[19] 中华人民共和国国家知识产权局

[51] Int. Cl[7]
G06N 3/02
G06F 17/30

[12] 发明专利申请公开说明书

[21] 申请号 02140286.8

[43] 公开日 2004 年 1 月 7 日

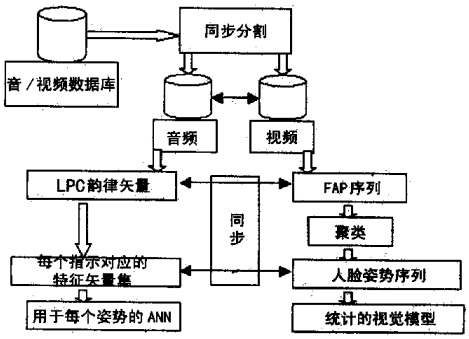
[11] 公开号 CN1466104A

[22] 申请日 2002.7.3 [21] 申请号 02140286.8
[71] 申请人 中国科学院计算技术研究所
地址 100080 北京市中关村科学院南路 6 号
[72] 发明人 陈益强 高 文 王兆其
[74] 专利代理机构 中科专利商标代理有限责任公司
代理人 戎志敏
权利要求书 2 页 说明书 13 页 附图 3 页

[54] 发明名称 基于统计与规则结合的语音驱动人脸动画方法

[57] 摘要

一种基于统计与规则相结合的语音驱动人脸动画方法，包括步骤：利用音视频同步切割方法得到音视频对应数据流；通过音视频分析方法，得到相应的特征向量；运用统计学习方法学习到音视频同步隐射关系模型；运用统计学习到的模型以及规则得到与用户给定语音序列相对应的人脸运动参数，并驱动人脸动画模型。 本发明使用视频采集，语音分析及图象处理等方法，记录真实人脸说话时的语音与人脸特征点运动数据，同时对语音和人脸特征点之间的关联模式进行统计学习。 当给定新语音，利用学习到的模型以及一些规则，可以得到与该语音对应的人脸特征点运动参数，驱动人脸动画模型。



ISSN1008-4274

知识产权出版社出版

1、一种基于统计与规则相结合的语音驱动人脸动画方法，包括步骤：

利用音视频同步切割方法得到音视频对应数据流；

通过音视频分析方法，得到相应的特征向量；

运用统计学习方法学习到音视频同步隐射关系模型；

运用统计学习到的模型加上规则得到与新语音相对应的人脸运动参数。

2、按权利要求 1 所述的方法，其特征在于所述的音视频同步分割方法包括步骤：

a、假设视频采集帧率为 Videoframecount/msec，音频帧率为 Audiosamplecount/msec，语音分析窗位移为 Windowmove，语音分析窗大小为 Windowsize，需要语音窗个数为 m，语音分析窗与语音分析窗位移比例为 n；

b、Windowmove=Audiosamplecount/(Videoframecount*m)

Windowsize=Windowmove*n

其中，m 与 n 为可调参数，根据实际情况设定。

3、按权利要求 1 所述的方法，其特征在于所述的音视频分析与特征提取方法包括步骤：

a、对于音频提取海明窗中语音数据的线性预测参数以及韵律参数（能量、过零率以及基频）作为语音特征向量

b、对于视频，提取人脸上与 Mpeg-4 一致的特征点，然后计算各特征点坐标与标准帧坐标的差值 Vel={V1,V2...Vn}，再计算按 Mpeg-4 定义的特定人脸上的各特征点对应尺度参考量 P={P1,P2,...,Pn}，通过公式（3）即可得到人脸运动参数。

$Fap_i=(V_{i(x|y)}/P_{i(x|y)})*1024$ （3） Fap_i 表示与第 I 个特征点对应的人脸运动参数，$V_{i(x|y)}$表示的 V_i 的 x 或 y 坐标，$P_{i(x|y)}$表示与 $V_{i(x|y)}$对应的尺度参考量。

4、按权利要求 1 所述的方法，其特征在于所述的音视频同步隐射关

系模型的统计学习方法包括步骤：

a) 首先得到同步分割特征集 Audio, Video;

b) 对 Video 集中视频进行无监督聚类分析，得到人脸运动基本模式，设为 I 类;

c) 利用统计方法得到两类或多类之间的转移概率，称为统计视觉模型，并用熵来评价模型的好坏，然后再进行 b)直到熵最小。

d) 将属于同一个人脸运动基本模式的对应的语音特征集 Audio 中的数据分成相应的子集 Audio(i), I 代表第几类。

e) 对每个子集 Audio(i)用一个神经网络进行训练，输入为子集中的语音特征 F(Audio(i))，输出为属于这个类别的近似程度 P(Video(i))。

5、按权利要求 1 所述的方法，其特征在于所述的得到与语音特征相对应的人脸运动参数包括步骤：

a) 对于给定新语音，提取语音特征;

b) 将语音特征作为输入送入每个人脸模式对应的神经网络，得到输出的属于这个类别的近似程度;

c) 当一个句子完成后，利用统计视觉模型及 Viterbi 译码算法得到一条最大概率的类的转移路线，连接起来就是与语音对应的人脸动画模式序列;

d) 对预测的人脸动画模式序列可以通过人脸运动知识库中的规则进行修订，使结果更加真实自然。

基于统计与规则结合的语音驱动人脸动画方法

技术领域

本发明涉及一种基于统计与规则结合的语音驱动人脸动画方法，尤指一种使用视频采集，语音分析及图象处理等方法，记录真实人脸说话时的语音与人脸特征点运动数据，建立一个初始的语音图像数据库；通过视频采集帧率与语音数据采样率可以计算出语音数据分析窗的位移量，同时利用这几个数据利用统计学习方法得到语音与视频帧对应的同步对应关系模型。利用这种模型，加上规则，可以得到任意语音对应的人脸运动参数，驱动人脸动画模型。

背景技术

在通过一幅或几副图象或视频序列恢复具有真实感的三维人脸的技术成为现实后，目前研究为具有真实感的三维人脸行为的模拟。和语音合成中遇到的问题一样，得到大量真实人脸运动图象和人脸合成基元并不难，困难在于如何编辑和重用这些存在的人脸动画数据。一种方法是提供一套方便的用于手工编辑的工具，将编辑好的关键帧做插值后生成动画序列，这种方法最直接，但需要熟悉动画的专家花大量的时间制作。第二种采用控制技术，用其他相关的信号比如文本，声音，视频，或传感器实现对人脸动画的控制。用文本控制，输出的声音是合成语音，而且同步很难掌握。通过视频控制，对视频图象的跟踪与特征提取是一个难点。采取传感器方案的话，设备造价太高，而且一些细节的特征点的变化只能估算出来。因此现在可行的而且很多研究者在做的是实现语音驱动人脸动画。人们对于人脸行为非常敏感，对于是否具有真实感很容易判断，并且也容易从声音信号找到对应的人脸运动行为。要实现语音驱动的人脸动画，语音与唇动以及人脸表情合成之间的关联模式对于人物的真实感和可信度是至关重要的。

认知学家与心理学家已经观察到有大量的相关信息存在语音和人脸行为中。脸部信息可以增加观察者对语音内容以及形式上的理解，并且被很多基于语音界面的系统考虑。相反，合成可信度较高的人脸被认为是生成可接受的虚拟人和动画人的主要障碍。人们对于解释人体运动行为有较高的敏感性，不真实自然的动画人脸通常会干扰甚至打断人们对语音的理解。目前的语音驱动研究可分为两类：通过语音识别和不通过语音识别。第一种方法是通过将语音分割成语言单元，如音素（Phoneme），视觉基元(Viseme)以及更进一步音节(syllable)，随后将这些语言单元直接隐射到嘴唇姿势后用拼接法合成。这种方法非常直接易于实现，但缺点是忽视了动态因素并且同步问题---潜在的语音段落与肌肉模型运动的相互作用及影响很难处理。到现在为止，几乎所有的同步问题上的努力集中在启发式规则以及经典的平滑方法上。比如 Baldy 是一个语音基元驱动的 3D 虚拟人脸系统，对于同步问题的处理采用心理学家认可的手工设计的语音同步模型。虽然视频重写（Video Rewrite）方法通过对三音子对应的视频段排列得到新的视频，结果比生成的动画模型自然，但值得指出的是，三音子所表示的是语音之间的过渡连接，并不代表人脸帧之间的运动。同时系统的好坏取决于提供三音子样本的数目以及平滑技术。当我们用离散的语音基元或图象基元表示音视频的基本单元时，很多必要的信息会被丢失。事实上，语音基元的设计仅满足区别发音高低以及可以传递语言内容的需要。语音基元表示对于识别而言非常有效但对于合成来说却不是最好的，这主要由于他们很难预测声音韵律和人脸表情之间，声音能量与姿势放大之间，以及声音段落与唇动同步之间的关系。第二种方法是绕过语音基元这种形式，找到语音信号与控制参数之间的隐射关系，然后直接驱动嘴唇运动。可以用神经网络进行训练，用前后各五帧语音信号去预测控制参数。但一般采用手工标定对应语音段控制参数的方法，虽然回避了人脸特征点自动获取的难题，但同时也导致系统难以描述人脸复杂的变化。也有将一些 3D 位置跟踪器安放在嘴唇旁边以及脸颊周围，虽然可以获得人脸运动的准确数据，但对于人脸上部如眼睛，以及眉毛等等的变化却没有实现。有人提出用一种用相关信号预测控制信号的方法（HMM），并将它用于语音驱动人脸动画中。但

用一个 HMM 处理复杂的音频数据将问题简化了。同时以上处理都是基于统计学习的，可以处理语音与唇动等关联性较强的隐射，但对于语音与眨眼，语音与头势等弱关联关系则难以通过学习得到。

发明内容

本发明的目的是提供一种采用基于统计和规则相结合的方法实现语音到人脸的映射的方法。

为实现上述目的，本发明提供的方法包括步骤：

利用音视频同步切割方法得到音视频对应数据流；

通过音视频分析方法，得到相应的特征向量；

运用统计学习方法学习到音视频同步隐射关系模型；

运用统计学习到的模型以及规则得到与新语音相对应的人脸运动参数，驱动人脸动画模型。

本发明使用视频采集，语音分析及图象处理等方法，记录真实人脸说话时的语音与人脸特征点运动数据，建立一个初始的语音图像数据库；通过语音分析可以得到语音特征，包括线性预测系数以及韵律参数（能量以及过零率和基频），从视频帧可以提取 MPEG4 定义的人脸动画参数对应的特征点，通过相对帧作差计算以及相对位移计算可以得到人脸动画参数。利用聚类，统计以及神经网络等方法完成从语音特征到人脸动画参数的学习映射。学习后，当新的语音进来，通过分析可以得到语音特征，语音特征通过映射可以得到人脸动画参数，在此基础上，运用人脸运动知识库，在结果上加入规则约束，实现真实感的动画。

附图说明

图 1 是学习阶段框架示意图；

图 2 是人脸特征点跟踪示意图；

图 3 是特征点检测与影响区域示意图；

图 4 是 MPEG4 中的部分 FDPFAP 对应点及 FAPU；

图 5 是 29 种 FAP 模式；

图 6 是应用阶段框架示意图；

图 7 是统计视觉模型方法与基于神经网络方法的比较（嘴唇高度参数的比较）；

图 8 是语音驱动人脸动画示例，上图为真实音视频，下图为根据本发明的利用音频得到的人脸运动序列。

具体实施方式

首先利用无导师聚类分析可得到视频的人脸运动参数（FAP）特征向量类。然后统计与语音事件同步发生的人脸动态模型（实质为 FAP 类别转移矩阵），我们称之为统计视觉模型，其原理与自然语言处理中的统计语言模型类似。最后学习训练多个神经网络（ANN）完成从语音模式到人脸动画模式的隐射。机器学习后对新语音数据通过计算可以得到一些人脸动画模式序列，利用统计视觉模型可以从中选出最佳人脸运动参数（FAP）序列，然后利用人脸运动规则对 FAP 序列进行修正和补充，完成平滑后，使用这些 FAP 可以直接驱动人脸网格模型。这种策略有以下独到之处：

1)整个过程建立可以用经典的 Bayes 规则给出描述，

$$\arg\max_{L} \Pr(L \mid A) = \arg\max_{L} \frac{\Pr(A \mid L).\Pr(L)}{\Pr(A)},$$

其中 A 可以看作语音信号。最大似然估计 Pr(A|L) 衡量对语音信号模型化的准确度，先验模型 Pr(L) 建立关于真实人脸运动的背景知识或称统计视觉模型。

2）语音信号的聚类分析是建立在对人脸姿势的分类学习上，这样做比考虑假设的通过语音感知分类要好。同时，由于同一唇形对应完全不同的语音特征，因此，采用神经网络对同一类的语音信号训练，可使得预测结果的鲁棒性提高。

3）统计视觉模型容许我们找到整句话优化的人脸运动轨迹，充分使用上下文信息同时避免了神经网络训练难以实现上下文相关的缺陷。

4）视频信号仅需分析一次，用来训练语音与人脸动画参数（FAP）的对应关系，结果模型可以被用来做其他人的人脸合成。

5）人脸运动规则的引入使原来与语音关联程度不高的部分的动画也能更加真实，如眨眼和头动等。

6）整个框架可以用于其他信号之间的相关预测和控制或合成。

上述基于统计与规则相结合的语音驱动人脸动画方法包括如下两个方面：学习与应用阶段：

1）学习阶段包括如下步骤（图 1）：

a）音视频同步录制与分割

通过摄象机可以同步的录制语音和视频数据，形成 AVI 文件，但为了以后分析需要，必须将音视频信号分为不同通道的音频和视频流。传统做法通常根据经验，对采用的某种摄象机固定设置，本发明提出音视频同步分割方法可用于任意摄象机采集视频。

假设视频采集帧率为 Videoframecount/msec，音频帧率为 Audiosamplecount/msec，语音分析窗位移为 Windowmove，语音分析窗大小为 Windowsize，需要语音窗个数为 m，语音分析窗与语音分析窗位移比例为 n；

Windowmove=Audiosamplecount/(Videoframecount*m) （1）；

Windowsize=Windowmove*n （2）；

其中 m 与 n 为可调参数，根据实际情况设定。按这种方法设置的同步参数可以使音视频同步精确到采样位。

为了覆盖尽量全的各种发音，方法选择 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 1268 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。记录下各种单字，词及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。设置摄象机按 10 帧/秒将采集的视频转为图象并利用跟踪程序处理得到图象特征序列。假设 m=6，n=2/3 我们采用语音采样率为 8040Hz，则语音分析的窗长为 8040/10*6=134，帧移为 134*2/3=89。

b）音视频特征提取。

对于音频提取海明窗中语音数据的线性预测参数以及韵律参数（能

量、过零率以及基频）作为语音特征向量

对于视频，提取人脸上与 Mpeg-4 一致的特征点，然后计算各特征点坐标与标准帧坐标的差值 Vel={V1,V2...Vn}，再计算按 Mpeg-4 定义的特定人脸上的各特征点对应尺度参考量 P={P1,P2,...,Pn}，通过公式（3）即可得到人脸运动参数。

$Fap_i=(V_{i(x|y)}/P_{i(x|y)})*1024$ （3） Fap_i 表示与第 I 个特征点对应的人脸运动参数，$V_{i(x|y)}$表示的 V_i 的 x 或 y 坐标，$P_{i(x|y)}$表示与 $V_{i(x|y)}$对应的尺度参考量。

对于语音特征，在语音分析中应用传统的海明窗，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。

对于人脸运动特征，使用基于 MPEG4 的人脸动画表示方案。MPEG-4 使用 FDP（人脸定义参数）和 FAP（人脸动画参数）指定人脸模型及其动画，使用 FAPU（人脸动画参数单元）标示 FAP 的位移活动。基于上述原理，获取人脸表情和唇动运动数据，就是要获取相应的 FDP 和 FAP 参数。为了获得人脸运动数据，开发了一套计算机视觉系统可以同步跟踪许多个性的人脸特征如：嘴角以及嘴唇线，眼睛以及鼻子等。图 2 显示我们可以跟踪和得到的特征点。由于获取精确的特征点运动数据比实验众多的跟踪算法对我们合成更重要。我们采用通过在脸上标记特定颜色的做法来获取数据并且要求话者尽量减少头部运动，图 3 显示最终获得的特征点以及影响区域。

通过特征点提取出来的数据是绝对坐标，而且由于话者头部运动或身体运动的影响，使得用简单图象处理得到的坐标值具有很大的噪音，因此需要进行归正预处理。我们假设不受 FAP 影响的特征点是相对不运动的，利用这种不变性完成从图象坐标到人脸模型相对坐标的变换，从而可以去除由话者运动引起的旋转和伸缩变化对数据的影响。对图 4 中 Mpeg4 定义的特征点，我们选取了 P0(11.2),P1(11.3),P2(11.1) 和 P3(多加在鼻尖上的一个点)形成正交坐标系(X 轴 P0 P1 ,Y 轴 P2 P3) ，根据这个坐标系，按照以下方法可计算出旋转角度以及伸缩尺度。假设这些参考点的坐标为 P0 (x0,y0), P1 (x1,y1), P2 (x2,y2) and P3 (x3,y3), 新坐标体系的原点坐标可以由它们连接成的两条直线交点算出，假设为

P(xnew, ynew) 同时还可以算出新坐标相对于正交坐标的旋转角度 Φ。这样任意点 (x, y) 在新坐标体系下的值（x', y'）可以按照如下公式计算：

$$x' = x \times \cos(\theta) - x \times \sin(\theta) + P(x_{new}) \quad (4)$$

$$y' = y \times \sin(\theta) - y \times \cos(\theta) + P(y_{new}) \quad (5)$$

为了避免伸缩影响，假设加在鼻梁上的点相对于第一帧是不运动的，任何其他点可以根据式（6）和（7）计算与这一点的相对位移，从而将图象坐标转为人脸模型坐标，得到特征点运动的准确数据：

$$x''_k = (x'_k - x_{k3}) - (x'_1 - x_{13}) \quad (6)$$

$$y''_k = (y'_k - y_{k3}) - (y'_1 - y_{13}) \quad (7)$$

其中(x_{13}, y_{13})表示第 1 帧的鼻尖点的坐标，(x'_1, y'_1)表示第 1 帧其他特征点的坐标，(x_{k3}, y_{k3}) 表示第 k 帧的鼻尖点的坐标，(x'_k, y'_k)表示第 k 帧其他特征点的坐标，(x''_k, y''_k)表示第 k 帧其他特征点的最后计算坐标。通过滤波后，每一特征点的坐标都可以参照图 4 定义的人脸动画参数单元(FAPU)计算出 FAP 值。假设图 4 中定义的 ESO 以及 ENSO 分别为 200 和 140，则 5.3（x, y）对应于两个 FAP 值分别可以计算为：

$$FAP39 = X \times 1024/200 \quad (8)$$

$$FAP41 = Y \times 1024/140 \quad (9)$$

c）音频特征到视频特征的统计学习。

①首先将音视频按 a)，b)所述得到同步分割特征集 Audio, Video;

②对 Video 集中视频进行无监督聚类分析，得到人脸运动基本模式，设为 I 类;

③利用统计方法得到两类或多类之间的转移概率，称为统计视觉模型，并用熵来评价模型的好坏，然后再进行 b)直到熵最小。

④将属于同一个人脸运动基本模式的对应的语音特征集 Audio 中的数据分成相应的子集 Audio(i)，I 代表第几类。

⑤对每个子集 Audio(i)用一个神经网络进行训练，输入为子集中的语音特征 F(Audio(i))，输出为属于这个类别的近似程度 P(Video(i))。

1. ②中人脸运动基本模式聚类分析方法

对于基本人脸模式，认知学家给出了一些研究成果，但一般都是定性给出 6 种基本表情或更多，这种定性表达合成结果的真实感不好。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，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 MPEG4 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合成复杂过程解释和理解打下基础。在聚类过程中，由于这样的基本模式的个数并不确定，一般采用无导师聚类。

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困难。从聚类数据的类间距或类内距虽可以得到用于指导聚类评价，但无法描述在实际系统中应用聚类可以达到的效果，通常效果的好坏对于动画系统是至关重要的，我们直接采用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以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（10）都进行方差计算，结果如表 1 所示：

$$\text{ErrorSquar e}(X, Y) = \sqrt{(X - Y) * (X - Y)^T} \Big/ \|X\| \qquad (10)$$

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。

	每类最小样本数	分离参数 P/合并参数 C	聚类数目	方差比较
1	32	P=0.5-1, C=1-1.5	18	3.559787
2	20	P=0.5-1, C=1-1.5	21	4.813459
3	10	P=0.5-1, C=1-1.5	23	2.947106
4	5	P=0.5-1, C=1-1.5	29	2.916784
5	3	P=0.5-1, C=1-1.5	33	2.997993

表 1：聚类结果比较

上述聚类是在 6200 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 和 C 都在[0.5,1]区间中变化。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。从方差估计可看出，第 3 行，第 4 行和第 5 行的聚类结果方差相差不大，可认为趋于平缓，由此将人脸基本表情模式的数目设为 29。图 5 显示出结果：

2. ③中的统计视觉模型建立方法

建立统计视觉模型的目的是为了容许找到整句话优化的人脸运动轨迹，充分使用上下文信息同时避免了单一神经网络训练难以利用上下文相关的缺陷。统计视觉模型可以计算出视频序列出现的概率。如果我们假设 F 是一特定语句的人脸动画序列，如，

$$F = f_1 f_2 \cdots f_Q$$

那么，$P(F)$ 可以由下列公式计算得到

$$P(F) = P(f_1 f_2 \cdots f_Q) = P(f_1)P(f_2 \mid f_1) \cdots P(f_Q \mid f_1 f_2 \cdots f_{Q-1}) \quad (11)$$

然而，对于任意人脸姿势以及所组成的序列，估计所有可能的条件概率 $P(f_j \mid f_1 f_2 \cdots f_{j-1})$ 是不可能的，在实际中，一般采用 N 元文法来解决这个问题，可以近似估计 、、、为

$$P(F)=\prod_{i=1}^{Q}P(f_i \mid f_{i-1}f_{i-2}\cdots f_{i-N+1}), \quad (12)$$

条件概率$P(f_i \mid f_{i-1}f_{i-2}\cdots f_{i-N+1})$ 可以通过简单的相对统计方法得到：

$$P(f_i \mid f_{i-1}f_{i-2}\cdots f_{i-N+1})=\frac{F(f_i,f_{i-1},\cdots f_{i-N+1})}{F(f_{i-1},\cdots f_{i-N+1})} \quad (13)$$

其中，F 是各种人脸姿势在给定的训练视频数据库中的同现次数。建立统计视觉模型后，我们采用困惑度来估计整个训练模型的性能好坏。假设θ_i是通过聚类分析得到的聚类集合 I 的聚类中心，对于$\theta=\{\theta_1,\theta_2\cdots\theta_n\}$，我们希望找到一个优化的视觉模型。对于模型$\theta$的困惑度可根据如下方法定义：

$$pp = 2^{H(S,\theta)} \approx 2^{-\frac{1}{n}\log p(S\mid\theta)} \quad (14)$$

其中 $S=s_1,s_2,\cdots,s_n$ 表示语句的人脸动画参数序列。$p(S\mid\theta)=\sum_i p(s_{i+1}\mid s_i\cdots s_1)$ 表示人脸动画参数序列 S 在模型$p(\theta)$下的概率. $p(\theta)$实质上表示我们对于人脸运动的背景知识，同时可以利用上述的统计方法获取。比如可以用自然语言处理中常用的二元文法或三元文法的方法，表 2 显示不同聚类结果得到的统计视觉模型的困惑度比较：

	Number of state	Bi-gram (PP)	Tri-gram (pp)
1	18	8.039958	2.479012
2	21	6.840446	2.152096
3	26	5.410093	1.799709
4	29	4.623306	1.623896
5	33	4.037879	1.478828

表 2：困惑度比较

通过统计视觉模型，我们得到一组状态转移的分布概率，当有多个人脸动画序列给出时，可以利用 Viterbi 算法求出在概率上最大可能发生的人脸动画序列。

3. ⑤中的神经网络学习方法

如果将语音到 FAP 模式的映射看作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。每一个聚类节点可以用两个神经网络完成训练，一个用于表征状态，取值为 0 或 1，另一个用于表征速度。这两种反馈神经网络可以统一描述为：

$$y_k = f_2(\sum_{j=0}^{n_2} w_{kj}^{(2)} f_1(\sum_{i=0}^{n_1} w_{ji}^{(1)} x_i)) \qquad (15)$$

其中 $x \in \Phi$ 是音频特征，$w^{(1)}$ 和 $w^{(2)}$ 是每一层的权值以及阈值，f_1 and f_2 是符号函数。训练非常简单，给定数据集后，采用 Levenberg-Marquardt 优化算法调整权值以及阈值来训练神经网络。

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。对于状态神经网络，输出节点个数定为 1 个，表示 0 或 1。对于中间隐层节点个数采用 30，同时神经网络的参数设为：学习率 0.001，网络的误差为 0.005。 对于速度神经网络，输出节点个数定为 18，表示 18 维 FAP 特征向量。对于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，网络的误差为 0.005.

2）应用阶段包括如下步骤（图 6）：

1）音频录制：

可直接利用麦克风或其他录音设备获取语音数据

2)音频特征提取

按照学习阶段的音频特征提取方法提取语音特征

3)基于统计学习模型的音频特征到视频特征的映射

将语音特征作为输入送入每个人脸模式对应的神经网络，每个状态神经网络都有一个输出，得到输出的属于这个类别的近似程度；当一个句子完成后，利用统计视觉模型及 Viterbi 译码算法得到一条最大概率的类的转移路线，连接起来就是与语音对应的人脸动画模式序列；

虽然主要还是由语音提供的信息起主要作用，但 viterbi 算法保证生成序列符合人脸的自然运动。虽然直接用每个聚类中心代表序列的每个状态就可以驱动人脸网格，但由于简化选取基本模式，人脸动画会出现抖动现象。传统方法一般用插值来解决，虽然可以消除抖动，但不符合人脸动画的动态特性，我们现在在每个状态下都有两个神经网络来预测，其中一个预测速度，这样利用转移矩阵得到的最终结果序列包含有足够的信息可以生成与自然人脸运动一致的动画， 整个公式非常简洁，令 $T=\{t_1,t_2\cdots t_n\}$ 为预测的人脸运动状态点， $V=\{v_1,v_2\cdots v_n\}$ 为每个状态点下的速度。

$$Y_{(t*i/m)->t+1}=Y_t+((Y_{t+1}-Y_t)/m)*v_t*i \qquad \text{If } i<=m/2 \qquad (16)$$

$$Y_{(t*i/m)->t+1}=Y_{t+1}-((Y_{t+1}-Y_t)/(i*m))*v_{t+1} \qquad \text{If } i>m/2 \qquad (17)$$

其中 $Y_{(t*i/m)->t+1}$ 表示从状态 t 到状态 t+1 的第 I 帧，m 表示从状态 t 到状态 t+1 需要插入的帧数. 由于有了速度参量，使得生成的人脸动画比插值方法更加符合人脸运动的多变性。

4)基于人脸运动规则的视频特征流修正

在得到基于统计模型的人脸运动参数序列后，由于学习预测结果的一点小的影响，会导致整个动画序列的真实感下降，同时有些人脸运动与语音特征的关联程度不大，如眨眼，点头，为此，在统计学习的基础上，加入人脸运动知识库的规则对序列进行修正，从而改善结果输出，使动画真实感更强。

5)音视频同步播出

得到语音以及动画播放文件，可在不同的通道直接播出，由于本身得到的数据是严格同步的，因此播出也是同步的。

四）实验结果比较

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，对很多机器学习系统，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 7 显示两句话中上嘴唇高度参数值的测试结果，并且与单个神经网络方法进行对比，上两图测试数据为训练数据，下两图测试数据为非训练数据，通过测试所有 FAP 参数并按式（10）计算出预测数据和真实数据的均方差，得到表 3 的结果。

测试数据	均方差 (VM+ANN)	均方差(ANN)
训练数据	2.859213	3.863582
测试数据	4.097657	5.558253

表 3：FAP 参数预测数据和真实数据的方差比较

对于多模式系统的评价至今没有统一标准，对于语音驱动人脸动画系统， 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，要求五个人视听系统，并从智能性，自然性，友好性以及人脸运动的可接受性进行评估。由于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，并有效解决同步问题，因此得到了较高的评价。

利用本文的系统，当给定一个人的语音后，神经网络可以实时预测每 帧语音特征对应的 FAP 模式，通过平滑后可直接驱动基于 Mpeg4 的人脸网格。图 8 给出语音驱动人脸动画的部分帧。

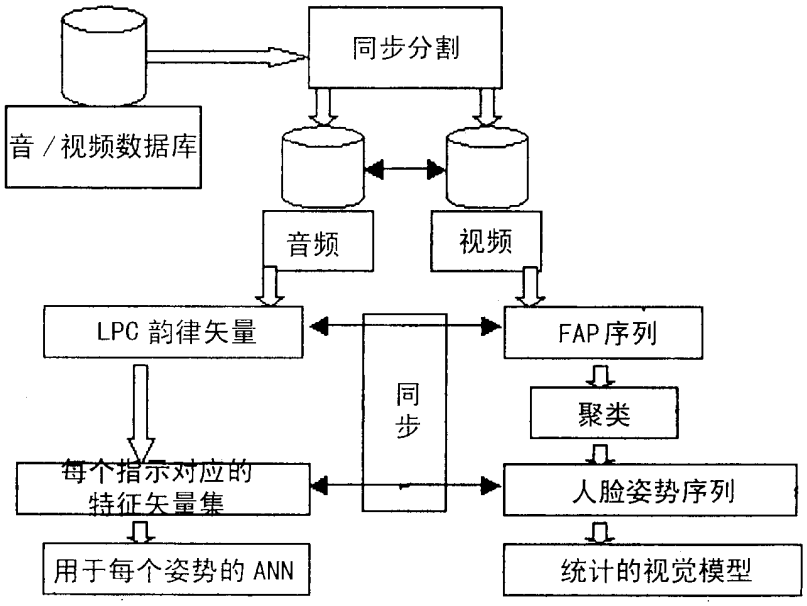

同步分割
音/视频数据库
音频
视频
LPC 韵律矢量
FAP序列
同步
聚类
每个指示对应的特征矢量集
人脸姿势序列
用于每个姿势的 ANN
统计的视觉模型

图 1

图 2

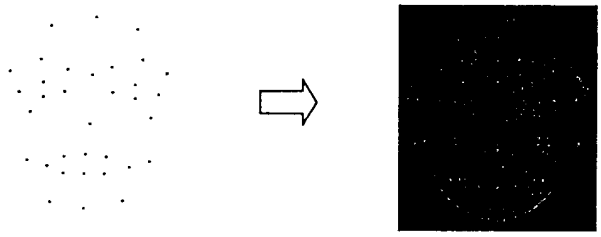


图 3

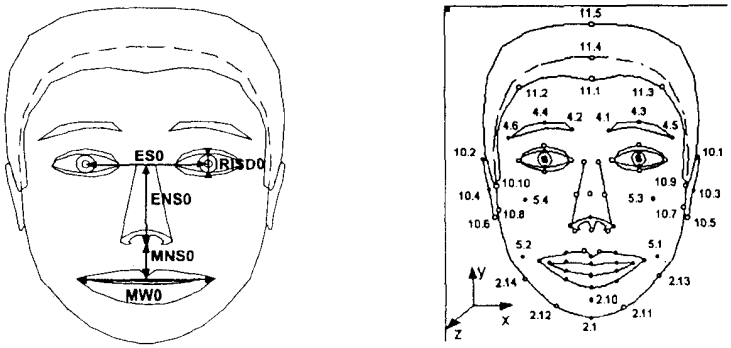

ES0
IRISD0
ENS0
MNS0
MW0
11.5
11.4
11.2
11.1
11.3
4.4
4.2
4.1
4.3
4.6
4.5
10.2
10.1
10.10
10.9
10.4
5.4
5.3
10.3
10.8
10.7
10.6
10.5
5.2
5.1
2.14
2.13
2.10
2.12
2.11
2.1
y
x
z

图 4

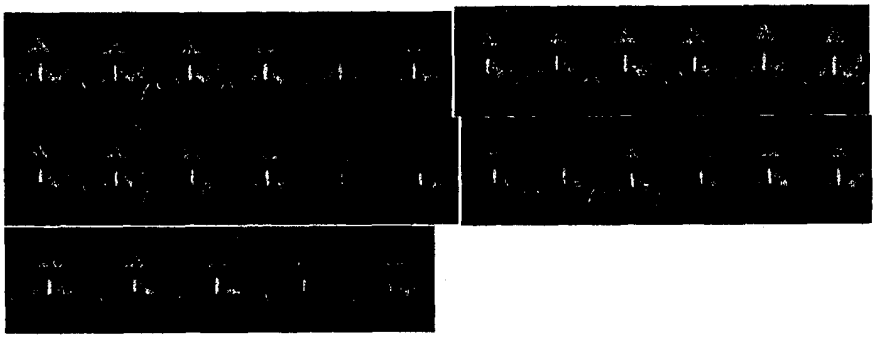


图 5

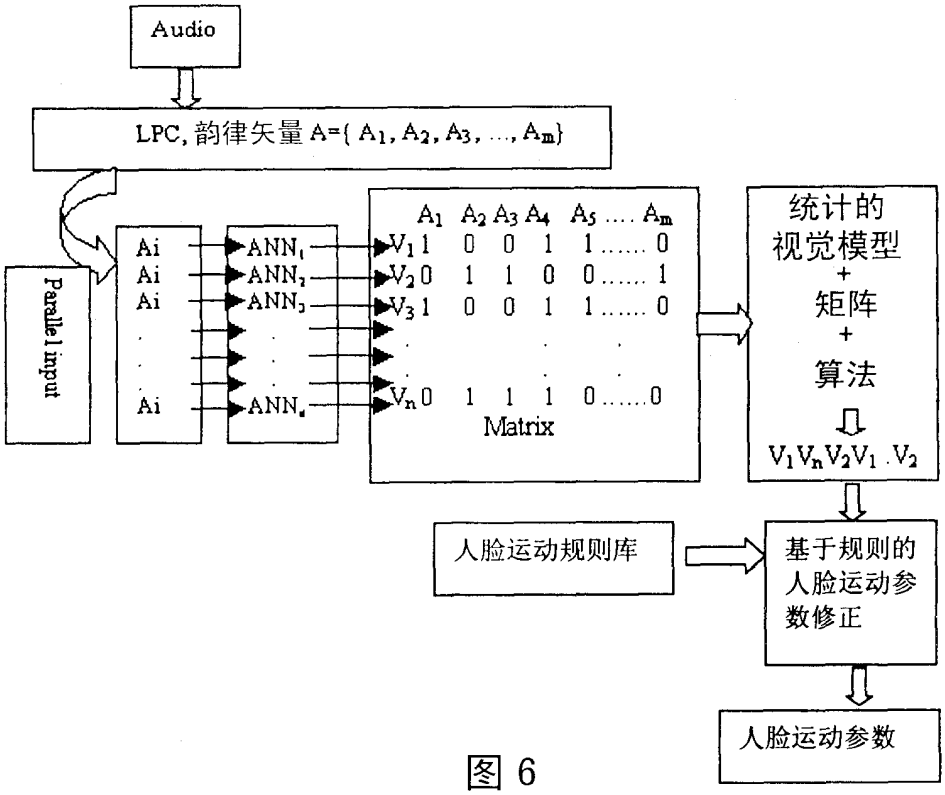

Audio
LPC, 韵律矢量 A={ A1, A2, A3, ..., Am}
Parallel input
Ai
Ai
Ai
Ai
ANN1
ANN2
ANN3
ANNn
A1 A2 A3 A4 A5 .... Am
V1 1 0 0 1 1 0
V2 0 1 1 0 0 1
V3 1 0 0 1 1 0
Vn 0 1 1 1 0 0
Matrix
统计的
视觉模型
+
矩阵
+
算法
V1 Vn V2 V1 . V2
人脸运动规则库
基于规则的
人脸运动参
数修正
人脸运动参数

图 6

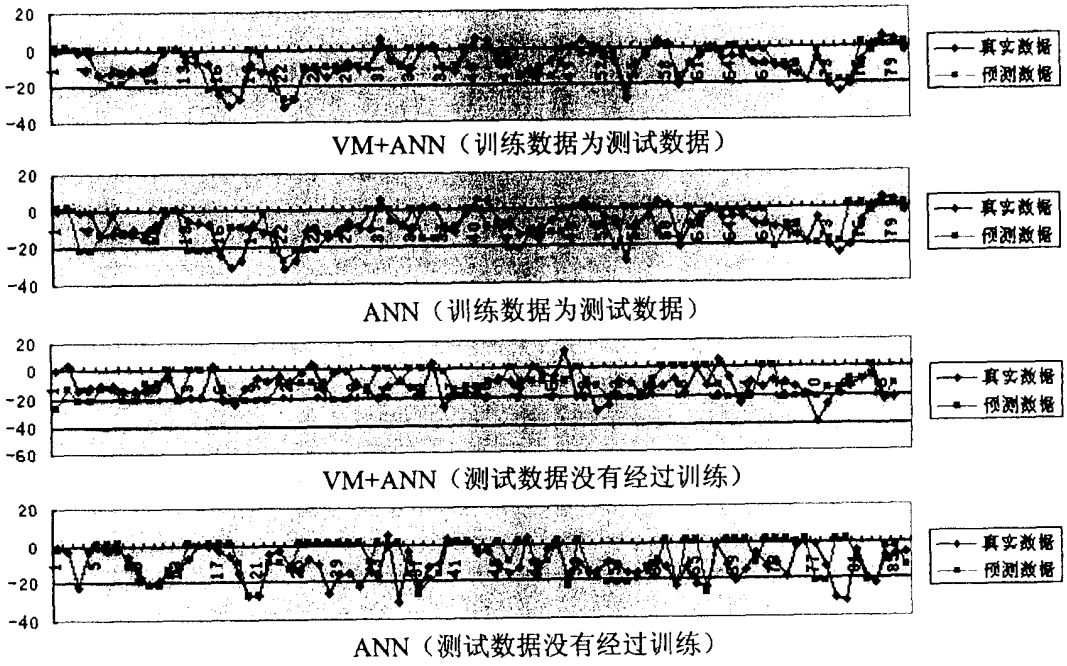

20
0
-20
-40
真实数据
预测数据
VM+ANN（训练数据为测试数据）
20
0
-20
-40
真实数据
预测数据
ANN（训练数据为测试数据）
20
0
-20
-40
-60
真实数据
预测数据
VM+ANN（测试数据没有经过训练）
20
0
-20
-40
真实数据
预测数据
ANN（测试数据没有经过训练）

图 7

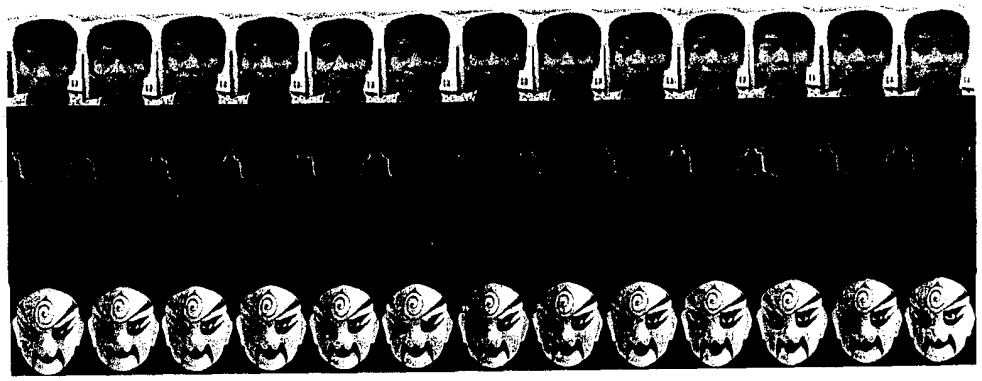


图 8