人脸建模与动画的研究*)

王 洵 董兰芳 万寿红
(中国科学技术大学计算机科学技术系 合肥 230027)

Study of Face Modeling and Animation

WANG Xun DONG Lan-Fang WAN Shou-Hong
(Department of Computer Sci. & Tech. ,University of Sci. & Tech. of China,Hefei 230027)

Abstract Face modeling and animation is one of the most challenging problems in Computer Graphics. In this paper, we describe our study of face modeling and animation, especially of three-dimensional model-based facial animation. Our study includes the following aspects: developing a face model editor; realizing face model calibration; generating a realistic face image; developing a MPEG-4 compliant facial animation system; developing two speech animation systems, one is based on KD2000, the other is based on SAPI5.0.

Keywords Face modeling, Facial animation, Speech animation, MPEG-4, Viseme

1. 引言

人脸建模与动画(face modeling and animation)是计算机图形学中最富有挑战性的课题之一[1,2]。这是因为:首先,人脸的几何形状非常复杂,其表面不但具有无数细小的皱纹,而且呈现颜色和纹理的微妙变化,因此建立精确的人脸模型、生成真实感人脸非常困难;其次,脸部运动是骨骼、肌肉、皮下组织和皮肤共同作用的结果,其运动机理非常复杂,因此生成真实感人脸动画非常困难;另外,我们人类生来就具有一种识别和理解脸部表情的神奇本领,任何微妙的表情变化都能够立即觉察出来,这就使得人脸建模与动画变得更加困难。

1.1 人脸动画的分类

人脸动画主要分为两种类型,即基于样本(sample-based)的人脸动画和基于三维模型的人脸动画。

基于样本的人脸动画也称为数据驱动(data-driven)的人脸动画,这种方法类似于语音合成中的波形拼接合成法,不需要建立三维人脸模型,而是通过对给定的样本(一般是一段说话的录像)重新进行组织来生成新的人脸动画[3~5]。由于这种方法直接取材于样本,因此其最大优点就是真实感非常强,缺点之一是需要一段真人说话的录像,数据获取不太方便,缺点之二是没有建立三维人脸模型,视点不能变化或只能小范围变化。这种方法一般用于只需要少量虚拟人物但是需要高度真实感的应用领域,如电影与电视的虚拟演员、网络虚拟主持人等。

基于三维模型的人脸动画首先建立三维人脸模型,接着驱动人脸模型生成语音与口型同步播出的语音动画(speech animation),人脸在说话时还可以带有各种表情。虽然这种方法目前在真实感方面还比不上基于样本的人脸动画,但是由于这种方法数据获取简单(只需要几张不同角度的照片),制作方便(不需要或只需要少量用户交互),能够生成真三维动画,因此这种方法可以广泛应用于需要大量虚拟人物但是真实感要求不太高的应用领域,如网络虚拟社区、三维语音动画聊天室、有声 Email 等。本文研究基于三维模型的人脸动画。

1.2 基于三维模型的人脸动画

基于三维模型的人脸动画主要有以下几种方法:(1)关键表情插值法[6,7];(2)参数化模型[7,8];(3)基于物理的肌肉模型[9~11];(4)伪肌肉模型,包括抽象肌肉动作模型[12]、有理自由变形[13]、径向基函数变形[14]等;(5)表演驱动的人脸动画[15,16];(6)参数化模型与肌肉模型相结合的混合方法[17]。

1.3 本文的组织结构

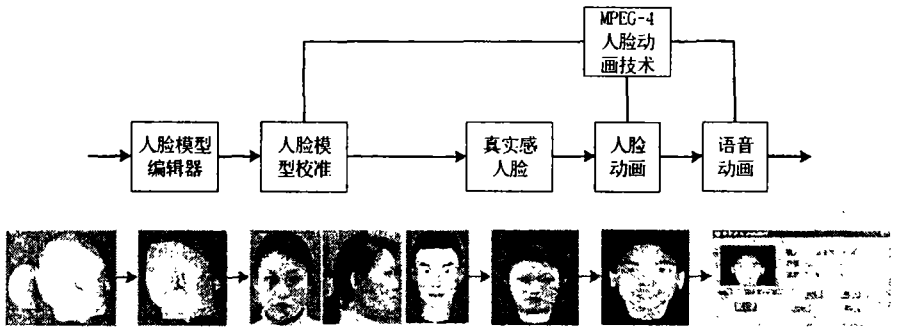


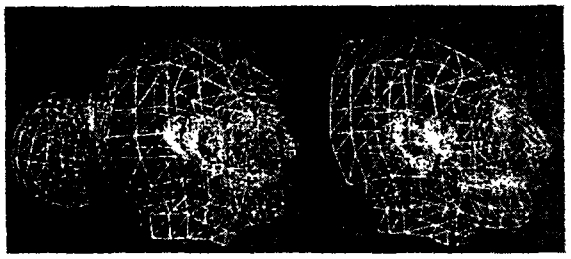
图 1 本文的组织结构

*)本课题得到安徽省自然科学基金(项目编号:01042203)、中国科学技术大学青年科学基金资助。王 洵 博士,安徽省信息产业厅副厅长、党组成员,中国科学技术大学兼职教授,主要研究方向为计算机图形学、多通道智能用户界面。董兰芳 硕士,讲师,主要研究方向为计算机图形学、软件体系结构。万寿红 硕士,讲师,主要研究方向为数字图象处理、计算机视觉。

图 1 既表示本文的组织结构，又表示我们开发的 MPEG-4 兼容的人脸动画系统[18]和语音动画系统[19,20]的流程，图 1 最下面一行彩色图像表示每一步处理的输入数据和输出结果。

2. 人脸模型编辑器

建立精细的三维人脸模型是人脸建模与动画的第一步，同时也是至关重要的一步[21]。我们的人脸动画系统和语音动画系统使用了从 Internet 下载的三维人脸网格模型(如图 2(a)所示)，该模型存在着冗余特征(如马尾巴辫)，某些特征还需要进一步加工(如上下牙齿需要分开)。因此，有必要开发一个人脸模型编辑器，使用模型编辑器对该模型进行增加、删除、修改等编辑操作，从而得到满意的可用的人脸模型。



(a)修改前的人脸模型　　(b)修改后的人脸模型

图 2　修改前后的三维人脸模型

我们首先对下载的三维人脸模型(VRML 文件)进行手工筛选，仅保留特征名称、点和面的信息等。接着，以筛选过的 VRML 文件作为输入，通过平移、缩放、旋转等几何变换以适当的位置、大小和方向显示该模型，选择需要编辑的点和面，对其进行增加、删除、修改等编辑操作，直到得到满意的结果为止。最后，将修改后的三维人脸模型以特定文件格式保存，供人脸动画系统使用。图 3 为人脸模型编辑器的流程图。

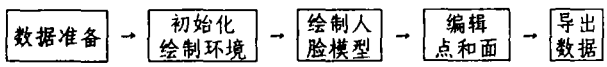


图 3　人脸模型编辑器的流程图

图 2(b)表示修改后的三维人脸模型，该模型被应用在我们已经实现的 MPEG-4 兼容的人脸动画系统和语音动画系统中，取得了良好的效果。

3. MPEG-4 人脸动画技术

MPEG-4 出现之前，在人脸建模与动画方面没有公认的国际标准。MPEG-4 是世界上第一个基于对象(object-based)的多媒体压缩标准[22~24]，该标准可以对各种自然的或合成的音频/可视对象(audio-visual objects，AVO)分别独立编码，将它们有机地集成在同一个场景中。考虑到人脸动画的重要性，MPEG-4 定义了专门的人脸动画工具。

3.1　中性状态人脸模型

图 4 表示 MPEG-4 定义的中性状态人脸模型。为了能够在任意人脸模型上定义人脸动画参数(facial animation parameter，FAP)，MPEG-4 定义了人脸动画参数单位(facial animation parameter units，FAPU)。FAPU 被定义为关键脸部特征之间距离的分数(fraction)，这些关键脸部特征之间距离(如两眼之间的距离)是在中性状态人脸模型上定义的(见图 4)。FAPU 使得人脸动画参数在任意人脸模型上具有一致的解释，能够在任意人脸模型上产生合理的表情与口型。

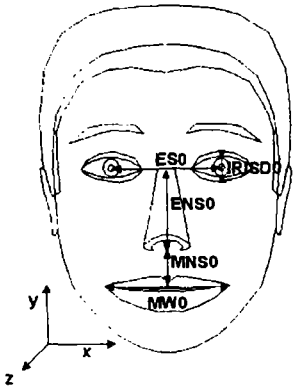


图 4　中性状态人脸模型

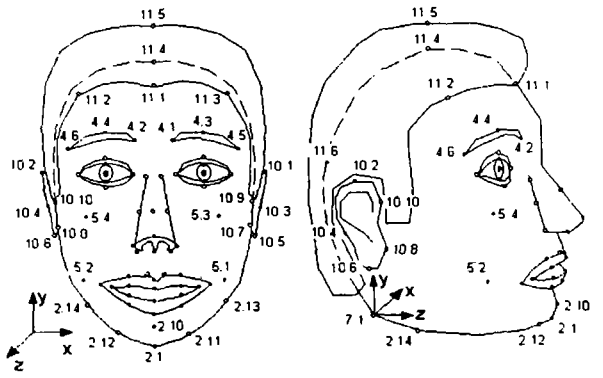


图 5　特征点

如图 5 所示，MPEG-4 在中性状态人脸模型上定义了 84 个特征点，这些特征点的主要作用是为定义人脸动画参数提供空间参考。有一些特征点(如 11.1、11.2、11.3 等)并不会受到人脸动画参数的影响，但是它们在校准私有人脸模型(proprietary face model)的时候需要用到(见 3.3 节)。这 84 个特征点被分为若干组(如脸颊、眼睛、嘴巴等)，所有符合 MPEG-4 的人脸模型都应当了解这些特征点的位置。

3.2　人脸动画参数

MPEG-4 共定义了 68 个人脸动画参数，这 68 个人脸动画参数被分为 10 组。组 1(FAP1 与 FAP2)定义了 2 个高层参数。FAP1 包括 14 个英文可视音素(visual phoneme，viseme)，用于生成语音动画。FAP2 包括喜悦(joy)、悲伤(sadness)、愤怒(anger)、恐惧(fear)、厌恶(disgust)、惊讶(surprise)等 6 种常用脸部表情。组 2-10(FAP3-68)定义了 66 个底层人脸动画参数。

3.3　人脸定义参数

每一个能够接受人脸动画参数并产生人脸动画的 MPEG-4 解码器应当具有符合 MPEG-4 标准的人脸模型，该模型通常是一个私有人脸模型，编码器并不了解该模型的构造和形状。使用人脸定义参数(facial definition parameter，FDP)，编码器可以对解码器的人脸动画进行不同程度的控制：

(1)不使用 FDP。解码器接收到的 FAP 流用于控制未经校准的私有人脸模型。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 5 所示全部或部分特

征点的位置，这些特征点用来校准解码器的私有人脸模型，然后使用 FAP 控制校准过的人脸模型。这样编码器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。这样可以利用纹理映射方法生成真实感人脸。

(4)使用三维人脸模型和人脸动画表(Facial Animation Table，FAT)。一个新的三维人脸模型从编码器传送到解码器，人部动画表用来定义 FAP 如何控制这个新模型。这时编码器对解码器的动画进行完全控制，能够全部预知动画的结果。

因为情况(4)只需要解码器做很少量的工作即可产生人脸动画，所以目前我们的人脸动画系统没有考虑这种情况，主要考虑前面三种情况。

MPEG-4 编码器可以给出中性状态人脸模型中全部或部分特征点的位置，这就要求解码器能够对私有人脸模型进行变形，使得私有人脸模型的特征点与编码器给出的特征点重合。该过程称为私有人脸模型校准(proprietary face model adaptation)。第 4 节专门介绍人脸模型校准。

除了特征点之外，MPEG-4 编码器还可以给出纹理图和特征点的二维纹理坐标，解码器可以利用纹理映射方法显著地提高人脸的真实感。第 5 节专门介绍生成真实感人脸。

需要指出的是：MPEG-4 关于人脸动画只是定义了一个标准格式，并没有给出具体问题的解决方案，这就给研究者留下了广阔的空间。以下各节将围绕着 MPEG-4 人脸动画技术，以实现 MPEG-4 兼容的人脸动画系统和语音动画系统为目标，分别介绍我们在人脸模型校准、真实感人脸、人脸动画、语音动画等各个方面遇到的问题和解决方案。

4. 人脸模型校准

人脸模型校准是人脸建模与动画中非常重要的一步，人脸模型校准实质上是一个空间变形(spatial deformation)问题，即：已知三维网格模型的所有网格点和 n+1 个控制点的位置，当控制点从老位置 p_i 移动到新位置 p_i'($0\leqslant i\leqslant n$，p_i' 可以等于 p_i，表示该控制点没有移动)时，如何求出网格点 p 的新位置 p'？也可以表达为：已知控制点的位移 $\Delta p_i=p_i'-p_i$($0\leqslant i\leqslant n$，Δp_i 可以等于 0，表示该控制点没有移动)，如何求出网格点 p 的位移 Δp?。

如前所述，MPEG-4 只是规定解码器应该能够进行人脸模型校准，并没有具体规定必须使用什么样的校准算法。目前比较常用的人脸模型校准算法主要有径向基函数变形(Radial Basis Function Deformation)[2]和 Dirichlet 自由变形(Dirichlet Free-Form Deformation，DFFD)[25]。

4.1 径向基函数变形

Pighin等采用散乱数据插值(scattered data interpolation)方法进行人脸模型校准，选择径向基函数作为插值函数[2]。径向基函数变形不具有局部性，也就是说，移动任何一个控制点都需要重新计算所有网格点的新位置。因此，当知道所有控制点的新位置时，可以采用径向基函数变形求所有网格点的新位置；当需要动态地、交互地调整控制点的位置并不断观察变形效果时，不适合采用径向基函数变形。

4.2 Dirichlet 自由变形

Dirichlet 自由变形算法的基本思想为：给定控制点集合 P 和其凸包内的一点 p，可以确定 p 的 Sibson 邻居集合 $P_n=\{p_i|0\leqslant i\leqslant n\}$，并计算出点 p 相对于 Sibson 邻居的 Sibson 局部坐标 $u_i(0\leqslant i\leqslant n)$。

当移动 P_n 中一个或多个控制点之后，假设控制点的新位置为 $p_i'=p_i+\Delta p_i(0\leqslant i\leqslant n)$，那么点 p 的新位置 $p'=p+\sum_{i=0}^{n}u_i\Delta p_i$。

使用 DFFD 算法对物体进行变形的基本步骤为：

(1)设计控制点集合。控制点可以在待变形物体的表面上，也可以在物体的内部或外部，但是物体需要变形的部分必须在控制点集合的凸包内。

(2)计算 Sibson 局部坐标。对物体上的每个点确定其 Sibson 邻居集合，并计算 Sibson 局部坐标。

(3)移动控制点。可以使用任意方法移动一个或一组控制点。

(4)对物体进行变形。根据控制点的新位置计算出物体上点的新位置，从而引起物体变形。

4.3 人脸模型校准的实现

我们采用 Dirichlet 自由变形来实现人脸模型校准[18]。首先根据 MPEG-4 关于人脸特征点的定义，在一般人脸模型上选择相应的网格点作为特征点，这些特征点就是 DFFD 算法的控制点。接着求控制点集合对应的 Voronoi 图和 Delaunay 图，对于每一个网格点，记录该网格点的 Sibson 邻居以及对应的 Sibson 局部坐标；对于每一个控制点，记录该控制点的影响区域，即受该控制点影响的网格点，将这些信息保存为新数据文件。

(a)正面照片　(b)侧面照片　(c)校准后的模型

图 6　用正面和侧面照片校准模型

我们在三维空间构造一个假想的包围一般人脸模型的最小长方体，该长方体称为包围体。如图 6 所示，特定人的正面照片表示 x-y 平面，侧面照片表示 z-y 平面，我们将人脸模型的包围体分别投影到正面和侧面照片上，得到正面和侧面照片上的红色长方形。分别在正面和侧面照片上调整红色长方形的位置和大小，使其包围正面和侧面人脸，这样就使得三维空间的包围体与正面和侧面照片上的红色长方形对齐并调整了包围体的大小，使得包围体的宽度、高度和厚度之间的比例符合特定人的比例。再将人脸模型的部分或全部特征点分别投影到正面和侧面照片上，在正面和侧面照片上动态地、交互地调整特征点的位置，使其与特定人的脸部特征相匹配。使用 DFFD 算法可以由控制点的移动计算出网格点的移动，这样一般人脸模型就变成了特定人脸模型。图 6(a)和(b)中的红色点表示右眼处的四个特征点(3.8、3.10、3.12、3.14)，图 6(c)表示校准后的人脸模型。

5. 真实感人脸

人脸模型校准可以让三维人脸网格模型在几何形状上接近特定人脸，但是这样绘制出来的人脸(如图 6(c))缺少表面细节，因而缺少真实感。我们在人脸模型校准的时候使用了特定人的正面和侧面照片来调整特征点，现在，我们还可以使用正面和侧面照片通过纹理映射(texture mapping)来生成真实感人脸。如果只使用正面照片，那么人脸的侧面纹理没有得到反映；反之，如果只使用侧面照片，那么人脸的正面纹理没有得到反映。因此，生成真实感人脸需要解决如何综合利用正面和侧面照片的问题。

Lee 等首先利用正面和侧面照片生成视点无关纹理图，再将该纹理图通过纹理映射贴到人脸模型上[25]。Lee 等使用的视点无关纹理图方法认为人脸的正面纹理完全取自正面照片，侧面纹理完全取自侧面照片。事实上，由于人脸的形状非常复杂，这种"非此即彼"的方法并不合适。例如，正面的某些部分(如鼻翼)主要朝向侧面，纹理信息应该主要取自侧面照片；还有一些区域接近 45 度，应该同时考虑正面和侧面照片。

我们可以将正面和侧面照片看作是人脸的正视图和侧视图，对于人脸模型上的某一点 P，假设其外法向量为(Nx，Ny，Nz)，在正面照片上对应点(正投影)的颜色为 C_1，在侧面照片上对应点(侧投影)的颜色为 C_2，那么该点的颜色为

$$C=k*C_1+(1-k)*C_2$$

其中，k=abs(Nz)/[abs(Nx)+abs(Nz)]，称为融合因子。

图 7　用纹理融合方法绘制的真实感人脸

在具体实现时，由于 OpenGL 提供了按照融合因子进行纹理映射的功能，因此我们可以用正面照片作为纹理图做一次纹理映射，设定好人脸网格模型各个顶点的融合因子之后，再用侧面照片作为纹理图做一次纹理映射。图 7 是我们用纹理融合方法绘制的真实感人脸。

6. 人脸动画

产生真实自然的人脸动画是一个人脸动画系统成败的关键。Lavagetto 等采用基于语义信息的方法来实现人脸动画[26]，需要同时给出人脸网格模型的几何信息和相关的语义信息，这种方法对人脸网格模型的依赖性比较大，更换人脸网格模型比较困难。

6.1　四层控制结构

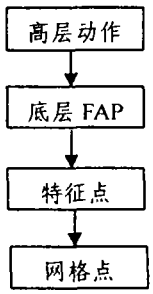


图 8　四层控制结构

我们提出了由高层动作、底层 FAP、特征点、网格点构成的人脸动画的四层控制结构(图 8)。

高层动作包括可视音素和表情，可视音素可以是 MPEG-4 的 FAP1 定义的 14 个英文可视音素，也可以是我们扩充的中文可视音素；表情可以是 MPEG-4 的 FAP2 定义的 6 种常用表情，也可以是我们扩充的其它表情。高层动作通过底层 FAP 的组合来实现。

底层 FAP 就是 MPEG-4 定义的 FAP3-68，一共 66 个。底层 FAP 通过一组相关的特征点的移动来实现。

特征点的位置按照 MPEG-4 的定义在中性状态人脸模型上确定，特征点的移动会驱动相关的网格点移动，从而产生人脸动画。

四层控制结构可以减少人脸动画对人脸网格模型的依赖性，使得更换模型比较方便。我们将高层动作抽象出来，使得使用和扩充都比较方便。

6.2　人脸动画的实现

下面我们自底向上地介绍人脸动画的实现：

首先确定特征点的相关网格点以及权因子。前面我们已经求出了特征点集合对应的 Voronoi 图和 Delaunay 图，记录了每一个网格点的 Sibson 邻居以及对应的 Sibson 局部坐标，也记录了每一个特征点的影响区域，并且根据这些信息采用了 Dirichlet 自由变形来实现人脸模型校准。由于特征点的影响区域完全由特征点的空间位置确定，没有考虑网格点之间的连接关系，因此直接采用 Dirichlet 自由变形来实现人脸动画时，特征点的影响区域偏大了一些，动画效果不佳。Noh 等提出的边距离度量方法考虑了网格点之间的连接关系，从特征点出发沿着网格边进行广度优先搜索，直到到达指定的边距离为止[14]。我们在实现人脸动画时将边距离度量方法结合进来，给特征点设定一个适当的边距离，从而缩小了特征点的影响区域，位于这个影响区域内的网格点就是特征点的相关网格点，网格点相对于该特征点的权因子就使用网格点相对

于该特征点的 Sibson 局部坐标。我们的方法综合考虑了特征点的空间位置和网格点之间的连接关系，能够得到比较好的动画效果。

接着我们就可以根据 MPEG-4 对底层 FAP(FAP3-68)的文字描述，手工移动相关特征点并观察效果，确定该底层 FAP 的相关特征点以及权因子。

最后，根据 MPEG-4 对高层动作(FAP1、FAP2)的文字描述，通过底层 FAP 的组合和适当的权因子，就可以实现高层动作。

每一帧动画最终都表现为若干特征点及其相关网格点的移动，这种移动是相对于中性状态人脸模型的位移。对于给定强度的某一高层动作，计算特征点以及网格点位移的方法为：

底层 FAP 的强度＝高层动作的强度 * 底层 FAP 的权因子

特征点位移＝底层 FAP 的强度 * 特征点的权因子

网格点位移＝特征点位移 * 网格点的权因子

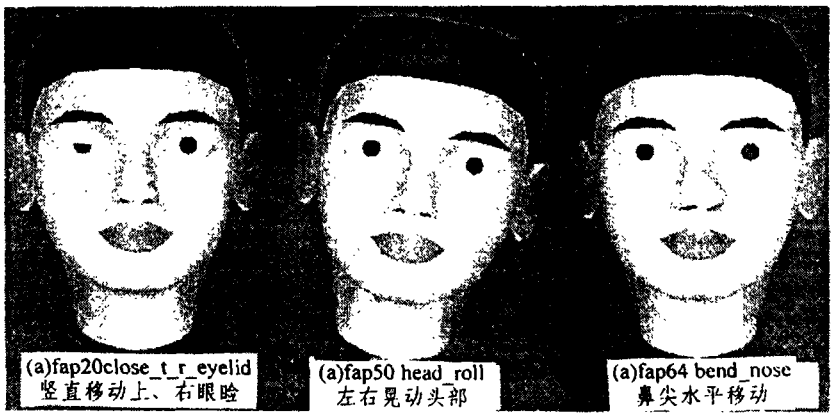


图 9　部分 FAP

我们采用上述方法实现了 MPEG-4 定义的 68 个 FAP，并且扩充了 15 个中文可视音素和多种其它表情。图 9 是我们实现的 FAP 中的一部分。

7. 语音动画

人脸和语音是人类相互交流的两种最重要的渠道，将人脸动画技术与语音处理(speech processing)技术相结合、用计算机生成语音与口型同步播出动画的技术，称为语音动画(speech animation)技术，也称为“说话的头像”(talking-head，talking-face)或“对口型”(lip-sync)[1]。

7.1　语音动画的分类

语音动画主要分为两种类型，即文本驱动(text-driven)的语音动画和语音驱动(speech-driven)的语音动画。

文本驱动的语音动画系统是语音合成(Text-To-Speech，TTS)系统与人脸动画系统相结合的产物：语音动画系统接受输入文本，调用语音合成系统进行分析处理，得到对应的音素流和时间信息，在语音合成系统产生语音输出的同时，人脸动画系统根据音素流和时间信息产生同步播出的人脸动画。文本驱动的语音动画系统也称为可视语音合成(Visual Text-To-Speech，VTTS)系统。

语音驱动的语音动画系统是语音分析(Speech Analysis)系统与人脸动画系统相结合的产物：语音动画系统接受用户的自然语音输入，调用语音分析系统进行分析处理，得到对应的音素流和时间信息，在播放自然语音的同时，人脸动画系统根据音素流和时间信息产生同步播出的人脸动画。本文研究文本驱动的语音动画。

我们在已经实现的一个 MPEG-4 兼容的人脸动画系统[18]的基础上，相继设计并实现了一个基于中国科大讯飞公司的 KD2000 的语音动画系统[19]和一个基于 Microsoft SAPI5.0[27]的语音动画系统[20]。

7.2　语音动画系统的结构

图 10 表示我们设计并实现的语音动画系统的结构。该系统接受带有表情标签的输入文本，经过 TTS 引擎分析处理，一方面进行语音合成、产生音频输出，另一方面产生带有时间戳的音素流和表情标签，经过时间计算，分别生成某一时刻的音素帧参数和表情帧参数，混合之后驱动特征点和网格点，产生与语音同步播出的人脸动画。其中的语音合成系统分别使用了中国科大讯飞公司的 KD2000 和 Microsoft 的 SAPI5.0，人脸动画系统是我们在 Windows 平台上使用 OpenGL 实现的[18]。

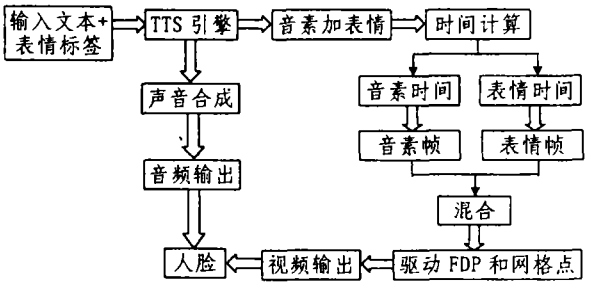


图 10　语音动画系统的结构

7.3　基于 KD2000 的语音动画系统

7.3.1　*中文可视音素*　MPEG-4 的高层人脸动画参数 FAP1 定义了 14 个英文可视音素，无法用于中文。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，定义了 15 个中文可视音素：a、bpm、dtnl、e、f、gkh、i、jqx、o、r、uv、zcs、zhchsh、n(韵尾)、ng(韵尾)。图 11 表示部分中文可视音素的口型。

7.3.2　*可视音素过渡*　语音动画系统中的 TTS 引擎分析处理输入文本，产生带有时间戳的音素流和表情标签。每个音素具有名称和持续时间两个参数，其中持续时间是该音素与下一个音素的间隔时间。利用查找表可以得到音素对应的可视音素。可视音素的强度随时间变化，所包含的底层 FAP 的强度也随时间变化。为了得到自然流畅的口型变化，需要对

两个相邻的可视音素进行平滑过渡，我们在底层 FAP 实现可视音素过渡。

图 11　部分中文可视音素的口型("a"、"zhchsh"、"gkh")

底层 FAP 在某一时刻的强度可以通过 Hermite 插值[28]得到。Hermite 插值一般需要四个参数，分别为起点和终点的位置和切向量。我们可以假设终点的切向量为 0，这样只需要三个参数就够了。假设可视音素的持续时间规范化为 1，那么插值公式为：

$$f(t)=(2t^3-3t^2+1)a_s+(-2t^3+3t^2)a+(t^3-2t^2+t)g_s$$

其中 $t\in[0,1]$，a_s 表示 FAP 在 $t=0$ 时的强度，a 表示 FAP 在 $t=1$ 时的强度，g_s 表示 FAP 在 $t=0$ 时的切向量。

7.3.3　*表情过渡*　我们的语音动画系统不但在高层动作定义了可视音素，还定义了表情。表情是通过在输入文本中插入表情标签来实现的，表情标签的格式为：

〈expression，a，T〉

其中表情名称 expression 可以是 MPEG-4 的 FAP2 定义的 6 种常用表情之一，也可以是我们扩充的其它表情，强度 a 是 0～63 之间的整数，时间 T 是达到强度 a 所需的时间。我们假设经过时间 T 之后，强度 a 会一直保持着，直到通过另一个表情标签指定另一个强度为止。强度为 0 的表情标签可以用来复位已经设置的表情。表情的开始时间由表情标签在输入文本中的位置来确定。

与可视音素类似，表情也是通过底层 FAP 的组合来实现的，因此表情过渡也在底层 FAP 用 Hermite 插值来实现。

当前表情所包含的底层 FAP 在开始时间($t=0$)的强度 a_s 依赖于前一个表情对该 FAP 的影响，有如下三种可能情况：

(1)0　如果当前表情是第一个影响该 FAP 的表情

(2)b　如果前一个表情的强度为 b，并且两个表情之间的时间间隔大于前一个表情的时间 T

(3)前一个表情实际达到的强度　如果两个表情之间的时间间隔小于前一个表情的时间 T

当前表情所包含的底层 FAP 在开始时间的切向量在前两种情况下都为 0，在第三种情况下为前一个表情在该时刻的切向量，即

$$f'(t)=(6t^2-6t)(a_s-a)+(3t^2-4t+1)g_s$$

其中 a_s、a 和 g_s 是前一个表情的 Hermite 插值参数。

7.3.4　*可视音素与表情混合*　上述可视音素与表情过渡算法分别生成某一时刻的音素帧与表情帧参数，需要对它们进行混合，才能驱动特征点和网格点。如果一个底层 FAP 只包含在音素帧或表情帧参数中，那么直接取该帧参数即可；如果一个底层 FAP 同时包含在音素帧与表情帧参数中，那么我们取其最大值。这种混合方法计算简单，适合生成实时语音动画，实践表明具有良好的效果。

7.3.5　*语音与动画同步*　我们的语音动画系统采用时间戳同步的方法，TTS 引擎分析处理好输入文本之后，每次开始播放一句话的时候都要给人脸动画系统发送"开始"消息。人脸动画系统接收到"开始"消息，就根据带有时间戳的音素流和表情标签进行可视音素过渡、表情过渡以及可视音素与表情混合，得到底层 FAP 的强度，驱动特征点和网格点产生与语音同步的人脸动画。由于我们只在一句话的开始才发送一次消息，因此系统通信开销很小。

7.4　基于 SAPI5.0 的语音动画系统

7.4.1　*定义中文可视音素*　SAPI5.0 定义了 408 个中文音素，每一个中文音素表示一种汉语拼音组合，但是没有定义中文可视音素。因为中文音素的数目比较多，并且一个中文音素(即一种汉语拼音组合)往往对应着一系列变化的口型，所以直接用 408 个中文音素作为可视音素来设计口型库是不现实的。更为严重的是：我们经过实验发现 SAPI5.0 不产生 SPEI_TTS_PHONEME 和 SPEI_TTS_VISEME 事件，这就意味着我们根本得不到中文音素，更不用说可视音素。因此，我们仍然使用基于 KD2000 语音动画系统中定义的 15 个中文可视音素。

7.4.2　*得到中文可视音素*　因为 SAPI5.0 不产生 SPEI_TTS_PHONEME 和 SPEI_TTS_VISEME 事件，所以我们必须自己分析输入文本才能得到中文可视音素。

SPEI_WORD_BOUNDARY 事件表示一个汉字的开始，我们在处理该事件时，使用 GetStatus 函数获得该汉字在输入文本中的位置，便可以取得该汉字；接着通过查找我们自己建立的汉语拼音表，可以得到该汉字的汉语拼音；再用我们自己定义的 AnalysisPhoneme 函数对该汉字的汉语拼音进行分析，将其分解为中文音素序列，从而得到对应的中文可视音素。

7.4.3　*估算可视音素的持续时间*　要进行可视音素过渡，除了可视音素之外，还必须知道可视音素的持续时间。在 SAPI5.0 中，可以通过 SetRate 函数设定播放速度，速度值范围在－10 到 10 之间，数值越大，播放速度越快。对于每一种播放速度，我们都进行长句子测试，以所有字的平均时间作为一个汉字在此播放速度下的持续时间的估算值。这样估算出的持续时间是一个平均值、经验值，而在实际播放时，每一个汉字的持续时间是不相同的，与估算值肯定有一个误差，我们

将在7.4.5节讨论这个问题。

我们已经估算出一个汉字的持续时间，并且将该汉字分解为可视音素序列，接下来我们就可以估算可视音素的持续时间。我们根据中文的发音特点，对声母、介母、韵母进行分类，根据可视音素位置和类型分配不同的持续时间。例如，对“a”、“o”等需要较大口型的韵母分配较长的持续时间，对“b”、“p”、“m”等需要较大口型的声母也分配较长的持续时间，对“zhuang”中间的介母“u”分配较短的持续时间，对“d”、“t”、“n”、“l”等主要靠舌头发音的声母也分配较短的持续时间。

7.4.4 *处理表情标签* 在调用SAPI5.0播放语音之前，我们要从输入文本中将表情标签提取出来，并记住表情标签在输入文本中的位置。SAPI5.0播放的是不带有表情标签的文本。我们根据表情标签在输入文本中的位置，在SAPI5.0播放到此处的时候，进行表情过渡并与可视音素混合。

7.4.5 *语音与动画同步* 一般情况下，我们估算的一个汉字的持续时间与实际播放语音的时间不会完全相同，有的时候差别可能还比较大。为了使得语音与动画同步，我们将语音动画系统分为语音与动画两个线程，通过共享音素缓冲区来实现同步。

语音线程首先启动SAPI5.0在异步方式下朗读输入文本，接着不断分析输入文本中的汉字，将其分解为音素序列，写入共享的音素缓冲区，同时清除缓冲区内原有的内容。动画线程不断读取音素缓冲区内的音素来生成动画。

如果估算的持续时间大于实际播放语音的时间，那么语音播放结束了，动画播放还没有结束。这时通过更新音素缓冲区，可以停止原有汉字的动画播放，从该时间点开始向下一个汉字进行动画过渡，从而使得动画能够赶得上语音。这样做的结果是最后一个或几个音素的动画无法得到完整的播放。

如果估算的持续时间小于实际播放语音的时间，那么动画播放结束了，语音播放还没有结束。动画线程会保持当前口型不变，等到语音播放结束、更新音素缓冲区之后，才会向新的口型过渡。如果等待时间太长(例如遇到句子之间间隔时)，人脸动画会显得很不自然。因此我们设定了一个阈值，当等待时间大于该阈值时，人脸就会恢复到中性状态。

因为我们无法精确地得到一个汉字的持续时间，只能进行估算，所以我们采用了这种一个汉字同步一次的方法。实验结果表明，生成的语音动画效果还是不错的。

结束语 本文介绍了我们对人脸建模与动画、特别是基于三维模型的人脸动画的研究工作，主要内容包括：

(1)开发了一个人脸模型编辑器。利用人脸模型编辑器对从Internet下载的三维人脸网格模型进行增加、删除、修改等编辑操作，从而得到满意的可用的人脸模型。

(2)实现了人脸模型校准。利用特定人的正面和侧面照片，采用Dirichlet自由变形方法实现了从一般人脸模型到特定人脸模型的变形。

(3)生成了真实感人脸。利用特定人的正面和侧面照片，采用纹理融合方法来生成真实感人脸。

(4)开发了一个MPEG-4兼容的人脸动画系统。提出了人脸动画的四层控制结构，采用Dirichlet自由变形与边距离相结合的方法实现了MPEG-4定义的68个人脸动画参数。

(5)开发了基于KD2000和SAPI5.0的语音动画系统。提出了一种语音动画系统的结构，实现了基于中国科大讯飞公司的KD2000和Microsoft SAPI5.0的语音动画系统。

进一步的工作可以沿着以下两个方向：

(1)基于网络的人脸和语音相结合的多通道用户界面系统。在基于网络的人-机交互方面，使用人脸和语音相结合的多通道用户界面可以给人们网上购物、检索信息、浏览信息提供方便。在基于网络的人-人交互方面，使用人脸和语音相结合的多通道用户界面可以创造一个类似于真实世界的虚拟环境，人们在这个虚拟环境中自由交流、协同工作，支持远程会议、网络虚拟社区等网络应用模式。

(2)基于样本的语种无关的语音动画技术。分为学习与合成两个阶段：在学习阶段，给定一段说话的录像，通过学习可以生成说话者的语音动画模型，该模型极大限度地保留了说话者的语音与口型、表情之间的联系，反映了说话者特定的语言习惯；在合成阶段，直接输入一段新的语音，或者输入文本、由语音合成系统生成一段新的语音，都可以驱动语音动画模型生成高度真实感语音动画。

参考文献

1 Parke F I, Waters K. Computer Facial Animation, A K Peters, Wellesley, Massachusetts, 1996

2 Pighin F, Hecker J, Lischinski D, et al. Synthesizing realistic facial expressions from photographs. In: Proc. SIGGRAPH'98, 1998. 75～84

3 Bregler C, Covell M, Slaney M. Video rewrite: driving visual speech with audio. In: Proc. SIGGRAPH'97, 1997. 353～360

4 Brand M, Voice Puppetry. In: Proc. SIGGRAPH'99, 1999. 21～28

5 Cosatto E, Graf H P. Photo-Realistic Talking-Heads from Image Samples. IEEE Transactions on Multimedia, 2(3), Sept. 2000. 152～163

6 Parke F I. Computer generated animation of faces. In: Proc. ACM annual conf. Vol. 1, Aug. 1972. 451～457

7 Parke F I. A paramateric model for human faces: [Ph. D. Thesis]. University of Utah, Salt Lake City, UT, Dec. 1974, UTEC-CSc-75～047

8 Parke F I. Parameterized models for facial animation. IEEE Computer Graphics and Applications, 1982, 2(9): 61～68

9 Platt S M, Badler N I. Animating facial expressions. Computer Graphics (Proc. SIGGRAPH'81), 1981, 15(3): 245～252

10 Waters K. A muscle model for animating three-dimensional facial expression, Computer Graphics (Proc. SIGGRAPH'87), 1987, 21(4): 17～24

11 Lee Y C, Terzopoulos D, Waters K. Realistic modeling for facial animation. In: Proc. SIGGRAPH'95, 1995. 55～62

12 Magnenat-Thalmann N, Primeau N E, Thalmann D. Abstract muscle actions procedures for human face animation. Visual Computer, 1988, 3(5): 290～297

13 Kalra P, Mangili A, Magnenat-Thalmann N, Thalmann D. Simulation of Facial Muscle Actors Based on Rational Free-Form Deformation. Computer Graphics Forum (Proc. EUROGRAPH'92), 1992, 2(3): 59～69

14 Noh J-y, Fidaleo D, Neumann U. Animated Deformations with Radial Basis Functions. ACM Virtual Reality and Software Technology (VRST'1999), 1999. 166～174

15 Williams L. Performance-driven facial animation. Computer Graphics (Proc. SIGGRAPH'90), 1990, 24(4): 235～242

16 Guenter B, Grimm C, Wood D, et al. Making Faces. In: Proc. SIGGRAPH'98, 1998. 55～66

17 Breton G, Bouville C, Pelé D. FaceEngine A 3D Facial Animation Engine for Real Time Applications. In: Proc. 3D Technologies for the World Wide Web (WEB3D'2001), Paderbon, Germany, 2001. 15～22

18 王奎武，王洵，董兰芳，陈意云. 一个MPEG-4兼容的人脸动画系统. 计算机研究与发展，2001，38(5)：529～535

19 王洵，张道义，董兰芳，陈国良. 一个MPEG-4兼容的语音动画系统. 系统仿真学报，已录用

(下转第64页)

$Gr_1 \subseteq Gr_2$。令对于任意 n 元组来说，几何表示都是完备的，n 元组是指这个类中几何个体的布局。几何个体定义域被分为布局的等价类，布局的等价类可以通过定义了特定几何结构的转换组相互转换。

对半代数集合和它们在 R^n 中的位置的抽象的几何结构表示是基于符号结构的，符号结构显式地表示了几何图形的不变性质：符号结构作为不变量显式地刻画转换组集合的等价类。这种用于几何配置的转换不改变它的符号表示。在图4中的两个图形结构表示两个不同的空间个体布局的几何结构。空间个体是通过平面上的点的子集来建模的。

α、β 和 γ 表示二维正则开集，A、B、C 和 E 表示一维开集，a、b 和 c 是平面上的点。α、β、γ、A、B、C、E、a、b、c 是显式地指定平面上的点和点集的符号。空间布局通过无向图来显式地表示：边由一维开集的名称来标注；节点由点的名称来标注，这里，点或者是一维开集的交集，或者是一维开集的起点，或者是一维开集的终点；面由通过边围绕的二维开集的名称来标注。

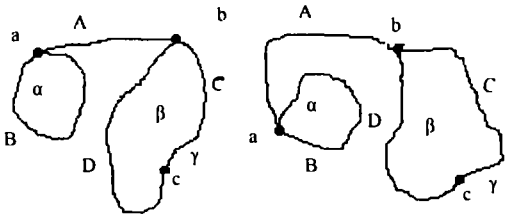


图4　两个等价的 PLA 结构[9]

这种图结构是一个显式地表示了在各向同性转换下保持不变的几何性质的符号结构，各向同性转换是一种平面的拓扑转换。图4中的两个布局是由同一个图来表示的。两个布局可以通过平面拓扑转换方法进行相互转换。对于各向同性转换来说，图表示是不完备的[9]，即布局虽然可以由同一图结构表示，但是不能由各向同性转换方法进行相互转换。PLA 结构是扩展的图表示，它既显式地表示了每个点的相邻边和相邻面的循环顺序（对于图4中的 a 来说：相邻边和相邻面的循环顺序是[α、B、γ、A、γ、B]），又显式地表示了无边界区域的名称（图4中的 γ）。对于各向同性的几何结构来说，PLA 结构是完备的[10]，它们可以完全并且显式地表示正则闭集合的各向同性。它们比半代数集合表示更加抽象，因为它们表示一个更大的集合类，并且辨别较少的性质。类似 PLA 的结构可以被广泛地使用。

位置定义域的抽象　位置定义域的抽象方法是选择一个基准框架并且定义位置，以使确定的个体集合被保存在更为一般的几何结构位置定义域中。位置定义域中的临界抽象点必须选择“合适”的基准框架，合适程度是根据有限个体集合的几何结构的表示来评价的。假设个体是可以辨别的，如果位置定义域可以充分地辨别所有选择的个体，并且保存了个体集合的预期的几何结构，则基准框架“适合于”表示个体集合的几何结构。

在图1中，在辨别 I_2 和 I_3 之间的区别时，位置定义域是不充分的。忽略 I_3，则位置定义域保存了布局的拓扑属性在集合 C 中：$C=\{(I_1,I_4),(I_4,I_2),(I_1,I_2)\}$。根据基准框架的结构、位置的定义和它们的位置知识，可以导出集合 C 中的布局的不相交关系。基准框架在保存平移-不变的几何性质时是充分的。根据位置知识和基准框架的嵌入知识，即经度和纬度 g_1、g_2 和 g_3、g_4 轴的排列，可以导出布局的集合 C。

考虑平面内的位置，一般来说，基准框架内的位置抽象是由二个实数轴的系统给出的，基准框架内的位置是由有限的平面区域的划分构成的。通常，区域划分可以是不规则的，也就是说，划分元素可以任意地共享区域。一个特例是由图1中的9个区域构成规则划分。通过区域划分给出的表示基准框架内的区域个体位置。

参 考 文 献

1 刘亚彬，刘大有. 空间推理与地理信息系统综述. 软件学报，2000，11(12)：1598～1606

2 刘亚彬. 空间推理的逻辑基础及空间信息的模糊性和不确定研究：[吉林大学博士学位论文]. 2001

3 Bittner T, Frank A U. On Representing Geometries of Geographic Space. In: Proc. of 8th Int. Symposium on Spatial Data Handling, SDH'98, (Poiker, T. K., & Chrisman, N., eds.), in Vancouver, Canada July 1998

4 Bittner T, Frank A U. On Representing Geometries of Geographic Space. In: Proc. of 8th Int. Symposium on Spatial Data Handling, SDH'98, (Poiker, T. K., & Chrisman, N., eds.), in Vancouver, Canada, Published by International Geographic Union, 1998. 111～122

5 Frank. Qualitative Spatial Reasoning about Distance and Direction in Geographic Space. Journal of Visual Languages and Computing, 1992, 3: 343～371

6 Kanellakis P C, Kuper G M, et al. Constraint Query Languages. 9th ACM PODS

7 Franklin W R. Cartographic Errors Symptomatic of Underlying Algebra Problems First International Symposium on Spatial Data Handling, Zurich, Switzerland

8 Hironaka H. Triangulation of Algebraic Sets. Proceedings of Symposia in Pure Mathematics 29: 165～186

9 Vandeurzen L, Gyssens M, et al. On the Desirability and Limitation of Linear Spatial Database Models. In: 4th Symposium on Large Spatial Database-SSD95, Portland, ME, Springer Verlag, Heidelberg, 1995

10 Hidders J. An Isotropic Invariant for Planar Drawings of Connected Planar Graphs, Eindhoven University of Technology. 1995

11 Leung Y. Intelligent Spatial Decision Support Systems. ISBN 3-540-62518-6 Springer-Verlin, Berlin Heidelberg, New York. 1997

（上接第11页）

20 王洵，张道义，董兰芳，陈国良. 一个基于 SAPI5.0的中文语音动画系统. 已投稿

21 王洵，宋阳，董兰芳，陈国良. 人脸模型编辑器的设计与实现，小型微型计算机系统，已录用

22 ISO/IEC IS 14496-1 Systems, 1999

23 ISO/IEC IS 14496-2 Visual, 1999

24 王洵，董兰芳，陈国良，许胤龙. MPEG-4人脸动画技术和一个基于MPEG-4的人脸动画系统的设计. 计算机科学，2002，29(1)：49～52

25 Lee W, Magnenat-Thalmann N. Head modeling from pictures and morphing in 3D with image metamorphosis based on triangulation. In: Proc. CAPTECH'98 (Modeling and Motion Capture Techniques for Virtual Environments), Springer LNAI LNCS Press, Geneva, 1998. 254～267

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

27 Microsoft Speech Technologies Web Site. http://www.microsoft.com/speech

28 Foley J D, van Dam A, Feiner S K, Hughes J F. Computer Graphics: Principles and Practice, Second Edition in C. Addison-Wesley, 1996