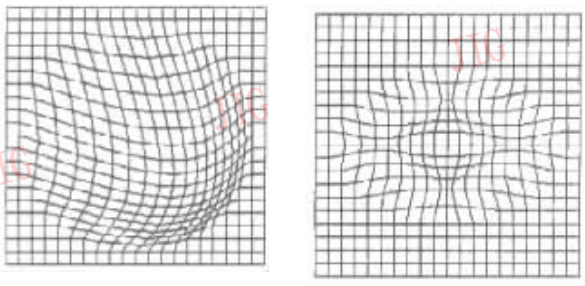
模型分成两种,分别对应于平行肌和轮匝肌.

如果控制点是平行肌附着在骨上的点,比如说眉毛的两端,那么,可以认为最大的位移发生在控制点处,而且位移量在控制点的周围组织扩散并衰减. 控制点周围的点(x,y)的位移表示为

$$D_x(x,y)=\Delta x\times\cos\left(\frac{d}{R}\times\frac{\pi}{2}\right) \tag{1}$$

$$D_y(x,y)=\Delta y\times\cos\left(\frac{d}{R}\times\frac{\pi}{2}\right) \tag{2}$$

其中,$\Delta x,\Delta y$ 是控制点的位移;d 是点(x,y)与控制点之间的欧拉距离;R 是最大的影响范围. 在二维状态下,平行肌收缩引起的效果如图 2(a)所示.



(a) 平行肌模型 (b) 轮匝肌模型

图 2 肌肉收缩效果模型

如果运动是由轮匝肌收缩或扩张引起的(比如嘴),那么,衰减即在椭圆的周围组织扩散,本文将控制点所在的椭圆称为控制椭圆,且认为控制椭圆上的点有最大的位移. 这里假设控制椭圆是

$$\frac{x^2}{a^2}+\frac{y^2}{b^2}=1 \tag{3}$$

则控制椭圆内部相邻部分的位移表示为

$$D_x(x,y)=\Delta x\times\sin\left(d\times\frac{\pi}{2}\right)\quad(d\leq 1) \tag{4}$$

$$D_y(x,y)=\Delta y\times\sin\left(d\times\frac{\pi}{2}\right)\quad(d\leq 1) \tag{5}$$

而控制椭圆外部相邻部分的位移是

$$D_x(x,y)=\Delta x\times\cos\left(\frac{d-1}{R-1}\times\frac{\pi}{2}\right)\quad(1<d\leq R) \tag{6}$$

$$D_y(x,y)=\Delta y\times\cos\left(\frac{d-1}{R-1}\times\frac{\pi}{2}\right)\quad(1<d\leq R) \tag{7}$$

其中,$\Delta x,\Delta y$ 是控制椭圆在长轴和短轴方向上收缩或扩张的位移量;且 d 的定义可修改为

$$d=\frac{x^2}{a^2}+\frac{y^2}{b^2} \tag{8}$$

且 $R>1$,即影响范围为比控制椭圆大的一个椭圆区域. 以上提到的位移仅适用于第一象限的点,而其它象限中点的位移符号,则应根据点的位置而作相应的修改. 一般在二维状态下,轮匝肌收缩引起的效果如图 2(b)所示. 上述两个模型很容易扩展到三维情况.

由于脸部表情的真实情况可能更复杂一些,所以,将上述的位移控制肌肉模型应用到头部线框模型上时,应再做一些修改. 比如,有些点虽然在控制点的影响范围之内,但因为落在其它五官上,不能随着控制点移动,或移动量要减小,所以,一些必要的知识应加到本文所提出的规则中去.

3.2 实验结果

为了说明使用 FAP 参数,结合本文提出的位移控制肌肉模型重绘图象的效果,本文以“蒙娜丽莎”图象为例,采用浙江大学的人脸线框模型,得到了视觉感觉自然的“喜悦”(图 3b)和“悲伤”(图 3c)的表情图象.

(a) “蒙娜丽莎”原图 (b) 嘴角上拉,眉毛内侧上抬 (c) 嘴角下拉,眉毛内侧上抬

图 3 脸部表情图象生成实例

4 结 论

本文针对人脸这一在视频序列中出现最多、最为人关注的对象,对 MPEG-4 中有关的句法及 FDP 和 FAP 参数的编码作了详细的分析,而且作为解码器中最重要的一环,重绘过程直接影响着重建图象的质量,因此本文在 Waters 的肌肉模型的基础上,对其作了修改,提出了新的位移控制肌肉模型,使之适应于用 FAP 所含的控制点位移进行表情图象的重建. 在这些研究的基础,本文已在实验室建立起一套极低码率下脸部视频通信的仿真系统,它可应用于可视电话、会议电视、远程教育等领域,也可拓展到虚拟现实、动画特技制作、人机界面等应用.

参 考 文 献

1 Musmann H G, Hotter M, Ostermann J. Object-oriented analysis-synthesis coding of moving images. Signal Processing:

Image Communication, 1989,1(2):117～138.

2 ISO/IEC 14496-2, INFORMATION TECHNOLOGY-CODING OF AUDIO-VISUAL OBJECTS:Visual, FDIS, 1999.

3 Rydfalk M. CANDIDE——A parameterized face. Dep. Electronics Engineering Rep. LITH-ISY-I-0866, Linkoping University, Sweden,1987.

4 周峰. 脸部知识基图象编码的研究及 34MB/s 复合彩色电视图象 DPCM 编码硬件系统的研制[博士学位论文],杭州:浙江大学,1991.11.

5 虞露,周峰,姚庆栋. 脸部序列图象的模型基编码. 通信学报,1997,18(10):1～6.

6 ISO/IEC 14496-1, INFORMATION TECHNOLOGY-CODING OF AUDIO-VISUAL OBJECTS: System, FDIS, 1999.

7 Lu Yu, Jingyu Zhang. Facial animation reconstruction from FAP. In: SPIE Visual and Image Communications and Processing '2000,San Jose, USA,2000,3974:986～993.

8 Yunhai Liu, Lu Yu, Qingdong Yao. Automatic extraction of facial features using deformable templates. In: Picture Coding Symposium'99,Portland, USA, 1999:189～192.

9 Waters K. A muscle model for animation three-dimensional facial expression. Computer Graphics, 1987,21(4):17～24.

虞　露　1969 年生,博士,副教授,1996 年 6 月获浙江大学取得通信与电子系统专业博士学位,现在浙江大学信息与通信工程研究所从事教学科研工作.主要感兴趣的研究领域有 MPEG Audio/Video 压缩编码、多媒体通信、DSP 和 ASIC 的设计开发和应用研究.