P.O.Box 8718, Beijing 100080, China	Journal of Software Feb. 2003,14(2):215-221
E-mail: jos@iscas.ac.cn	ISSN 1000-9825, CODEN RUXUEW, CN 11-2560/TP
http://www.jos.org.cn	Copyright © 2003 by The Editorial Department of Journal of Software

基于机器学习的语音驱动人脸动画方法

陈益强, 高 文, 王兆其, 姜大龙

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)
第一作者: 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.
联系人: 陈益强 Telephone: 86-10-82649008, Fax: 86-10-82649298, E-mail: yqchen@ict.ac.cn
Received 2001-06-04; Accepted 2001-08-01

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.

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.
http://www.jos.org.cn/1000-9825/14/215.htm

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,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.

基金项目：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)

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.