语义脸像的分析与生成

王晓慧　张申　贾珈　蔡莲红

普适计算教育部重点实验室

清华信息科学与技术国家实验室（筹）

清华大学计算机科学与技术系，北京，100084

摘　要：本文提出了一种基于语义维度的虚拟说话人脸像生成算法。针对虚拟说话人“语义脸像”生成，提出 7 个语义维度，对文本、脸像所表达的语义信息进行量化描述。建立虚拟脸像库，对虚拟脸像进行语义维度标注。基于对语义维度和脸像局部状态特征参数的相关性分析，采用基于决策树的 MRT 回归算法，建立了语义维度到局部脸像状态的参数预测模型，实现了语义驱动的参数化虚拟脸像生成算法。定义了语义维度的“脸像表达权重”和“典型脸像状态”，描述了不同语义维度对于脸像认知表达的重要程度。感知评测实验表明，本文提出的语义脸像生成算法，能够生成与情境语义相配合的虚拟脸像，从而为具有丰富语义表现力的虚拟说话人合成奠定了基础。

关键词：人脸；表情；语义；决策树

1．引言

脸像是人类非言语交流中最富表现力的沟通要素[1]。对于脸像的研究吸引了生物学、心理学、认知科学以及计算机科学等领域的众多研究者。Ekman等人建立的“人脸表情系统”认为脸像是情感的外部表现，能够表达 5-7 种基本情感[2]。Russell提出的“维度-情境”(dimensions-context)脸像认知模型，认为情境语义是影响脸像表达和认知的关键因素[3]。计算机科学对于脸像的研究多集中于人脸表情的识别、理解与生成[4]，对于脸像在表情功能之外的研究尚比较欠缺。

本文在情感脸像的研究基础上，通过引入言语语义，使得虚拟说话人脸像除了表情之外，同时具有丰富的语义表现力。本文将能够贴切表达言语语义的面部状态称之为“语义脸像”。围绕着虚拟说话人语义脸像生成，本文提出并着重解决以下三方面问题：1) 脸像辅助表达了言语中的哪些语义要素？2) 这些语义要素应如何量化描述？3) 如何利用语义信息指导脸像生成，建立语义脸像的可计算模型？首先，借鉴认知语言学的研究成果，本文认为脸像表达了言语中的“情态语义”(Modality)，即语言文字中所蕴含的代表说话人主观态度的语义成分[5]。在言语交流中，情态语义可以通过语调、语速、眼神、表情、肢体动作等多种手段表达，而本文所关注的正是其中的脸像表达。其次，借鉴心理学中的情感维度模型[6]和语言学中的语义特征分析[7]，提出描述脸像情态语义的 7 个维度，并对文本情境进行了语义维度标注，建立了语义的量化描述方法。再次，建立虚拟语义脸像库，进行 7 维语义参数标注，并同时提取脸像的局部状态参数。最后，采用基于决策树的回归算

资助项目：国家重点基础研究发展计划（973）（2006CB303101），国家自然科学基金（90820304）

联系作者：王晓慧，E-mail：xiaohui0506@gmail.com

法，建立语义维度到脸像局部状态的参数预测模型，实现了语义驱动的脸像生成；并定义语义维度的“脸像表达权重”和“典型脸像状态”，对语义维度与局部脸像状态的关系进行了分析。感知评测实验表明，采用本文提出的语义脸像生成算法，所生成的虚拟脸像能够较好地与情境语义相配合，为实现具有丰富语义表现力的虚拟说话人合成奠定了基础。

2．语义维度选取

在分析语义脸像之前，要明确脸像表达了哪些语义信息，如何对其进行量化。从语言学的角度来看，自然语言中往往蕴含了代表说话人主观态度的信息，包括信念、观点、情绪、态度、观察角度、意图等等，被称为“情态语义”。作为一种语义范畴，情态语义被定义为“说话者的说话方式，说话者用这种方式表达对交际中谈及的情况所持的态度”[5]。语言学家利用语义特征分析法得到情态语义的基本特征包括意志、态度、评价三个方面[7]。

本文所研究的“脸像”是指说话人在言语交流中表现出的面部状态。借鉴情感计算中的维度分析方法[8]，参考语言学家对情态（Modality）的语义特征分析，选取了 7 个维度对情态语义进行量化。“愉悦度”、“激活度”和“紧张度”描述心理状态，“优势度”和“关注度”刻画对他人和环境的态度，“确信度”描述观点，“力度”表达意志和影响力。

对中科院心理所设计的 80 个常见心理情境[8]进行以上 7 个语义维度标注。按照语义特征分析的二元偶分标注法[7]，在每个维度上采用+1，-1 二元值，分值为 0 代表情境没有体现出相应的语义维度，即不选择该语义维度进行标注。标注结果显示，平均每个情境采用 5 个语义维度进行描述，平均每个语义维度描述了 63 个情境，占情境总数的 78%。因此本文认为上述的 7 个语义维度，可以较为充分地描述说话人内心的情态信息。

本文提出了语义脸像分析生成的工作流程如图 1 所示。建立虚拟脸像库，得到脸像的语义维度标注和局部脸像状态参数，采用基于决策树的学习算法，训练得到“语义-脸像”预测模型。通过对言语情境标注语义参数，“语义-脸像”预测模型将给出脸像状态参数，进而生成与情境相适应的语义脸像。

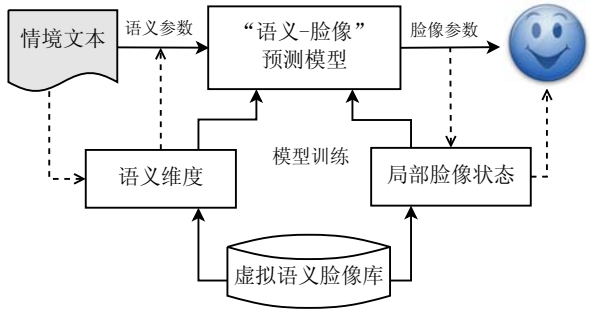


图 1　语义脸像分析与生成流程图

3．脸像语义分析

3.1 虚拟脸像库

为了建立脸像语义与人脸动作之间的映射模型，本文建立了一个虚拟脸像库。“虚拟”是指脸像是基于三维虚拟说话人模型[1]自动合成的，并非真实的人脸照片。在人脸表情合成研究工作的基础上[9]，提出了人脸局部状态参数PFP(Partial Facial Parameter)，如表 1 所示。每个参数的范围为[-1,1]，对应面部器官的连续变化过程。利用三维人脸表情编辑器[9]，可

[1] 该三维人脸模型由清华大学人机语音交互实验室（HCSI）同香港中文大学人机通讯实验室（HCCL）共同开发.

以将人脸局部状态参数(PFP)转换为人脸动画参数(FAP)，驱动虚拟说话人表现脸像动作。

表 1　局部脸像状态参数定义

局部脸像状态描述	参数编号	中性样例(=0)	典型样例(PFP=+1)	典型样例(PFP=−1)
皱眉	PFP1			
八字眉	PFP2			
扬眉	PFP3			
眼神（左右看）	PFP4			
眼神（上下看）	PFP5			
嘴张开/闭合	PFP6			
嘴角上翘/下弯	PFP7			
嘴角伸展/收缩	PFP8			

设置每个局部状态参数在[-1，+1]的区间内连续变化，得到面部器官典型状态的组合参数，在三维虚拟说话人模型上做出相应动作。由于参数的随机组合，导致生成一些奇怪脸像，即普通人一般不会或难以做出的动作，如对眼等等。因此对生成的虚拟脸像进行人工校准，删除奇怪脸像，最终选定 330 个脸像组成该脸像库，基本覆盖了人脸常见的脸像。

3.2 语义标注

如何描述特定脸像状态所表达传递的语义信息，是语义脸像分析首要解决的问题。维度观的描述方法已经在情感计算等研究领域得到了应用，并取得了较好的结果[8,10]。本文利用所提出的 7 个语义维度来对脸像进行标注量化。

实验邀请了三名大学生对 330 幅脸像样本进行 7 个语义维度上的评价打分。由于采用了二元偶分的标注方法，对于标注结果中的分歧，采用投票方式最终确定。对标注结果进行统计，平均每个脸像选用 4 个语义维度进行描述，平均每个语义维度描述了 224 幅图像，占样本总数的 67.8%。以上结果表明，该脸像库不仅描述了常见的脸像动作，而且在 7 个语义维度上都有充分的表达，为语义脸像的分析与生成提供了数据支持。

3.3 相关性分析

为了研究局部脸像状态与语义维度的关系，采用统计方法对脸像动作参数和语义参数进行量化分析。首先对局部脸像状态参数向量进行相关性分析，相关系数矩阵如图 2(a)显示，平均相关系数为 0.156，这是因为局部脸像状态参数描述了各个面部器官的局部运动，相互之间不存在明显的依赖关系。图 2(b)显示了脸像语义参数向量的相关系数矩阵，其平均相关系数为 0.364。在脸像语义认知过程中，同一个脸像可能会在多个语义维度上都有所表达，然而各个维度又分别描述了不同的语义信息，因此各语义维度之间存在一定的关联，但相关性不高，从而可以较为全面准确的描述脸像所表达的语义信息。

同时采用 CCA 典型相关分析（Canonical correlation analysis）对局部脸像状态特征与语义维度特征之间的关联性进行分析。计算结果显示，脸像语义参数与局部状态参数之间存在着较强的关联性，其 CCA 相关系数为 0.875。分析结果表明，7 个基本语义维度可以有效的对脸

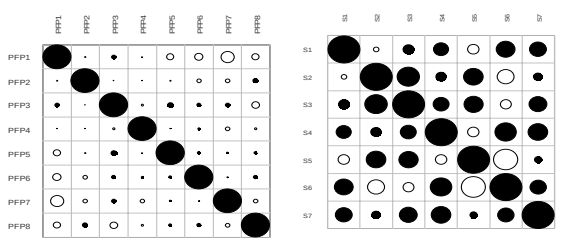


图 2 (a) 脸像参数相关矩阵　(b) 语义参数相关矩阵
(圆面积代表相关系数绝对值，空心圆代表负相关[11])

像动作状态表达的语义信息进行描述。

4．语义脸像生成

基于 3.3 节中对脸像与语义表达关联性的分析，建立语义参数到脸像状态参数的预测模型，实现语义参数驱动的虚拟脸像生成。通过情境的语义标注，生成与其相贴切的脸像。

4.1 多元回归树 MRT 算法

语义参数驱动的脸像生成，实际上是建立多维语义到面部状态参数的映射模型。考虑到目前的语义标注采用二元离散值，对应的脸像状态参数则在连续的实数空间内取值。因此将预测模型的建立转换为多元回归问题，采用基于 CART 的非参数回归算法，建立一棵根据语义参数预测面部动作的二叉树。回归树的每一个节点都对应一个语义维度的二值问题，所有脸像样本将根据其语义参数取值进入回归树的不同分支。

本文采用了基于CART的多元回归扩展算法MRT(multivariate regression tree)[12]，实现对脸像动作特征(PFP)多维变量的回归。每一个叶子节点对应了MRT算法的回归结果，该节点下所有脸像状态参数(PFP)的均值向量作为该节点的输出预测值。MRT算法通过在语义-脸像数据集上训练，学习得到树形结构的映射函数，能够在未见的语义特征上对动作参数进行准确的预测，其树形结构能够对语义与动作之间的映射关系给出明确的解释。

4.2 “语义-脸像”预测模型

采用De’ath提供的MRT算法建立“语义-动作”的参数预测模型[12]。330 幅虚拟脸像样本的语义参数作为MRT根结点的输入，对应的局部脸像参数状态作为预测输出。训练过程中采用交叉验证对MRT树进行剪枝(k=10)。根据Breiman提出的 1-SE规则[13]，选择在交叉验证最优结果的单位标准差范围内规模最小的树作为最终预测模型。

根据 3.3 小节中对 PFP 参数维度间的相关性分析，实验中为每个局部脸像状态分别建立 MRT 回归树。统计结果显示，MRT 树的平均规模为 31 个叶子节点，每个叶子节点内的样本个数约为 15 个，树的平均深度约为 9 层，平均交叉验证误差为 0.2892。

为了分析语义维度在脸像认知与表达中的作用，实验中对各语义维度在回归树建立过程中的出现频率和平均层级进行了统计，定义了语义维度的“脸像表达权重”。权重由该语义维度在MRT树中的出现频率和平均层级加权平均得到；权重越大，表明该语义维度在局部脸像所表达的语义中占比重越大。对语义维度S_i，脸像表达权重系数SW_i定义如式 1，

$$SW_i = \frac{LW_i + FW_i}{2} \quad \text{其中} \quad LW_i = 1 - \frac{\sum_{j=1}^{N_i} \text{level}(S_i^j)}{N_i \bullet \text{level}(tree)} \quad FW_i = 1 - \frac{N_i}{size(tree)} \qquad (1)$$

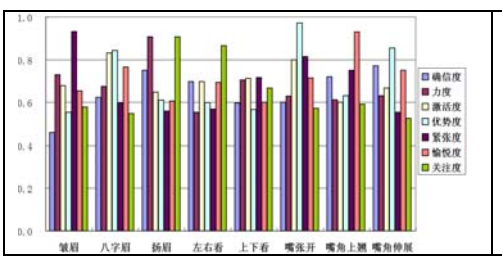
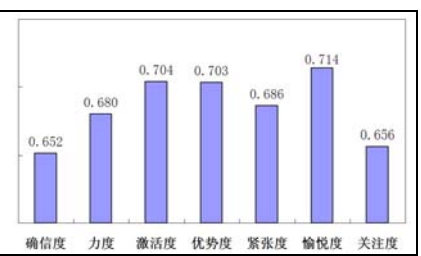


图 3(a) 语义维度在局部脸像中的表达权重



(b) 语义维度的平均脸像表达权重

其中N_i是MRT树中以S_i为分裂标准的节点总数，level(S_i^j)是S_i所在的第j个节点在MRT树中的层级值

（根结点值为 0），level(tree)和size(tree)分别是MRT树的最大深度和叶子节点总数。7 个语义维度在不同局部脸像状态中的表达权重和平均权重，如图 3 所示。MRT树根结点所对应的语义维度，是对特定脸像状态最具区分能力的语义特征，其所对应的局部脸像状态也最能够表达该特征语义信息，本文称之为语义维度的“典型脸像状态”，如表 2 所示。由于力度（S2）和激活度（S3）往往通过动作的幅度体现，在各种局部脸像状态中都会有所表达，因此没有对应的典型脸像状态。综合对语义维度的表达权重和典型脸像状态的分析，可以得出，各种局部脸像状态分别传递了特定的语义信息。

表 2　语义维度对应典型脸像状态

语义维度	确信度	力度	激活度	优势度	紧张度	愉悦度	关注度
典型脸像状态	嘴角伸展	—	—	八字眉、张嘴	皱眉	嘴角上翘	扬眉、眼神

4.3 “情境-脸像”主观评测实验

本文设计了“情境－脸像”感知评测实验来评价语义脸像生成算法的有效性。从中科院心理所提供的 80 个情境文本中选择 10 个测试样本，邀请 13 名被试者参加感知评测。图 4 给出了评测样例，被试者对脸像与情境的匹配程度进行打分，包括“生动”(5 分)，“自然”(4 分)，“一致”(3 分)，“不当”(2 分)，“奇怪”(1 分)。实验结果表明，主观评测的平均打分（MOS）为 3.45(标准差为 0.41)，即语义脸像生成算法能够生成与情境语义相适应的脸像，但自然度和表现力还需提高。

情境 1：你打开房门，发现许多多年未见的亲戚朋友，他们大喊道：“生日快乐!”。

情境 2：你刚接到一个邮件，通知你要缴纳前 3 年的个人所得税。

图 4　“情境-脸像”感知评测实验样例

5．结论与展望

本文在情感脸像的研究基础上，提出了“语义脸像”的概念，并提出了 7 个基本语义维度，对文本、脸像中所蕴含或表达的语义信息进行量化描述。建立虚拟语义脸像数据库，标注了脸像的语义参数，在对语义参数和局部脸像状态参数相关性分析的基础上，采用基于决策树的回归算法，建立了语义参数到脸像参数的映射模型，实现了语义驱动的参数化脸像生成算法。感知评测实验表明，本文提出的语义脸像生成算法，能够生成与情境语义相配合的脸像动作，但在自然度和表现力方面还有所欠缺。本文的研究成果为具有丰富语义表达能力的虚拟说话人合成奠定了基础，进一步的工作将对语义脸像生成算法进行改进完善，并将其应用在虚拟说话人合成系统中，为人机语音交互提供一种智能化的交互界面。

参考文献

[1]　A. Mehrabian, Nonverbal communication. Aldine-Atherton, Chicago, Illinois, 1972.

[2]　P. Ekman et al. Emotion in human face. New York：Pergamon, 1972

[3]　J. A. Russell, J. M. Fernández Dols. The psychology of facial expression. Cambridge University Press, 1997

[4] A. K. Jain, S. Z. Li. Handbook of Face Recognition. Springer-Verlag New York, Inc., Secaucus, NJ, 2005

[5] 彭利贞． 现代汉语情态研究． 复旦大学中国语文学系，博士论文，2005.4

[6] A. Mehrabian, Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. Current Psychology：Developmental, Learning, Personality, Social, 1996. 14: 261-292

[7] 俞咏梅． 《汉语语义学纲要》． 东北师范大学出版社，2006.12

[8] X.M.Li, H.T. Zhou, S.Z.Song, T. Ran, X.L.Fu. The Reliability and Validity of the Chinese Version of Abbreviated PAD Emotion Scales. Int. Conf. on Affective Computing and Intelligent Interaction, 513-518, 2005

[9] 张申，吴志勇，蔡莲红． 基于局部表情参数化的三维表情脸像合成． 第二届全国人机交互学术会议．574~581. 2006.

[10] 崔丹丹． 情感语音分析与变换的研究． 清华大学工学博士学位论文，2007

[11] 魏太云．相关矩阵的可视化及其新方法探究．http://cos.name/2009/03/correlation-matrix-visualization/

[12] De'Ath G.. Multivariate regression trees: a new technique for modeling species environment relationships. Ecology, 83:1105-1117, 2002

[13] Breiman, L., J. H. Friedman, R. A. Olshen, and C. G. Stone. Classification and regression trees. Wadsworth International Group, Belmont, California, USA, 1984

Semantic Facial Expression Analysis and Synthesis

Xiaohui Wang[+], Shen Zhang, Jia Jia, Lianhong Cai

Key Laboratory of Pervasive Computing, Ministry of Education

Tsinghua National Laboratory for Information Science and Technology

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

[+] Author: Tel: +86-13488797964, E-mail: xiaohui0506@gmail.com

Keyword: Facial Expression, Semantic, CART

Abstract: In this paper, a semantic-based facial expression synthesis method is proposed. Based on the previous work on emotional facial expression, we use the term of “semantic face” to indicate those facial expressions that convey the modularity information in face-to-face communications. An virtual facial expression database is established, which is annotated using 7 basic semantic-dimensions. Based on the correlation analysis between the semantics and partial facial movements, a CART-based regression algorithm is adopted to build the semantic-expression predict modal, which is used to synthesize the semantic facial expression. For each semantic-dimension, an express weight and typical partial facial state is defined, both of which can be used to measure the importance of semantic dimension in facial expression. The perceptual evaluation shows the proposed algorithm can synthesize proper facial expression for a given context, which enable us to build a virtual talking avatar with rich semantic expressivity.