基于机器学习的语音驱动人脸动画方法*

陈益强[1+], 高 文[1,2], 王兆其[1], 姜大龙[1]

[1](中国科学院 计算技术研究所,北京 100080)

[2](哈尔滨工业大学 计算机科学与工程系, 黑龙江 哈尔滨 150001)

A Speech Driven Face Animation System Based on Machine Learning

CHEN Yi-Qiang[1+], GAO Wen[1,2], WANG Zhao-Qi[1], JIANG Da-Long[1]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[2](Department of Computer Science and Engineering, Harbin Institute of Technology, Harbin 150001, China)

+ Corresponding author: Phn: 86-10-82649008, Fax: 86-10-82649298, E-mail: yqchen@ict.ac.cn

http://www.jdl.ac.cn

Received 2001-06-04; Accepted 2001-08-01

Chen YQ, Gao W, Wang ZQ, Jiang DL. A speech driven face animation system based on machine learning. *Journal of Software*, 2003,14(2):215~221.

Abstract: Lip synchronization is the key issue in speech driven face animation system. In this paper, some clustering and machine learning methods are combined together to estimate face animation parameters from audio sequences and then apply the learning results to MPEG-4 based speech driven face animation system. Based on a large recorded audio-visual database, an unsupervised cluster algorithm is proposed to obtain basic face animation parameter patterns that can describe face motion characteristic. An Artificial Neural Network (ANN) is trained to map the cepstral coefficients of an individual's natural speech to face animation parameter patterns directly. It avoids the potential limitation of speech recognition. And the output can be used to drive the articulation of the synthetic face straightforward. Two approaches for evaluation test are also proposed: quantitative evaluation and qualitative evaluation. The performance of this system shows that the proposed learning algorithm is suitable, which greatly improves the realism of face animation during speech. And this MPEG-4 based learning are suitable for driving many different kinds of animation ranging from video-realistic image wraps to 3D Cartoon characters.

Key words: machine learning; facial animation; speech driven

摘 要: 语音与唇动面部表情的同步是人脸动画的难点之一.综合利用聚类和机器学习的方法学习语音信号和唇动面部表情之间的同步关系,并应用于基于 MEPG-4 标准的语音驱动人脸动画系统中.在大规模音视频同步数据库的基础上,利用无监督聚类发现了能有效表征人脸运动的基本模式,采用神经网络学习训练,实现了从含韵律的语音特征到人脸运动基本模式的直接映射,不仅回避了语音识别鲁棒性不高的缺陷,同时学习的结果还可以直接驱动人脸网格.最后给出对语音驱动人脸动画系统定量和定性的两种分析评价方法.实验结果表明,

* Supported by the National Natural Science Foundation of China under Grant No.60103007 (国家自然科学基金); the National High-Tech Research and Development Plan of China under Grant No.2001AA114160 (国家高技术研究发展计划)

第一作者简介: 陈益强(1973－),男,湖南湘潭人,博士生,主要研究领域为数据挖掘及其应用,智能人机交互,生物信息学.

基于机器学习的语音驱动人脸动画不仅能有效地解决语音视频同步的难题,增强动画的真实感和逼真性,同时基于 MPEG-4 的学习结果独立于人脸模型,还可用来驱动各种不同的人脸模型,包括真实视频、2D 卡通人物以及 3 维虚拟人脸.

关键词: 机器学习;人脸动画;语音驱动

中图法分类号: TP181 **文献标识码**: A

语音与唇动面部表情是人际之间两种重要的交互模式.二者之间存在着非常密切的同步关系.在人脸动画中,如何实现二者之间的同步,并应用于语音驱动人脸动画,一直是多模式人机交互的研究热点和难点,引起了国内外众多研究者的兴趣.

目前的研究方法一般可分为通过语音识别和不通过语音识别两种.前者建立在一种中间基本单位形式的表示如音素(phoneme)、视觉基元(viseme)以及更进一步音节(syllable)的基础上,通过语音识别得到这种基本单位后,按照一套规则驱动与其基本单位对应的唇形[1~4].此方法直接、有效,但规则通常由认知专家人工定义,定性并不精确,同时实现准确的非特定人语音识别存在较大的困难.后者不通过语音识别,而是直接将语音特征映射到人脸动画参数上.这种方法不仅可以回避语音识别遇到的问题,同时又能有效地实现同步,增加真实感和逼真度.Hani Yehia[5]提出训练非线性预测器完成从语音到人脸运动的学习,输入为 LSP 相关系数,输出为人脸上安放的位置跟踪器的相对坐标值.Massaro 和 Beskow[6]提出一种采用人工神经网络的视觉语音合成方案.通过先验知识标定语音段与人脸运动参数的关系,得到了较好的结果.但人工标注工作量大且精度不高.Matthew brand[7]介绍了一种基于双隐马尔可夫(HMM)的相关控制方法,并用于语音驱动人脸动画,首先利用计算机视觉技术获取真实人脸运动行为,然后通过学习获得人脸控制模型.这种方法虽然很好地解决了同步以及上部人脸动画,但却不能达到实时,同时,HMM 中状态节点个数、每个状态节点下的高斯混合项个数等参数的选取存在经验性,对这些参数的选取直接影响了动画的效果.虽然以上技术取得了较好的效果,但要实现逼真的实时的语音驱动人脸动画,还存在很多困难.

本文综合利用聚类和机器学习的方法学习语音信号和唇动人脸表情之间的精确同步关系,并应用于基于 MEPG4 标准的语音驱动人脸动画系统中,以提高系统的自然度和逼真性.与已有的方法相比,本文提出的方法有如下特点:(1) 通过聚类,我们可以发现有效表征人脸运动的基本模式,这些基本模式不仅包含唇动,而且还包含脸部表情,突破以往系统仅能实现唇动未含表情的缺陷.同时,聚类是在语句级上完成,相对于手工调整唇形或在音素级上聚类,该方法不仅可以得到更多的用于刻画人脸运动的模式,而且也利于后续学习机的训练收敛.(2) 通过神经网络,利用从真实视频中获取的运动数据训练语音驱动人脸动画模型,可以提高系统的真实感和逼真性,同时避免了采用人工规则或手工标注的缺陷.(3) 直接采用加入韵律的语音特征到人脸基本模式映射的方法,可以实现非特定人语音输入甚至非符号语音输入,回避了语音识别.(4) 基于 MPEG-4 的学习结果独立于后端的人脸模型,因此可以用来驱动各种不同的人脸模型,包括真实视频、2D 卡通人物以及 3 维虚拟人脸.

本文第 1 节介绍基本方法.第 2 节给出数据预处理方法.第 3 节给出聚类算法以及神经网络模型.第 4 节是实验结果.最后给出结论和将来的工作.

1 基本方法

本文采用不通过语音识别的方法,即不考虑获取离散的且与音素或词语相对应的人脸模式,而是直接将语音特征映射到 MPEG-4 定义的人脸运动参数(FAP)模式中.

1.1 人脸动画参数模式

MPEG-4 定义的人脸运动参数 FAP(facial animation parameter)是实现人脸动画的一组参数[8].FAP 建立在人脸细微运动的基础上,通过对人脸模型各个部位动作的详细描述和量化,可以再现绝大多数自然的人脸表情和唇动.MPEG-4 标准中共定义了 68 个 FAP,其中包含有唇形(viseme)FAP 和表情(expression)FAP.对于这两种 FAP 来说,可以预先存储好一些基本的、不同的唇形或表情数据,其他唇形或表情可以由这些基本的唇形或表情线

性组合而成.唇形和表情 FAP 的作用是准确、方便地表现简单的唇动和表情,但是对于复杂、不规则的唇动和表情,唇形和表情 FAP 则很难较好地描述,因此在 MPEG-4 标准中,除唇形和表情 FAP 外,为方便用户刻画更加细微的人脸运动,还给出一般 FAP 定义.一般 FAP 主要用于人脸某一特定区域的运动,如眉毛上扬,或上嘴唇往上运动等.由于一般 FAP 包含了人脸各部分细节的运动描述,这样可生成比标准定义的 14 种唇形和表情 FAP 更完善和复杂的动画,因此本文的 FAP 模式建立在一般 FAP 的基础上.同时,与 MPEG-4 定义的基本唇形和表情不同,本文所描述的 FAP 模式不是参照假设的语音感知分类,如音素或词语进行聚类分析[9],而是直接从大量的真实图像数据中进行聚类分析得到,这样得到的结果可更加有效地用于人脸动画中.我们假设所有可能的人脸姿势都属于高维空间平面的某一种状态,而复杂人脸运动可视为成千上万种状态之间的转移.

1.2 同步学习策略

针对这种多样复杂的人脸运动轨迹,我们的学习策略是用一些可有效表征人脸运动特点或代表某一状态集合的 FAP 模式来分段近似这种多样化,然后用线性插值粘合这些分段,从而实现逼真度较高的人脸动画.当然,要实现语音驱动的人脸动画,这些 FAP 模式序列的获取必须通过某种预测方法得到.同时由于人脸运动的复杂性,只有通过从大规模音视频库中学习得来的结果,才能实现更为自然的人脸动画,采用手工或规则都无法达到要求.由于神经网络在处理输入输出映射关系时有较好的特性,而且符合人脑结构的特点,我们采用神经网络学习这种映射关系.

1.3 系统框架

图 1 简要给出了语音驱动唇动人脸动画系统整个过程的处理框架.从大规模音视频同步库中抽取出语音和图像序列.利用计算机视觉中的跟踪技术以及图像处理中的检测技术,可以将与 MPEG-4 定义的人脸定义参数(FDP)以及人脸动画参数(FAP)相对应的特征点从视频序列中提取出来.根据特征点坐标数据可以计算出 FAP 数据,利用无监督聚类算法就可以获取典型的 FAP 模式.同时,对同步录制的语音数据进行分析,利用语音识别中采用的语音特征提取技术,可以得到 LPC 系数(线性预测系数)以及 RASTA-PLP 和一些韵律参数:如能量、基频等,从而获取与 FAP 模式一一对应的语音特征序列.最后在帧层次上训练神经网络,学习从含上下文及韵律的语音特征到 FAP 模式上的映射,利用学习到的这种映射关系,就可以实现语音与人脸动画的有效同步.

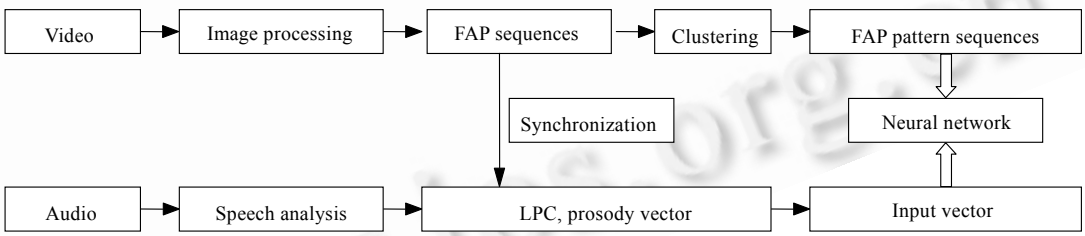


Fig.1 Synchronization processing framework of speech driven face animation

图 1 语音驱动人脸动画同步处理框架

2 数据预处理

2.1 语音信号分析

提取语音特征有多种方法,如 MFCC,LPC,PLP 等,但这些语音特征是为语音识别中音素分析设计的,对人脸行为预测并不完全确切有效.语音与人脸行为之间的关系可能比语音和音素之间的关联更加复杂而且松散.为了获取有效的语音特征表示,我们计算了混合的 LPC 以及 RASTA-PLP 语音特征.这两个语音特征已成功应用于语音识别中,对于环境以及话者变化有较强的鲁棒性[9].同时,由于语音中蕴涵的一些韵律特征与人脸表情也有较强的关联,如语音能量与表情夸张度之间、基频与头势之间等,因此在实验中,还提取了语音段的基音频率、能量等韵律特征,将这些为语音识别设计的语音特征与一些韵律特征结合起来形成语音分析得到的特征向量.

2.2 视频信号分析

为了获得人脸运动数据,我们开发了一套计算机视觉系统,可以同步跟踪许多个性化的人脸特征,如嘴角、嘴唇线、眼睛和鼻尖等.图 2 显示了系统的跟踪结果.但现行的系统无法跟踪低纹理区,如面颊.对系统而言,获取精确的特征点运动数据比实验跟踪算法更重要,我们回避计算机视觉中人脸特征点精确定位这一难点,采用在人脸的低纹理区上标记特定颜色点的做法来获取 MPEG-4 中定义的 FDP 与 FAP 相对应的人脸特征点的准确运动数据,利用跟踪技术以及图像处理技术,每一帧图像的人脸特征点都能被准确定位和提取出坐标,通过坐标换算以及比例变换可以得出一组 FAP 值,从而形成视觉特征向量.具体的 FAP 定义和计算可参考文献[8].图 3 显示了系统的跟踪结果以及每个特征点的影响区域.

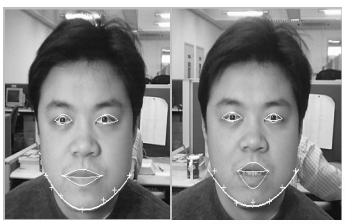


Fig.2 Face feature points tracking

图 2 人脸特征点跟踪

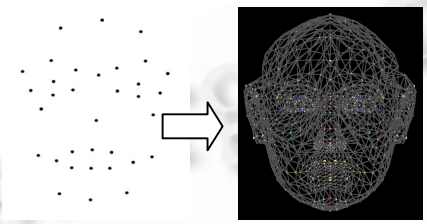


Fig.3 Feature points detection and effected area

图 3 特征点检测与影响区域

3 聚类和学习

3.1 聚类算法

对于基本人脸模式,认知学家给出了一些研究成果,但一般都是定性地给出 6 种基本表情或更多,也有研究人员通过对真实数据聚类来发现模式,但目前大多聚类分析都是在音素基础上进行的[10,11],忽略了语句级人脸运动的动态性.我们希望从大量真实语句中发现一组有效表达人脸运动的模式,这种发现的模式可以具有很明显的意义,如 MPEG-4 定义的 14 种唇形,也可以只是一种可有效用于人脸合成的基本模式.通过模式我们发现,不仅利于神经网络训练的收敛,同时也为后续对唇动人脸运动复杂过程的解释和理解打下了基础.在聚类过程中,由于这样的基本模式的个数并不确定,一般采用无监督聚类.通过数据预处理得到具有统一长度而且较为精确的 FAP 模式序列后,采用改进的 ISODATA(iterative self-organizing data,迭代自组织数据分析方法)算法[12]进行聚类分析.

对于聚类算法,存在很多参数的设置问题,参数设置对于聚类结果影响很大,对于唇动人脸基本模式聚类,由于没有已知类别的实验样本集作为错误率评价,同时又无法直接观察高维空间的几何特征,因此评价聚类结果存在一定的困难.我们直接采用聚类数据与真实数据求方差的做法来衡量聚类结果是否已达到描述主要运动模式的要求.通过调整聚类算法参数,如希望聚类数目、最大训练次数、每类最小样本数、分离参数 P 以及合并参数 C 等,可以得到不同的聚类结果,对这些结果按式(1)都进行方差计算,结果见表 1.

$$\text{ErrorSquare}(X,Y)=\sqrt{(X-Y)*(X-Y)^T}\Big/\|X\|, \tag{1}$$

其中 X 为真实数据矩阵,Y 为真实数据向类别映射后的矩阵,$\|X\|$表示矩阵大小.

Table 1 Comparison of clustering result

表 1 聚类结果比较

	The minimal number of objects in one class	Partition parameter P/ Combination parameter C	Number of expected classes	Errors square
1	32	P=[0.5,1], C=[1,1.5]	18	3.559 787
2	20	P=[0.5,1], C=[1,1.5]	21	4.813 459
3	10	P=[0.5,1], C=[1,1.5]	23	2.947 106
4	5	P=[0.5,1], C=[1,1.5]	29	2.916 784
5	3	P=[0.5,1], C=[1,1.5]	33	2.997 993

上述聚类是在 6 200 个样本数据上进行的,希望聚类的数目设为 64,最大训练次数设为 200,其余参数人工调节,*P* 表示分离参数,*C* 表示合并参数,*P* 在[0.5,1]区间变化,*C* 在[1,1.5]区间内变化.我们发现方差比较并没有呈平缓的下降,而出现某种抖动,这主要由于不同聚类参数选取如初始类中心选择以及聚类算法的删除步骤对结果产生的影响.从方差估计可以看出,第 3 行、第 4 行和第 5 行的聚类结果方差相差不大,趋于平缓,由此将人脸基本表情模式的数目设为 29.图 4 显示出结果.

Fig.4 Tewnty-Nine classified FAP patterns

图 4 29 种 FAP 模式

3.2 人工神经网络

如果将语音到 FAP 模式的映射看做是一个模式识别的任务,有很多学习算法可被使用,如隐马尔可夫模型(HMM)、支持向量机(SVM)以及神经网络等等.由于神经网络对于学习输入输出映射体现出较强的效率和鲁棒性,我们选择一种神经网络(BP 网)来学习大量记录的句子.整个 3 层反馈神经网络构造如图 5 所示.

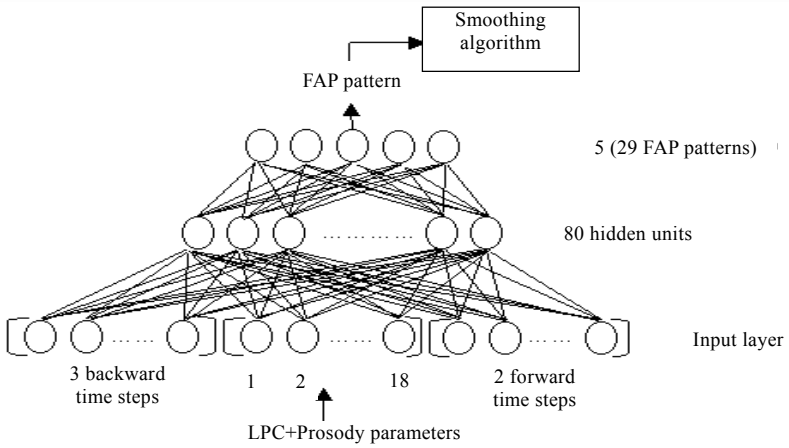


Fig.6 The model architecture of the ANN

图 5 神经网络结构示意图

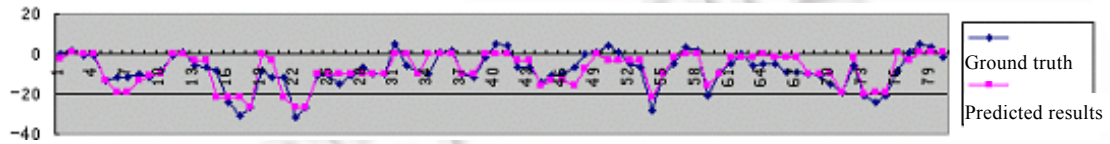
对于语音每一帧都计算 16 维 LPC 与 RASTA-PLP 混合向量加上 2 维韵律参数,形成 18 维语音特征向量,取前后 6 帧合为一个输入向量,这样,每次神经网络的输入是 108 维的向量.由于有 29 类,因此输出节点个数定为 5 个,这样可以覆盖 29 个类.中间隐层节点个数采用 80,同时神经网络的参数设为:学习率 0.001,网络的误差 0.001.由于只是简单地将人脸基本表情划分为 29 种,在用神经网络的输出驱动人脸动画时,会出现抖动现象,因此对于神经网络的输出序列,采用插值平滑算法以使整个动画过程更加逼真.

4 实 验

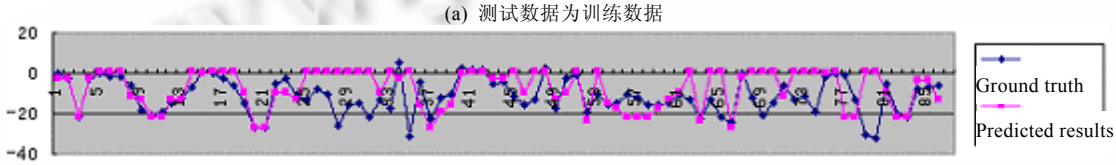
为了覆盖尽量全的单个人的发音,本文选择 863 中国语音合成库 CoSS-1 总结的文本资料作为话者发音的文字材料.CoSS-1 包含所有汉语 1 268 个独立音节的发音,也包含大量 2~4 字词的发音以及 200 个语句的语音.

为了有利于语句级的合成,聚类结果应从真实的语句中发现并获取,因此与很多研究者仅仅记录下词语和音节不同,我们主要记录的是语句的同步音视频库.通过标记特征点,可获取嘴唇、脸颊、眼皮等位置的运动数据.摄像机按 15 帧/秒将采集的视频转为图像,并利用跟踪程序处理得到图像特征序列.语音采样率为 8040Hz, 语音分析的窗长为 156,帧移为 90,在语音分析中应用海明窗,这样每一帧得到 16 阶 LPC 与 RASTA-PLP 混合系数以及一些韵律参数.用反馈神经网络训练从语音特征到 FAP 模式的映射.这些 FAP 模式可以用来驱动合成虚拟人脸模型.

对系统采用了定性和定量两种估价方法.定量测试是基于计算衡量预测数据与真实数据之间的误差.多数机器学习系统都应采用定量方法.定性测试是通过感知来判断合成出的人脸运动是否真实.对于合成而言,定性测试是非常重要的.在定量测试中,衡量了预测数据与真实数据的误差,包括闭集(训练数据为测试数据)和开集(测试数据没有经过训练)两组.图 6 显示了两句话中,上嘴唇高度参数值的测试结果,图 6(a)中的测试数据为训练数据,图 6(b)中的测试数据为非训练数据,通过测试所有 FAP 参数并按式(1)计算出预测数据和真实数据的均方差,得到表 2 的结果.



(a) Test data set is training data set
(a) 测试数据为训练数据



(b) Test data set is non-training data set
(b) 测试数据为非训练数据

Fig.6 Comparison of lip height parameter value

图 6 嘴唇高度参数的比较

Table 2 Comparison of mean errors squared of training data and non-training data

表 2 FAP 参数预测数据和真实数据的方差比较

Test data	Errors square between ground truth and predicted results
Selected from training data set (five sentences)	1.209 910
Selected from non-training data set (five sentences)	3.258 232

对于多模式系统的评价至今没有一个统一的标准,对于语音驱动人脸动画系统,由于无法得到任何人的与语音对应的人脸分析数据,无法计算预测数据与真实数据的误差,因此单纯定量结果并不能代表系统的实用性能.对于非特定人的语音测试评价,一般只能采用定性的方法.在实验中,要求 5 个人视听系统,并从智能性、自然性、友好性以及人脸运动的可接受性比较语音驱动和我们以前的文本驱动系统[13].由于现在的系统不仅可以解决人脸上部的动态变化,而且使用的是录制的原始语音,并可以有效地解决同步问题,因此得到了较高的评价.

利用本文所提系统,当给定一个人的语音后,神经网络可以实时预测每一帧语音特征对应的 FAP 模式,通过平滑后可直接驱动基于 MPEG-4 的人脸网格.图 7 给出了语音驱动人脸动画的部分帧.

5 结论和将来的工作

本文综合利用聚类和机器学习的方法学习语音信号和唇动人脸表情之间的同步关系,并将学习结果应用于基于 MEPG-4 标准的语音驱动人脸动画系统中.利用聚类发现了人脸运动的 29 种基本模式,利用神经网络完成从语音到人脸模式的映射.同时,给出系统定量的和定性的两种评价方法.所有实验表明,基于学习的语音驱动人脸动画不仅可以有效地解决语音视频同步的难题,增强动画的真实感和逼真性,同时基于 MPEG-4 的学习结果独立于人脸模型,可以用来驱动各种不同的人脸模型.目前,基于学习的人脸动画在国内研究不多,但国外已有较强的基础.在现有较好工作的基础上,将来的工作包括使系统对于环境以及话者的变化更加鲁棒、对聚

类和学习算法的改进,以及科学的多模式人机接口评价方法.

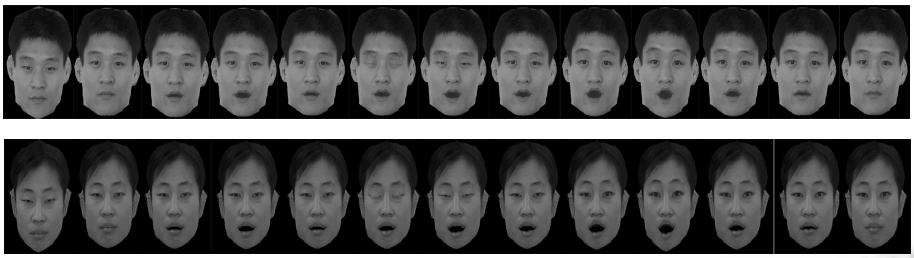


Fig.7 Example of speech driven face animation key frames

图 7 语音驱动人脸动画的示例

致谢 感谢中国科学院计算技术研究所领域前沿青年基金(20026180-17)提供资助.感谢北京工业大学尹宝才教授对本文工作的帮助.中国科学院计算技术研究所博士生左力、山世光博士对本文的完成提出了很多有益的建议,在此一并表示感谢.

References:

[1] Beskow J. Rule-Based visual speech synthesis. In: Proceedings of the 4th European Conference on Speech Communication and Technology. 1995. 299~302. http://www.speech.kth.se/~beskow/papers/es95rul.pdf.

[2] Waters K, Levergood, TM. DECface : an automatic lip-synchronization algorithm for synthetic face. Technical Report, CRL 93-4, Digital Equipment Corporation, Cambridge Research Laboratory, 1993. ftp://crl.dec.com/pub/DEC/CRL/tech-reports/93.4.ps.Z.

[3] Hong PY, Wen Z, Huang TS. IFACE: a 3D synthetic talking face. International Journal of Image and Graphics, 2001,1(1):1~8.

[4] Ezzat T, Poggio, T. Visual speech synthesis by morphing visemes. International Journal of Computer Vision, 2000,38(1):45~57.

[5] Yehia H, Kuratate T, Vatikiotis-Bateson E. Using speech acoustics to drive facial motion. In: Proceedings of the 14th international congress of phonetic sciences (ICPhS'99). 1999. 631~634. http://trill.berkeley.edu/ICPhS/frameless/acceptance.html.

[6] Massaro DW, Beskow J, Cohen MM. Picture my voice: audio to visual speech synthesis using artificial neural networks. In: Proceedings of the 4th Annual Auditory-Visual Speech Processing Conference (AVSP'99). 1999. 105~111. http://mambo.ucsc.edu/pdf/avsp9922.pdf.

[7] Brand M. Voice puppetry. In: Proceedings of the SIGGRAPH'99. 1999. 21~28. http://www.cs.cmu.edu/~ph/869/papers/Brand-sigg99.pdf.

[8] Ostermann J. Animation of synthetic faces in MPEG-4. Computer Animation, 1998. 49~51. http://www.research.att.com/projects/AnimatedHead/pimages/companim3.pdf.

[9] Zhen B, Wu XH, Liu ZM, Chi HS. An enhanced RASTA processing for speaker identification, In: Huang TY, ed. Proceedings of the International Symposium of Chinese Spoken Language Processing. Beijing: China Military Friendship Publish,2000. 251~255.

[10] Wang AH, Bao HQ, Chen JY. Primary research on the viseme system in standard Chinese, In: Huang TY, ed. Proceedings of the International Symposium of Chinese Spoken Language Processing. Beijing: China Military Friendship Publish, 2000. 215~218.

[11] Chen T, Rao R. Audio-Visual integration in multimodal communication. In: Proceedings of the IEEE, Vol 86. 1998. 837~852. http://citeseer.nj.nec.com/chen98audiovisual.html.

[12] Chen YQ, Gao W, Zhu TS, Ma JY. Multi-Strategy data mining framework for mandarin prosodic pattern. In: Yuan BZ, ed. Proceedings of the 6th International Conference on Spoken Language Processing. Beijing: China Military Friendship Press, 2000, II:59~62.

[13] Shan SG, Gao W, Yan J, Individual 3d face synthesis based on orthogonal photos and speech-driven facial animation. In: Proceedings of the International Conference on Image Processing (ICIP 2000), Vol III. 2000. 238~242. http://www.jdl.ac.cn/user/sgshan/pub/Shan-ICIP00.pdf.