面向中国手语合成的口型与表情库构建

李敬华，王立春，王 振，孔德慧，尹宝才
（北京工业大学 计算机学院 多媒体与智能软件技术北京市重点实验室，北京 100124）

摘 要：为提高手语合成的真实感和可懂度，分析了手语表达中唇动和表情运动的特点，基于MPEG-4的参数化表达模型，提出了兼容于MPEG-4的口型和表情库构建方法，并基于该库完成了包含手势、唇动和表情的虚拟人手语动画. 实验结果表明，包含唇动和表情的合成手语的真实感得到增强，进而说明了口型和表情库的有效性.

关键词：手语合成；多模式；人脸定义参数(FDP)；人脸动画参数(FAP)
中图分类号：TP 391.9 **文献标志码**：A **文章编号**：0254-0037(2012)11-1665-05

Lip Movement and Expression Database Construction for Chinese Sign Language Synthesis

LI Jing-hua, WANG Li-chun, WANG Zhen, KONG De-hui, YIN Bao-cai
(Beijing Key Laboratory of Multimedia and Intelligent Software Technology, College of Computer Science, Beijing University of Technology, Beijing 100124, China)

Abstract: To synthesize realistic and intelligible lip motion and expression for Chinese sign language, based on MPEG-4 parametric model, a kind of method for constructing lip motion and expression database was proposed. When the lip motion and expression database were used to synthesize Chinese sign language animation, compared with the previous synthesized result, the animation including gesture, lip motion and expression was more realistic and intelligible. Experimental results show that the method for constructing lip motion and expression is valid.

Key words: sign language synthesis; multi-modal; facial definition parameters(FDP); facial animation parameters (FAP)

第1个美国手语合成系统研究开始于1982年，随着手语合成技术的日趋成熟，近来人们开始更多地关注提高手语合成的真实感和可懂度，致力于多模式手语合成研究[1-2]、手语韵律描述的研究[3]、手语手势动作表达力的研究[4]等. 其中多模式手语合成除合成手势这一主体模式外，也考虑头、面部、口型等非手势行为的辅助作用，进而增强虚拟人手语动画的表现力. 多模式手语合成通常基于多模式数据库采用拼接的方法合成，目前中国手语手势词库已比较完备，不过由于唇动和表情等面部运动的细微性，受其三维数据获取比较困难的限制，表情和口型库还不存在. 建立中国手语口型和表情库的关键是数据捕获，关于面部运动数据采集的方法主要有2类：一类是通过分析处理图像或视频获取数据，另一类是通过使用各种基于光学、传感器的运动捕捉设备获取数据. 前者采集的是二维数据，不能从中提取得到人脸在纵深方向的运动. 本文在分析中国手语表达中唇动与表情运动表现特点的基础上，基于MPEG-4的参数化表达模型表示口型和表情运动，并采用Motion Analysis公司的Hawk数字动作捕

收稿日期：2011-02-22.
基金项目：国家自然科学基金资助项目(60825203)；北京市自然科学基金资助项目(4112008).
作者简介：李敬华(1979—)，女，博士研究生，主要从事人机交互方面的研究，E-mail：lijinghua@bjut.edu.cn.

捉系统进行唇动和表情数据的采集,进而建立了面向中国手语合成的口型和表情库,最后将该库直接应用到中国手语合成中.

1 中国手语表达中的口型与表情

1.1 中国手语表达中的口型

中国手语表达中的口型指中国手语表达过程中与手势伴随的汉语词的口型动作[5],为此本文针对汉语的发音特点建立口型库,即以汉语的音节为基元,建立口型库.

汉语发音是拼读音节的过程,而音节是由声母和韵母拼接而成的,其中声母23个,韵母38个,汉语音节共有398个[6]. 汉语拼音音节的许多声母和韵母的发音口型非常相似,如果对每个音节都建立口型库,数据库的规模就会很大,并且有冗余信息. 王志明等[6]根据汉语发音特点和音位组成对声母口型和韵母口型进行了分类. 本文以其声母发音口型的分类为基准,依据汉语拼音音节表找到以各类声母开头的所有汉语音节,然后依据其韵母发音口型分类,去掉韵母相同或相似的冗余音节,得到简化的汉语音节发音口型分类表,比如/ba/、/pa/、/ma/只需采集/ba/即可,再比如/ba/、/bang/只需采集/ba/即可. 这样做排除了不同韵母对声母协同发音的影响,也解决了多个声母或韵母对应一个口型的问题. 本文针对表1的汉语音节发音口型分类表,建立了汉语音节口型库,口型库包含148个汉语音节的发音口型.

表1 汉语音节发音口型分类表

Table 1 Classification of Chinese syllables

声母	以该类声母开头的汉语音节
b,p,m	ba,bai,bao,bo,beng,bei,bi,bu,bin,bing,biao,bian,bie,me,miu,mou
f	fa,fan,fo,fou,feng,fei,fu
d,t,n	da,dai,dao,dou,de,dei,di,du,dong,dia,die,ding,diao,diu,dian,duo,dui,duan,dun,nin,nong,nü,nüe
l	la,lai,lao,lo,lou,le,lei,li,lu,lü,long,lie,lin,ling,liao,liu,lian,liang,luo,luan,lun,lüe
g,k,h	ga,gai,gao,gou,ge,gei,gu,gua,guo,guai,gui,gun,gong
j,q,x	ji,ju,jia,jie,jin,jing,jiao,jiu,jian,juan,jun,jue,jiong
zh,ch,sh,r	zha,zhai,zhao,zhou,zhe,zhen,zhi,zhu,zhong,zhua,zhuo,zhuai,zhui,zhun
w	wu,wa,wo,wai,wei,weng
y	yi,ya,ye,yao,you,yan,yin,ying,yong,yu,yue,yuan,yun
z,c,s	za,zai,zao,zou,ze,zei,zi,zu,zong,zuo,zui,zuan,zun
	a,ai,ao,o,ou,e,ei,er

1.2 中国手语表达中的表情

表情在中国手语合成中起着重要作用,中国手语表达中的表情指在打手势的同时脸上做出的各种各样的配合手势的表情或者是单独出现的表情[5]. 本文从中国手语合成的角度出发,重点考虑聋人使用中国手语时的表情,一方面是为表达喜、怒、哀、乐等内心情感而表现出的面部表情,另一方面源于《中国手语》[7]对部分手语词根描述的表情.

1.2.1 基本表情

MPEG-4标准中定义了人脸的6种基本表情,包括高兴、悲伤、愤怒、恐惧、厌恶、惊讶. 文献[8]对这6种基本表情做了比较全面的分析,本文据此进一步分析,总结出人脸基本表情主要表现在面部的眉毛、眼睛、下巴和唇部.

1.2.2 手语词根级表情

MPEG-4标准中定义的6种基本表情不能涵盖中国手语表达的面部动作. 本文对《中国手语》中的196个含有面部表情的手语词根进行研究和分析,首先按照《中国手语》对词根级表情的描述涉及的部位如唇部、眼睛、面部进行粗分类,然后基于同一部位的不同表情进一步细分类. 粗分类后的结果为:25个词的表情体现在唇部,28个词的表情体现在眼睛,143个词的表情体现在整个面部. 然后对上述粗分类结果进一步细分,唇部动作可分为16类,眼睛动作可分为7类,面部动作可进一步分为61类,合计84类. 84类词根级表情包含了与6种基本表情重复的部分,所以最终整理出78种词根级表情.

2 中国手语口型与表情库的构建

2.1 面部运动数据采集

本文采用基于MPEG-4的参数化的唇动和表情表示模型,基于Motion Analysis公司的Hawk数字动作捕捉系统完成数据的采集工作. Hawk数字动作捕捉系统包含22个Hawk数字捕捉镜头,采集到的数据是标记点的三维坐标,采样率为50帧/s. 采集过程主要包括:Hawk系统坐标系标定、面部贴点和特征点跟踪,其中标记点位置的选择是关键环节.

本文参考MPEG-4标准[9]定义的人脸定义参数(FDP)进行特征点标记,MPEG-4在中性人脸上

定义了 84 个特征点，即 FDP. 对于口型，MPEG-4 在口型外边缘定义了 10 个特征点，其中 8 个是会受 FAP 影响的，所以标记点选择这 8 个；对于面部表情，中国手语表达中的表情主要集中在眉毛、眼睛、唇、面颊，而在眼睛上贴点有一定难度，并且从其水平、垂直方向运动的角度讲，眼睛与眉毛几乎有相同的运动趋势，所以荧光点贴在眉毛、唇、面颊等主要部位受 FAP 影响的 19 个 FDP，其中唇部贴 8 个点，腮部 2 个，脸颊 2 个，左右眉毛 6 个，下巴 1 个.

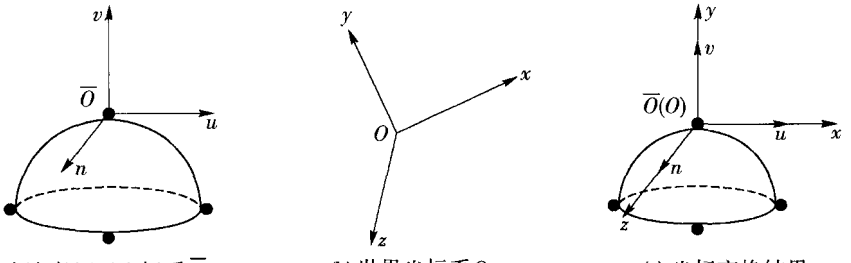
数据采集时，每个样本采集 3 遍，然后选择最优数据进行后续处理. 对于表情数据的采集，特别是手语词根级表情数据的采集，邀请了有中国手语专业背景、能熟练使用手语的聋校手语老师进行演示. 由于采集设备基于光学原理，而面部数据捕获时，脸上贴的荧光点都是直径为 3 mm 的小点，所以受光照条件、空气杂质和镜头稳定性的影响会出现误差和错误，因此需要对采集数据进行降噪、平滑等处理. 本文使用 EVART 软件应用插值方法，修正采集数据中存在的缺帧、变化突兀等问题.

2.2　人脸动画参数的提取

MPEG-4 虽然定义了 FDP 和 FAP，但没有规定 FAP 的提取方法. 对于 MPEG-4 定义的 2 个高级 FAP：视位（viseme）FAP 和表情（expression）FAP，是通过对预先定义的唇形和表情的线性组合得到的，本文试图通过其余的 66 个普通 FAP 定义的人脸一些小区域的运动和头部的转动来描述唇动和人脸表情的细致变化，所以本文只考虑其他 66 个 FAP. 本文从 FAP 定义出发，通过计算运动序列中人脸标记点的位移得到 FAP 的值. 由于采集者在表演面部表情动作的时候可能存在微小的摇头和摆动，而 FAP 参数是根据一组顶点的绝对位移来计算的，所以数据采集时头的旋转运动会使 FAP 参数的计算产生误差，因此在 FAP 提取前要进行去头部旋转处理，该方法计算复杂度低，效果可以接受.

1）去旋转

基本思想是以数据采集时相对位置不变的头部帽子上的 4 个点建立表演者帽子坐标系 $\overline{O}uvn$，本文以第 1 帧头部帽子上的 4 个点建立坐标系，其中头顶点设为坐标原点，如图 1（a）所示；计算世界坐标系 $Oxyz$（见图 1（b））到帽子坐标系 $\overline{O}uvn$ 的变换[10]，记为 $M_{xyz\to uvn}$，用于将后续其他帧的特征点坐标变换到 $\overline{O}uvn$ 坐标系，如图 1（c）所示. 设变换后的特征点 p' 坐标为 (x',y',z')，变换前对应点 p 坐标为 (x,y,z)，则 $(x',y',z',1)^{\mathrm{T}}=M_{xyz\to uvn}\cdot(x,y,z,1)^{\mathrm{T}}$，这里 $M_{xyz\to uvn}=R\cdot T(-\overline{O}_x,-\overline{O}_y,-\overline{O}_z)$，$(\overline{O}_x,\overline{O}_y,\overline{O}_z)$ 为 $\overline{O}uvn$ 坐标系的原点 $\overline{O}$ 在 $Oxyz$ 坐标系中的坐标，R 为从 $Oxyz$ 到 $\overline{O}uvn$ 的正交变换.



(a) 表演者帽子坐标系 $\overline{O}uvn$　(b) 世界坐标系 $Oxyz$　(c) 坐标变换结果

图 1　坐标系变换图

Fig. 1　Coordinate transformation

2）FAP 提取

66 个 FAP 中有 3 个用于控制头部的转动，通过去旋转时保留的坐标信息来计算，对于其他 63 个 FAP，其提取方法如下：首先标记模型的 FDP，手工标记 FDP 一方面不够准确，另一方面非常费时，Xface 是一个兼容于 MPEG-4 的基于关键帧的三维虚拟人动画创建的开源工具，具有友好的人机交互界面，可以帮助进行 FDP 的标记，因此本文基于 Xface 工具设定 FDP. 然后通过计算每帧相对第 1 帧各特征点的相对位移，结合 FAP 参数描述的定义得到 FAP 值，比如 FAP3 描述下颌的垂直偏移（OPEN_JAW 动作），对应本文中标记的下巴特征点向下方向的位移.

2.3　中国手语口型与表情库

本文基于 2.1～2.2 的方法对 1.1～1.2 中总结归纳的发音口型和表情基元数据进行数据采集和处理，得到描述唇动和表情的兼容于 MPEG-4 的 FAP 关键帧序列，从而建立了包含 148 个汉语音节的发

确定了中国手语的口型和表情基元.

2）建立了基于 MPEG-4 的参数化的唇动和表情表达模型，经数据采集、参数提取后，建立了中国手语的口型和表情库.

3）实现了基于中国手语口型与表情库的中国手语动画合成，取得了比较好的效果，并且该库中的口型数据和部分表情数据具有一定的通用性，可以推广到其他虚拟人动画应用中.

参考文献：

[1] CHEN Yi-qiang, GAO Wen, WANG Zhao-qi. Text to avatar in multimodal human computer interface [C] // Proceeding of the Fifth Asia Pacific Conference on Computer Human Interaction. Beijing: [s. n.], 2002: 636-643.

[2] SUMIHIRO K, TAKAO K. Facial and head movements of a sign interpreter and their application to Japanese sign animation [C] // Proceedings of the 9th International Conference on Computers Helping People with Special Needs (ICCHP 2004). Berlin: Springer-Verlag, 2004: 1172-1177.

[3] YE Ke-jia, YIN Bao-cai, WANG Li-chun. CSLML: a markup language for expressive chinese sign language synthesis[J]. The Journal of Computer Animation and Virtual Worlds, 2009, 20(2/3): 237-245.

[4] HARTMANN B, MANCINI M, PELACHAUD C. Implementing expressive gesture synthesis for embodied conversational agents [C] // Lecture Notes in Artificial Intelligence. Berlin: Springer-Verlag, 2006, 3881: 188-199.

[5] 吴铃. 汉语手语语法研究[J]. 中国特殊教育, 2005(8): 15-22.
WU Ling. Research on Chinese sign language grammar [J]. Chinese Journal of Special Education, 2005(8): 15-22. (in Chinese)

[6] 王志明，蔡莲红. 汉语语音视位的研究[J]. 应用声学, 2002, 21(3): 29-34.
WANG Zhi-ming, CAI Lian-hong. Study of Chinese viseme [J]. Applied Acoustics, 2002, 21(3): 29-34. (in Chinese)

[7] 中国聋人协会. 中国手语[M]. 北京：华夏出版社, 2003: 1-1176.

[8] ISO/IEC. 14496-2 Information technology-coding of audio-visual objects: part 2-visual [S]. Switzerland: International Organization for Standardization, 2001.

[9] ISO/IEC. JTCI/SC29/WG11 N2323 Overview of the MPEG-4 standard [S]. Geneva: International Organization for Standardization, 1998.

[10] 倪明田，吴良芝. 计算机图形学[M]. 北京：北京大学出版社, 2002: 143-144.

[11] 尹宝才，王恺，王立春. 基于 MPEG-4 的融合多元素的三维人脸动画合成方法[J]. 北京工业大学学报, 2011, 37(2): 266-271.
YIN Bao-cai, WANG Kai, WANG Li-chun. A synthesis method of three-dimensional facial animation with multiple elements blending based on MPEG-4 [J]. Journal of Beijing University of Technology, 2011, 37(2): 266-271. (in Chinese)

（责任编辑　梁　洁）