中国科学技术大学计算机学院
《计算机图形学理论和应用》调研报告

2021.05.26

调研题目：人脸动画及表情合成技术发展综述

学生姓名：胡毅翔

学生学号：PB18000290

计算机实验教学中心制

2019 年 9 月

目录

1 引言 **3**

2 基于几何模型 **3**

2.1 **整体法** 3

2.2 **分治法** 4

3 基于二维图像 **5**

3.1 **表情映射法** 5

3.2 **表情流法** 5

3.3 **替换法** 5

4 基于深度神经网络 **6**

4.1 FLAME **模型** 6

4.2 **音频驱动** 7

4.3 **文本驱动** 7

5 总结 **7**

1 引言

具有真实感的人脸脸表情动画合成一直是计算机图形学和计算机视觉领域的研究热点和难点之一，且被广泛应用于数字娱乐、视频会议、医疗、辅助教育等领域。在人脸动画领域，高端方法和低端方法之间存在很大的差距。高端方法生成的面部动画令真正的人类难以区分，但是需要人脸建模的专家付出大量的体力劳动。而低端方法，依托传感器的面部捕捉等，得到人脸动画参数，生成人类动画，又不足以表现人脸自然变化中的细节。为此，研究者们不断钻研，在鲁棒性，性能，易用性，实时性等方面得到了优化。

目前，人脸表情合成方法主要分为三大类：基于几何模型（三维网格模型）的人脸表情合成，基于二维图像的人脸表情合成和基于深度神经网络的人脸表情合成。随着应用场景不断拓宽，人脸表情动画的驱动方式也不断增加，从最初的参数驱动，到图像数据驱动，再到音频，文本驱动。

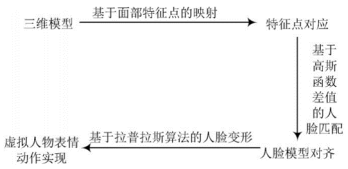
2 基于几何模型

基于三维网格的方法是通过跟踪一系列人脸表情变化的曲线和基于全局统计的模型生成所需的目标图像。

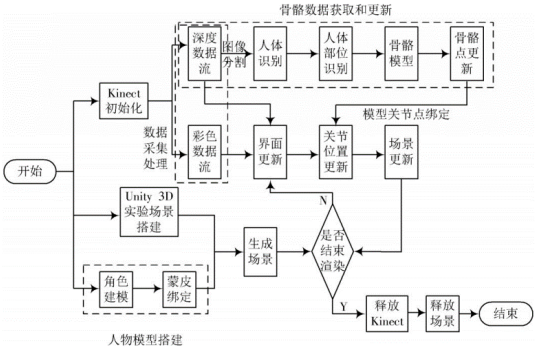
2.1 整体法

基于这种传统方法，有以下这些主要成果：Pighin 等 [1] 提出了具有逼真纹理的三维人脸建模的表情合成系统。Wang 等 [2] 构建了一个 MPEG-4 兼容的人脸动画系统，使用脸部定义参数（facial definition parameter, FDP）来构建人脸模型，通过脸部动画参数（facial animation parameter, FAP）来驱动该模型生成各种表情。Blanz 等 [3] 提出了从单张图片或视频中恢复人脸的三维模型，尝试从带纹理的三维模型中进行表情合成，但创建一个好的模型相当困难，因为必须对面部的所有细节进行建模，如眼睛、头发、牙齿等。Vlasic 等 [4] 提出的多线性模型（Multilinear Models）将分离参数化不同的属性（如顶点信息、形状、视位、表情等）建立在同一个数据张量空间中，这些属性之间用笛卡尔积来构造并且相互独立。将这些独立属性参数进行任意组合，最终得到不同的人脸表情。Lv 等 [5] 提出一种面向同一人脸表情转移的方法，即对目标人脸进行三维建模，生成特定的混合形状（blendshape）模型，利用该模型生成与输入人脸图像匹配的三维人脸模型，并对图像进行扭曲融合，生成所需的目标人脸图像。

Xiong 等 [6] 基于几何模型构建了一个虚拟人物表情动作系统，利用系统硬件模块完成图像数据及人体骨骼数据的采集、处理，以及人体骨骼点定位、关节点与 Kinect 骨骼点映射关系构建、虚拟人物模型的搭建。在此基础上，通过面部特征点映射完成特征点的对应，并通过对齐视频人脸与三维动画虚拟人脸，完成人脸模型对齐；采用拉普拉斯坐标恢复模型重建人脸表情动作，完成三维动画虚拟人物表情动作的模拟，实现基于三维动画的虚拟人物表情动作系统设计。



(a) 1.1



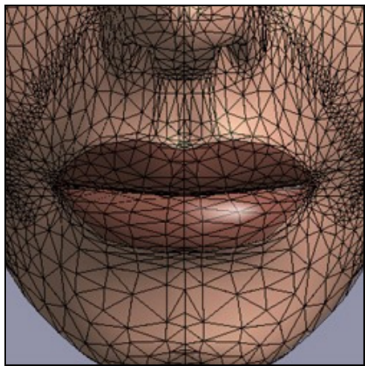
(b) 1.2

图 1: 虚拟人物表情动作系统

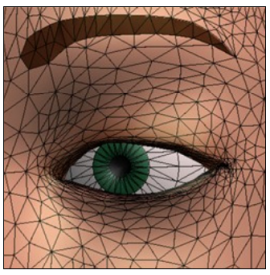
2.2 分治法

尽管上述整体法较为稳定，且便于计算，但这些方法通常不允许细粒度变形，并且在困难区域（嘴巴，眼睛）难以生成逼真的效果。而人们往往是通过这些五官特征来区分人脸的。为了解决五官特征不明显，不真实的问题，人们又基于分而治之的思想，提出了对五官，皱纹等细节部位先进行单独处理，再通过一定的约束合成人脸模型，以提高人脸的真实感。

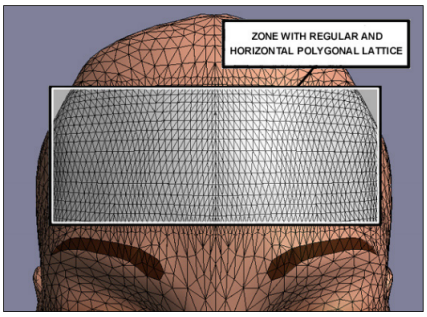
Pasquariello 等 [7] 在三维人脸模型中加入皱纹等细节因素，将人脸模型进行网格化，并按照人的生理结构将人脸分成嘴、眼睛、眉毛、额头等区域，使得每一区域的网格数量和拓扑等都不同，进而实现表情的模拟。



(a) 2.1



(b) 2.2



(c) 2.3

图 2: 分治法

Zhang 等 [8] 在三维形变模型中也加入了皱纹等细节，将其划分为 14 个子区域，避免了表情皱纹超过分区的边界。Joshi 等 [9] 提出的基于物理的分割方法可以自动将人脸分割成多个区域，每一个区域都表示成混合形状的线性组合，从混合形状中学习约束条件和参数。Park 等 [10] 将给定的每一个源关键模型分成 3 个子区域，每一个子区域都包含人脸的关键特征，实现了面部表情的合成。Joshi[11] 把每一个表情看成其他表情的线性组合，通过改变这些线性组合的权重来合成比较完整的面部表情。Garrido[12] 提出了一种基于图像的人脸视频再现方法，对于输入的两个不同人脸的面部表情视频，将源序列的表情

传递给目标序列，同时尽可能地保留灯光、背景等因素。该方法的缺点是：需要一些特定的人脸数据库，且依赖于输入图像的一些表情信息，如是否是中性人脸或者有无标记点等等。文献 [13, 14] 首先对人脸进行建模，然后得到特定用户的混合形状模型，进而求解出混合形状系数，最后通过对系数的改变来合成目标表情。Huang 等 [15] 提出了基于非联合学习的人脸表情合成方法：通过一种无监督回归的算法，将具有相同属性的三维人脸模型映射到同一个低维空间，对其进行重建，实现人脸表情的合成。

3 基于二维图像

基于二维图像的方法是用已知的表情数据来合成新的表情，或者直接将已有的表情传递到目标图像上。

3.1 表情映射法

Williams[16] 提出通过表情映射的方法来合成新的目标表情，首先提取两幅图像不同的面部特征，然后计算特征之间的矢量差，最后利用特征向量来进行图像的扭曲。该方法的鲁棒性不足，虽然实现了不同表情之间的转换，但是不适用于戴眼镜和头部位姿有较大变化的情况。

3.2 表情流法

Yang 等 [17] 提出了基于表情流的方法：首先提取两幅图像的特征点并分别进行三维人脸重建；然后计算两个三维模型之间的差异，将差异映射到二维图像上得到表情流，再利用所得表情流进行图像扭曲；最后进行图像融合 [18]。

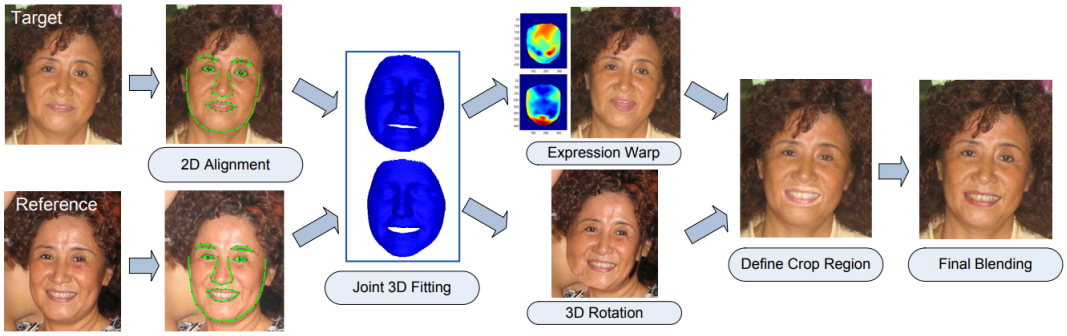


图 3: 表情流法

3.3 替换法

表情流的计算方法往往比较复杂，鲁棒性很差，得到的效果不是很逼真，而且只能在同一人脸之间进行，不具有普遍性。文献 [19, 20] 提出了一种人脸图像自动替换系统：首先对输入图像进行人脸检测；然后将提取的每个人脸进行对齐，并从大量的人脸数据库中找到与其相近的人脸；最后通过图像融合实

现目标输入图像的表情合成。这些方法可以有效、快捷地合成人脸表情，但由于人脸具有特异性，且表情复杂、丰富，在表情合成过程中合成具有真实感的表情比较困难。

图 4: 替换法

4 基于深度神经网络

基于深度神经网络的方法是在上述两种方法的基础上，结合深度神经网络，提供一种端到端的人脸表情生成方法。这些方法相较传统方法在实时性，鲁棒性等方面有所提高。

4.1 FLAME 模型

Li 等 [21] 提出了 FLAME（Faces Learned with an Articulated Model and Expressions）模型，通过从数千个精确对齐的 3D 扫描中学习人脸模型。该模型以下巴，脖子，眼球，姿态相关校正等线性组合而成，模块化处理，可以对各个部分进行精细化处理，并且最终得到的是低维表示。

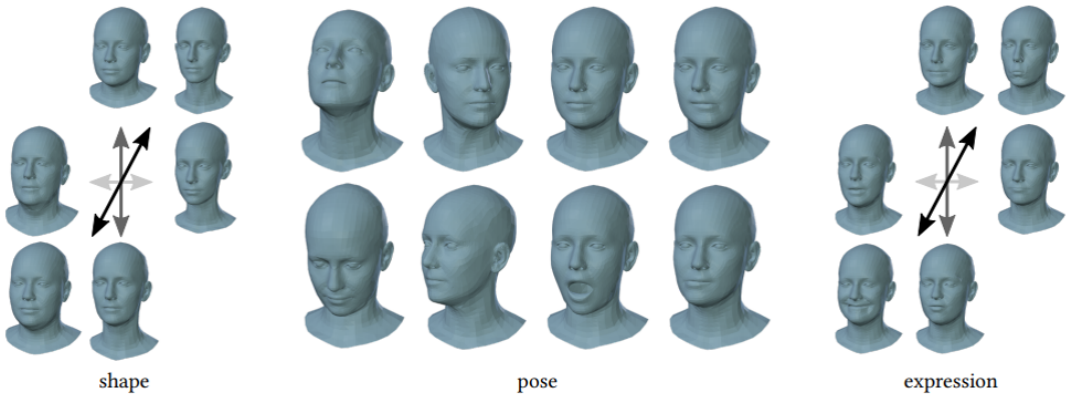


图 5: FLAME 模型

Sanyal 等 [22] 在 FLAME 模型的基础上假设脸部形状是恒定的，与表情，状态，光线等无关，实现了从单张 2D 图像得到 FLAME 模型的神经网络。这种方法进一步摆脱了对 3D 扫描的依赖，是一种无 3D 监督的学习，使得模型构建更为简便，还维持了很好的鲁棒性。

Paier 等 [23] 提出了一种混合动画空间，该框架利用深度学习来提供一个交互式动画引擎，可以提供简单直观的可视化将其用于面部表情编辑。此外，该框架还训练了这个变分式自动编码器，以学习用于交互式面部动画的人脸表情的低维潜在空间。

4.2 音频驱动

Hai[24] 等提供了一个用原始波形进行实时语音驱动生成 3D 人脸动画的深度学习框架。该网络直接映射输入语音序列到一系列面部动作单元激活和头部旋转用以驱动 3D 混合形状人脸模型。该模型还能学习语音的上下文中的潜在情感变化，使得人脸动画的情绪强度与语音的情绪强度相当。

Cudeiro[25] 等则提供了一个鲁棒性更强的深度神经网络。该网络同样由音频驱动，输入为任何语言和一个人脸模型，而输出为一段连续的人脸动画。

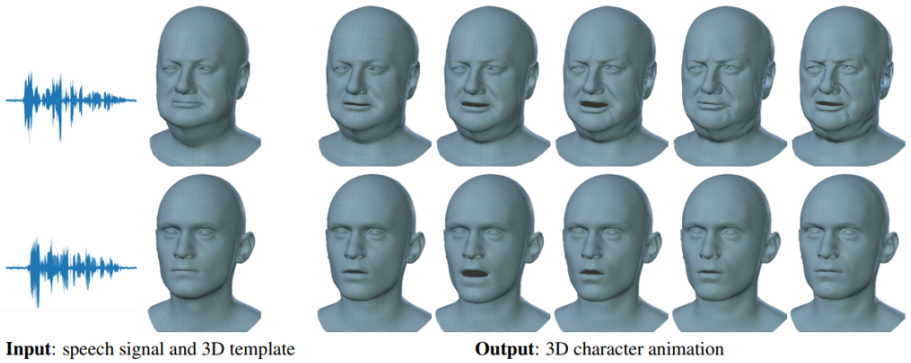


图 6: 音频驱动

4.3 文本驱动

给定文本作为输入时，Yu 等 [26] 提供了采用深度学习方法合成高真实感、认以身份以及唇音同步的人类动画。

5 总结

人脸动画表情技术的发展经历了从早期的几何模型生成，到基于二维图像生成，再到混合使用几何模型及二维图像，结合音频，文本输入，构建端到端神经网络。人脸的真实感不断增强，计算效率不断提高，算法的鲁棒性，实时性，可用性不断增强。随着技术的不断完善，人们对数字人的接受程度也越来越高，人脸动画在 VR，AR，游戏，CG 电影等领域得到了广泛应用。

在未来的研究过程中，基于深度神经网络的生成方法将成为主流，设计更为简单易用的，对计算性能需求更低的算法，使得个人用户可以在 PC 甚至移动设备端实现人脸动画的生成。同时，对于细节（眼部，脸部，鼻子等）的处理也更加真实。

此外，目前的研究中生成的人脸模型，以单纯的表面效果为主。未来的研究中，为使人脸更加真实，对口腔，鼻腔等原本被忽略的可视脸部器官模型也需要进一步完善。

参考文献

[1] Frédéric Pighin, Jamie Hecker, Dani Lischinski, Richard Szeliski, and David H. Salesin. 1998. Synthesizing realistic facial expressions from photographs. In Proceedings of the 25th annual conference on Computer graphics and interactive techniques (SIGGRAPH '98). Association for Computing Machinery, New York, NY, USA, 75–84.

[2] 王奎武，王洵，董兰芳，陈意云. 一个 MPEG-4 兼容的人脸动画系统 [J]. 计算机研究与发展,2001(05):529-535.

[3] Blanz, V. and Basso, C. and Poggio, T. and Vetter, T.Reanimating Faces in Images and Video.Computer Graphics Forum

[4] Daniel Vlasic, Matthew Brand, Hanspeter Pfister, and Jovan Popović. 2005. Face transfer with multilinear models. ACM Trans. Graph. 24, 3 (July 2005), 426–433.

[5] LV P, XU M L. Expression of Face Expressions Unrelated to Expression Database [J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(1).

[6] 熊伟. 基于三维动画的虚拟人物表情动作系统设计 [J]. 现代电子技术,2020,43(20):97-101.

[7] PASQUARIELLO S, PELACHAUD C. GRETA: A Simple Facial Animation Engine[M]// Soft Computing and Industry. London ; Springer, 2002 .

[8] ZHANG Q,LIU Z,GUO B, et al. Geometry-Driven Photorealistic Facial Expression Synthesis [J]. IEEE Transactions on Visualization & Computer Graphics, 2005, 12(1) : 48 − 60.

[9] JOSHI P, TIEN W C,DESBRUN M, et al. Learning controls for blendshape based realistic facial animation [C]// Proceedings of ACM SIGGRAPH Eurographics Symposium on Computer Aimation. 2003 : 187 − 192.

[10] PARK B, CHUNG H, NISHITA T, et al. A feature-based approach to facial expression cloning: Virtual Humans and Social Agents [J]. Computer Animation and Virtual Worlds, 2005 , 16(3/4) : 291 − 303.

[11] JOSHI P, TIEN W C, DESBRUN M, et al. Learing Cnotrols for Blend Shape Based Realistic Facial Animation [C]// ACM Transactions on Graphics. 2006.426 − 433. GARRIDO P, VALGAERTS L, REHMSEN O, et al. Automatic face reenactment

[12] Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014:4217-4224.

[13] WESIE T, BOUAZIZ S,LI H, et al. Realtime performance-based facial animation [J]. ACM Transactions on Graphics, 2011 , 30(4) : 1.

[14] CAO C, WENG Y, LIN S, et al. 3D shape regression for real- time facial animation[J]. ACM Transactions on Graphics, 2013 , 32(4) : 1.

[15] HUANG X Q, LIN Y X, SONG M L. Three-dimensional facial expression synthesis method based on nonlinear joint learning Journal of Computer-Aided Design &. Computer Graphics, 2011, 23(2).

[16] WILLIAMS L. Performance-driven facial animation [C]//ACM SIGGRAPH Computer Graphics. 1990 : 235 − 242.

[17] YANG F, WANG J, SHECHTMANE, et al. Expression flow for 3D-aware face component transfer [J]. ACM Transactions on Graphics, 2011, 30(4) : 1.

[18] PEREZ P,GANGNET M,BLAKE A. Poisson image editing [J]. ACM Transactions on Graphics, 2003, 22(3) : 313 − 318.

[19] BITOUK D. Face Swapping : Automatically Replacing Faces in Photographs [J]. ACM SIGGRAPH, 2008, 27(3) : 1 − 8.

[20] DALE K, SUNKAVALLI K, JOHNSON M K, et al. Video face replacement [J]. ACM Transactions on Graphics, 2011, 30(6) : 1.

[21] Tianye Li,Timo Bolkart,Michael J. Black,Hao Li,Javier Romero. Learning a model of facial shape and expression from 4D scans[J]. ACM Transactions on Graphics (TOG),2017,36(6).

[22] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, Michael J. Black. Learning to Regress 3D Face Shape and Expression From an Image Without 3D Supervision. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pages 7763-7772, Computer Vision Foundation / IEEE, 2019.

[23] Paier Wolfgang,Hilsmann Anna,Eisert Peter. Interactive facial animation with deep neural networks[J]. IET Computer Vision,2020,14(6).

[24] Hai Xuan Pham,Yuting Wang,Vladimir Pavlovic. End-to-end Learning for 3D Facial Animation from Speech[P]. Multimodal Interaction,2018.

[25] Cudeiro D, Bolkart T, Laidlaw C, et al. Capture, learning, and synthesis of 3D speaking styles[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 10101-10111.

[26] 于灵云. 基于文本/语音驱动的高自然度人脸动画生成 [D]. 中国科学技术大学,2020.