



AI驱动的面部机器人表情交互综述

生钦丞, 唐威*, 秦浩, 孔昱捷, 戴浩凯, 仲一丁, 王咏浩, 邹俊*, 杨华勇

浙江大学机械工程学院, 流体动力基础件与机电系统全国重点实验室, 杭州 310058

* E-mail: weitang@zju.edu.cn; junzou@zju.edu.cn

收稿日期: 2025-08-11; 接受日期: 2025-09-29; 网络版发表日期: 2025-10-15

国家自然科学基金(批准号: 52305074)、浙江省自然科学基金(编号: LZYQ25E050001)和国家自然科学基金国际合作与交流项目(编号: 52261135542)资助

摘要 随着人工智能(artificial intelligence, AI)将人形机器人的应用扩展到教育、医疗和家居等社会领域, 人们对其情感交互能力的需求日益增长. 面部表情作为情感交流中最关键的渠道之一, 承载了约55%的情感信息. 因此, 具备逼真面部表情能力的人形机器人, 即“面部机器人”的重要性日益凸显. 本综述系统回顾了AI驱动的面部机器人表情交互技术. 首先, 概述了面部机器人的硬件架构, 为理解其表情交互能力奠定基础. 随后, 基于“感知-推理-生成”框架, 对传统方法与前沿技术进行了比较分析. 传统方法主要依赖视觉与语音信号进行情感识别, 并基于离散或维度情感模型开展情感计算, 从而驱动面部运动、眼神交流与唇语同步等面部表情生成; 而先进方法则融合多模态感知, 结合大语言模型或多模态大语言模型进行情感推理, 并引入智能体的规划能力、记忆机制和工具使用, 从而显著提升机器人在适应性、真实感与情商方面的表现. 最后, 探讨了面部机器人的潜在应用场景、面临的技术挑战及未来研究方向. 本文通过整合硬件设计、算法机制、实际应用以及亟待解决的问题, 为构建具备同理心与社会适应能力的面部机器人提供了系统性的研究基础.

关键词 面部机器人, 人工智能, 人机交互, 大语言模型, 表情生成

1 引言

AI技术的快速发展加速了人形机器人的演进, 使其应用场景从传统工业领域拓展至教育、医疗、陪伴与客户服务等多样化的社会与家庭环境^[1~3]. 随着应用场景的不断丰富, 用户对机器人进行细腻和具有情感智能的人机交互的需求日益增强. 仅依赖语音进行交流已难以满足用户期望, 当今的机器人应能够通过具有社会意义的视觉表达行为传递情感, 展现对人类情

绪的真实理解与同理心^[4,5]. 在非语言情感交流中, 面部表情起着核心作用. 研究表明, 人类情绪信息的传达中, 约55%来自面部表情与姿态, 38%源于语音语调, 而真正由语言内容传达的仅占7%^[6]. 这凸显了面部表达在传递同理心、建立情感联系与增强信任感等社会性互动中的关键作用. 为满足这一需求, 面部机器人应运而生——它们通常为配备逼真可动面部的人形机器人, 旨在提升与人类的情感交互能力, 弥合人机之间的情感鸿沟.

引用格式: 生钦丞, 唐威, 秦浩, 等. AI驱动的面部机器人表情交互综述. 中国科学: 技术科学, 2025, 55: 1603–1637

Sheng Q C, Tang W, Qin H, et al. A review on AI-driven face robot for human-robot expression interaction (in Chinese). Sci Sin Tech, 2025, 55: 1603–1637, doi: [10.1360/SST-2025-0217](https://doi.org/10.1360/SST-2025-0217)

近年来,面部机器人技术在硬件平台与交互方法上均取得了显著进展。在硬件方面,多种驱动机制被用于实现逼真的面部表情。例如, SAYA等机器人采用气动驱动装置,通过人工肌肉实现表情变化^[7]; Sophia^[8]和 Ameca^[9]等主流系统则普遍采用高精度伺服电机与机械连杆结构实现复杂表情;此外,形状记忆合金(shape memory alloys, SMA)^[10]与电介质弹性执行器(dielectric elastomer actuators, DEA)^[11]等柔性驱动技术也展现出更自然、柔和的面部运动潜力。在交互方面,传统面部机器人主要遵循“感知-处理-生成”流程:通过识别人类的面部表情、眼神或语音等情绪线索^[12],借助情感建模方法进行内部情绪表示^[13,14],并结合语境或语言生成相应表情^[15,16]。已有综述对面部机器人的感知模态与机械结构进行了深入评述^[17,18],但关于表情交互流程,特别是基于新兴人工智能技术的系统性探讨仍属空白。当前,大语言模型(large language model, LLM)及其多模态扩展多模态大语言模型(multimodal large language model, MLLM)的能力不断提升,正在重塑机器人情绪处理机制。这些前沿技术带来了诸多变革:多模态情感识别通过融合视觉、听觉、文本与生理数据,实现更丰富、稳健的情绪感知; LLM

则支持更具上下文理解能力的深层情感推理^[19,20];基于智能体的框架结合了计划、记忆与工具使用能力,进一步增强了面部表达的意图性与适应性。鉴于人工智能的飞速发展,系统梳理这些技术如何推动人机面部表情交互的演进具有重要的理论与应用意义。

本综述旨在系统探讨面部机器人技术与表情交互方法的最新进展,尤其聚焦于人工智能驱动的关键技术路径。文章首先介绍面部机器人的硬件架构,包括驱动方式、机械设计、系统结构与仿生皮肤材料;随后,基于“感知-推理-生成”框架,分别回顾传统方法与前沿方法;第3节分析人类情感识别、情绪模型驱动的表情生成与表情合成技术;第4节探讨多模态感知、LLM驱动的情感推理,以及具备规划、记忆与工具使用能力的智能体式表情生成方法;最后,文章展望面部机器人在教育、医疗、陪伴与服务等场景中的应用前景,并提出当前面临的关键挑战与未来研究方向。图1展示了AI驱动的人机面部表情交互的技术发展路径。

2 面部机器人平台设计与开发

本节旨在为AI驱动的面部机器人控制建立物理实

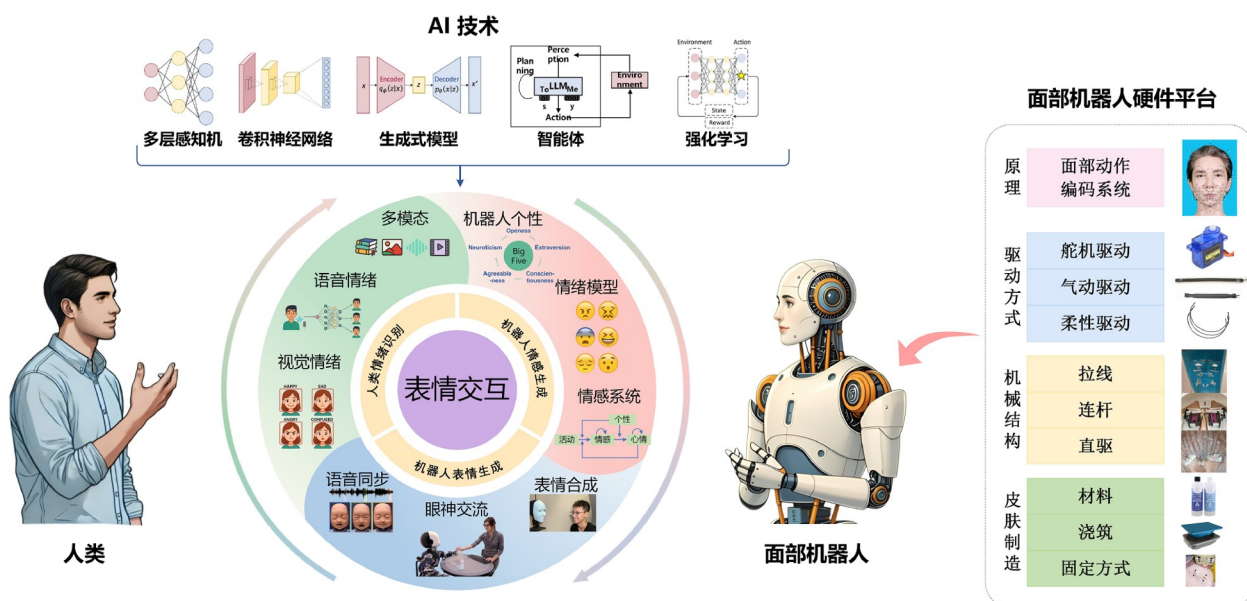


图1 (网络版彩图) AI驱动的人与面部机器人表情交互技术路线图。情绪模型, 引自文献[14]; 情感系统, 引自文献[20]; 表情合成, 引自文献[43]; 眼神交流, 引自文献[157]; 语音同步, 引自文献[16]

Figure 1 (Color online) Technological roadmap of AI-driven human-face robot expression interaction. Emotion model [14]. Copyright©2022, Elsevier. Emotion system [20]. Copyright©2024, The Author(s). Expression synthesis [43]. Copyright©2024, AAAS. Eye contact [157]. Copyright©2020, The Author(s). Speech Sync [16]. Copyright©2024, IEEE.

现基础, 系统阐述面部驱动原理、驱动方式、机械系统及皮肤制造技术之间的相互作用机制. 深入理解机械系统, 尤其是结构自由度与材料特性, 对于阐明基于AI的表情合成所面临的核心制约因素至关重要. 各子章节共同构建了后续AI算法的物理基础. 通过剖析这些工程问题, 本节阐明了机械结构与AI能力之间的内在关联.

2.1 面部驱动原理

面部表情的真实感是决定面部机器人能否克服“恐怖谷”(uncanny valley)效应的关键因素之一^[21]. 所谓“恐怖谷”现象, 是指当机器人在外观与行为上高度逼近人类, 但在某些细节上(特别是面部表情)存在轻微不协调时, 会引发人类的厌恶或不适感. 因此, 面部机器人设计中的核心挑战在于理解人类面部表情的生理机制, 并精确重现其动态特征.

人类面部表情是由多组面部肌肉的协同作用所产生. 例如, 愤怒表情源自皱眉肌和降眉间肌的协同收缩, 使眉毛向下拉拽. 面部表情的解剖与功能基础被系统地编码于“面部动作编码系统”(facial action coding system, FACS)中^[22], 如图2(a)所示. 该系统将面部表情分解为44个离散的“动作单元”(action units, AUs). 每个AU对应一个或一组面部肌肉的收缩或放松, 例如, AU12代表提颧肌, AU4代表皱眉肌. 复杂表情通过多个AU的动态组合而形成, 如“惊讶”表情通常由AU1+2+5+26的联合激活生成. 这些动作单元构成了面部机器人设计的关键原则, 并为其表情生成的准确性评估提供了客观基准. 对于基于AI驱动的表情交互而言, AUs在指导生成精确、自然的机器人表情以及识别用户情绪方面发挥着核心作用. 此外, AUs提供了一套标准化的评估指标, 用以衡量机器人面部表情的质量与有效性. 这一标准化评估对于训练和验证基于AI的表情控制算法不可或缺.

在具体实现中, 研究人员首先识别机器人面部的关键控制区域, 对应于人类面部的主要表达区域: 眉、眼、颊和唇. 通过分析人类面部肌肉的解剖结构及其在不同情绪下的运动模式, 研究人员将每种表情分解为多个控制点的协同动作. 例如, 日本开发的SAYA机器人^[7]中, 研究者依据FACS选取了14个关键AU, 并配置了19个控制点用于驱动机器人的面部皮肤. 这些控制点被精准对齐于相应人类面部肌肉的附

着位置^[23]. 这些AUs控制多个气动人工肌肉(McKibben驱动器), 同步牵引额部与口周区域, 从而有效再现人类“恐惧”表情中扬眉紧唇的生理特征.

然而, FACS作为静态编码系统, 在动态时序控制方面存在局限性^[24]. 尽管AU组合可描述表情的最终配置, 但该系统并未提供肌肉收缩的时间序列、速度或强度等动态参数的指导. 因此, 在表达生成过程中, 仍需额外策略来实现自然流畅的表情转换. 尽管如此, FACS凭借其解剖学精度, 仍是当前机器人面部表情建模的基础框架.

2.2 驱动方式

面部机器人的表情生成依赖于对控制点的精确操控. 表情的质量, 尤其在柔性、自然度与稳定性方面, 在很大程度上取决于所采用的驱动方式, 即实现面部运动的动力传递机制. 驱动方法的选择直接影响机器人执行精确面部动作的能力, 并进而制约AI算法的有效性. 例如, 柔性驱动系统较慢的响应速度可能限制AI算法实时更新表情的频率, 从而损害表情的自然度. 每种驱动方法都具有独特的控制特性、响应特性及自由度, 这些特性共同影响着AI驱动方法生成表情的精确性. 目前, 主流的面部驱动技术主要包括电机驱动、气动驱动与柔性驱动三类. 本节将系统分析这3种驱动方式.

(1) 电机驱动是面部机器人中最为常用的驱动技术之一, 通常采用伺服电机或步进电机, 通过机械连杆或线缆将旋转运动转化为控制点的线性或非线性位移. 例如, Faraj等人^[25]开发的低成本开源机器人Eva配备了25个电机, 其中12个用于面部表情控制, 能够实现最大15 mm的面部皮肤位移, 支持微笑、皱眉等基本表情, 并控制总头部重量在3 kg以内. 又如, Sophia使用23个舵机驱动面部表情, 5个舵机驱动脖子运动, 整个面部达到27个自由度. 配合Frubber弹性皮肤材料, 实现高自由度的眉毛、嘴角与眼部等多部件配合动作. 尽管电机驱动具备响应快、控制精度高等优势, 但也存在结构笨重、噪声大、易发热与机构复杂等缺陷, 在对空间与重量敏感的应用场景中存在限制.

(2) 气动驱动则常以McKibben人工肌肉为代表, 依靠压缩空气驱动柔性气动装置, 模拟人类肌肉的动态收缩与拉伸^[26]. 例如, 日本的SAYA机器人^[7]利用McKibben气动肌肉驱动19个关键面部控制点, 能够表

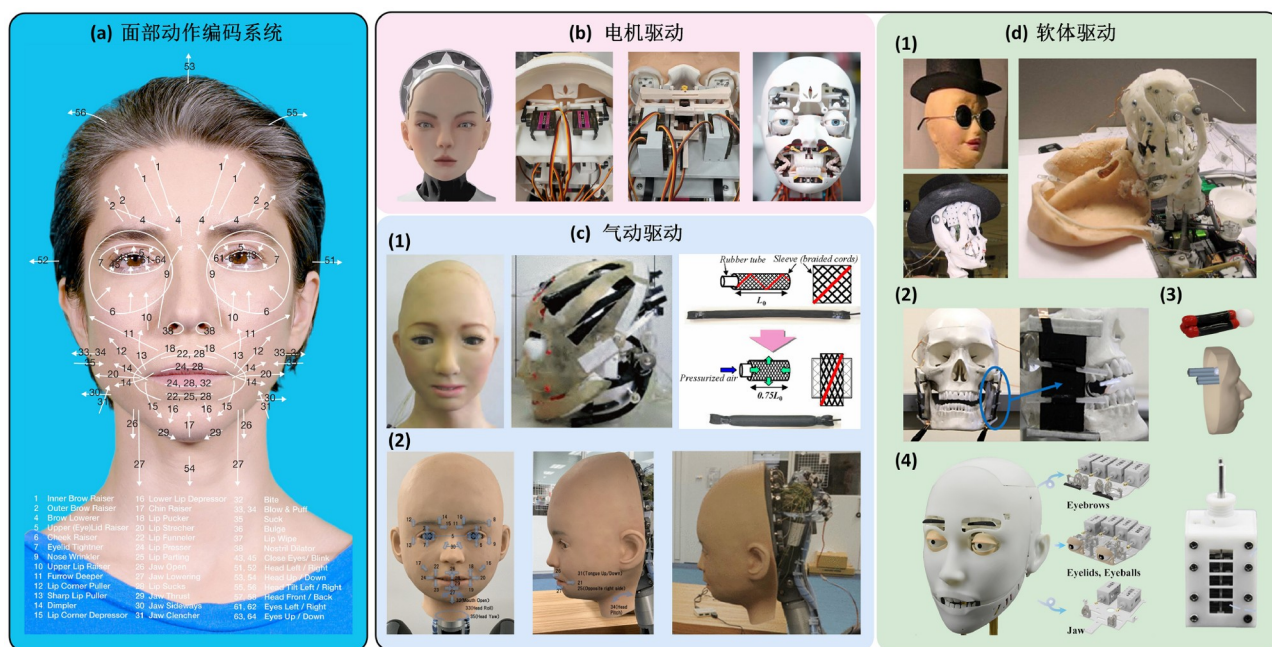


图 2 (网络版彩图)面部机器人的驱动原理和方法。(a) 面部动作编码系统, 将面部表情分解为44个离散动作单元^[22]。(b) 电机驱动的面部机器人: 小瑶, 配备26个伺服电机^[3]。(c) 气动驱动的面部机器人: (1) SAYA机器人采用McKibben气动人工肌肉^[7]; (2) Nikola机器人使用35个气动执行器^[29]。(d) 软执行器驱动的面部机器人: (1) 基于SMA执行器的面部机器人^[10]; (2) 使用DEA的人工下颌^[11]; (3) 由DEA驱动的人工眼球^[34]; (4) 由多层PVC凝胶执行器驱动的面部机器人^[38]

Figure 2 (Color online) Actuation principles and methods of face robots. (a) Facial action coding system, which decomposes facial expressions into 44 discrete action units ^[22]. Copyright©2025, Coralie Vogelaar. (b) Motor-actuated face robots: XiaoYao, equipped with 26 servo motors ^[3]. Copyright©2025, Qincheng Sheng. (c) Pneumatic-actuated face robots: (1) The SAYA robot employs McKibben pneumatic muscles ^[7]. Copyright©2006, IEEE. (2) The Nikola robot uses 35 pneumatic actuators ^[29]. Copyright©2022, IEEE. (d) Soft-actuated face robots: (1) A face robot based on SMA actuators ^[10]. Copyright©2011, ASME. (2) An artificial jaw using DEAs ^[11]. Copyright©2015, Elsevier. (3) Artificial eyeballs powered by DEAs ^[34]. Copyright©2019, IEEE. (4) A face robot actuated by multi-layer PVC gel actuators ^[38]. Copyright©2024, Elsevier.

达6种基本情绪(图2(c)-1)。而日本的Ishiguro团队^[27]开发的机器人主要使用气动执行器。例如, Geminoid-H1机器人的面部分配有约13个气动执行器, 加上躯干与四肢的其他气动致动器, 总共50个驱动单元用于整个机器人表情与动作的生成。其团队的另一款ERICA机器人^[28]使用13个气动驱动器, 精确模拟眉毛、眼睑与嘴角的运动, 同时具备更紧凑的设计结构。Diel等人^[29]进一步展示了气动驱动在生成复杂表情方面的潜力, 例如, Nikola机器人使用35个气动驱动器实现了“犹豫”、“自豪”等13种复杂情绪的表达^[30](图2(c)-2)。不过, 气动驱动系统通常存在响应延迟(从几十到数百毫秒不等), 且需依赖外部空气源, 压缩空气系统庞大, 气路控制复杂, 使其系统构建与维护较为繁琐。

(3) 柔性驱动技术通过仿生柔性材料模拟生物肌肉特性, 涵盖多种新型驱动机制, 如SMA、DEA、电活性聚合物(electroactive polymer, EAP)等。SMA驱动

具有结构紧凑的优势, 如Hara等人^[31]开发的机器人可通过SMA生成6种基本表情。Tadesse等人^[10]优化了SMA布线设计, 使12自由度的婴儿机器人头部在下颌运动上更加自然(图2(d)-1)。然而, SMA存在响应慢(通常需1~3 s)、成本高、能耗大等问题, 正逐步被其他技术替代。DEA因响应速度快而受到关注^[32], 例如, Kwak等人^[33]开发的机器人可生成6种基本表情; Wang和Zhu^[11]对其进行增强设计, 引入塑料纤维增强, 实现精确的下颌运动, 运动幅度接近人类自然下颌(图2(d)-2); Li等人^[34]开发的DEA驱动眼球机构可实现水平、垂直与圆周轨迹运动(图2(d)-3)。EAP具备低能耗优势, Hanson和White^[35]使用EAP驱动器实现轻力度表情动作。此外, 扭线聚合物(twisted-coiled polymer, TCP)人工肌肉^[36]结构紧凑, 适用于狭小空间; 电流体泵驱动的柔性弯曲执行器可以实现复杂的运动, 可用于面部机器人唇部机构的驱动^[37]。多层聚氯乙烯(polyvinyl

chloride, PVC)凝胶驱动器具备超过20%的收缩率与约45 ms的响应速度, 为轻量化面部机器人提供了前沿解决方案^[38](图2(d)-4)。然而, 柔性驱动技术仍面临长期稳定性不足与控制复杂度高等挑战。

综上所述, 当前的面部驱动技术各有优劣: 电机驱动稳定但结构庞大; 气动驱动更具仿生特性但系统复杂; 柔性驱动轻巧灵活但仍需提升可靠性与控制效率。面部机器人不同的硬件设计方式对比如表1所示。未来的发展趋势可能在于融合多种驱动技术的优势, 开发更轻量、耐用且响应频率更高的驱动器。

2.3 机械系统设计

机械系统是实现面部机器人有效表情交互的核心。其设计直接决定了面部动作的真实性与复杂度, 并深刻影响着AI算法的性能表现。设计精良的机械系统可以使AI算法能够精确地执行复杂的表情指令。该设计的关键在于面部控制点的合理分布与动作的高效传

递。本节将从控制点布局和机械传动机制两方面进行深入分析。

控制点的布局通常依托于解剖学知识, 尤其是FACS。根据FACS的指导原则, 控制点主要分布于额肌、眼睑、眼球、鼻部、面颊、口唇与下颌等关键肌肉节点或其末端位置(图3(a))。

常见布置为: 眉部使用4~6个控制点模拟额肌与皱眉肌的运动, 实现扬眉、皱眉等表情; 眼睑与眼球通过2~4个控制点控制眼睑的开闭, 眼球则配置2~3个自由度以支持水平与垂直转动; 面颊与鼻部配置4~6个控制点(其中鼻部2个, 面颊2~4个), 用于模拟微笑时的颊部提拉以及愤怒时的鼻翼扩张; 口唇与下颌使用12~14个控制点驱动口轮匝肌与咀嚼肌等肌肉, 支持嘴型变化与下颌开闭。

例如, Hashimoto等人^[39]在开发“Face Robot”时, 通过分析动态面部表情, 设计了39个控制点: 眉部6个、眼睑10个、口唇12个、下颌2个, 以高保真地再现目标

表1 面部机器人不同硬件设计方式对比

Table 1 Comparison of different hardware design methods for face robots

驱动方式	面部自由度 (DoF)	响应速度	表情逼真度	维护成本	优点	缺点	代表机器人
电机驱动	高(25~30)	极快(毫秒级)	中-高(精度高, 但自然度不足)	中(机构复杂, 需定期维护)	控制精度高、响应快、技术成熟	噪声大、重量大、结构复杂、发热严重	Sophia ^[8] , Ameca ^[9] , Albert HUBO ^[40]
气动驱动	中-高(13~35)	快(几十至数百毫秒)	高(运动柔和, 更仿生)	高(需外接气源, 系统复杂, 易泄漏)	输出力大、运动平滑柔和、仿生性好	响应延迟高、系统笨重、噪声大(气泵)	SAYA ^[7] , ERICA ^[28] , Nikola ^[29]
柔性驱动	低-中(e.g., 12)	SMA极慢(几秒) DEA快	中(受限于材料应变)	中-高(材料制备复杂, 驱动电路复杂)	结构紧凑、无噪声、质量小	SMA响应慢、滞后严重; DEA成本高、驱动电压高	基于SMA的面部机器人 ^[10] , 基于DEA的面部机器人 ^[33]

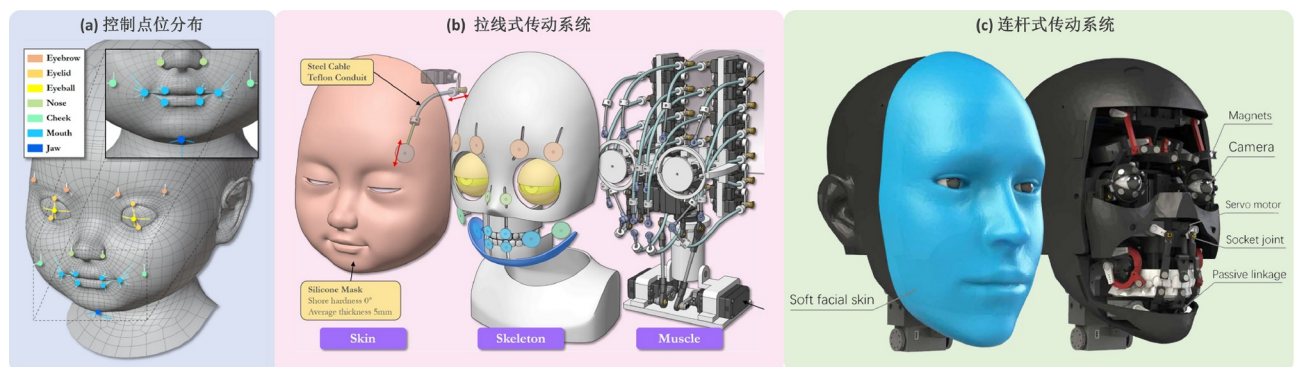


图3 (网络版彩图)面部机器人机械系统设计。(a) 控制点布局, 分布在眉毛、眼睑、眼球、鼻子、脸颊、嘴巴和下颌等关键面部区域^[16]。(b) 线拉机构系统^[16]。(c) 连杆驱动机构系统^[43]

Figure 3 (Color online) The hardware system design of face robots. (a) Layout of control points distributed across key facial regions, including the eye-brows, eyelids, eyeballs, nose, cheeks, mouth, and jaw ^[16]. Copyright©2024, IEEE. (b) The wire-pulling mechanism system ^[16]. Copyright©2024, IEEE. (c) The linkage-actuated mechanism system ^[43]. Copyright©2024, AAAS.

人物的面部特征. 控制点的布置需在面部皮肤柔性与机械结构限制之间取得平衡, 以在保证自然表情的同时, 控制系统复杂度.

控制点的传动方式直接影响面部表情的平滑性与可靠性. 主流的传动机制包括: 拉线式传动、连杆式传动与直接传动. 表2对不同面部机器人所采用的机械结构进行了比较.

拉线式传动系统是目前应用最广泛的机制之一. 该系统通过电机拉动与人工皮肤内部连接的柔性钢丝, 进而实现控制点的位移并引发局部表情变化(图3(b)). 例如, Albert HUBO机器人^[40]使用28个伺服电机驱动涂氟尼龙丝牵引面部锚点, 以实现精准表情控制; EveR-4机器人^[41]则通过33个电机与拉线机制实现眉部、眼睑与口唇的同步协调动作. Li等人^[16]进一步优化了拉线系统, 采用1.5 mm钢丝与特氟龙导管, 使唇部控制点的响应时间控制在0.04 s内. 尽管该系统成熟稳定, 但仍面临线缆磨损、控制精度低与维护频繁等问题, 且大部分线缆只能实现单向牵引, 回位需

依赖皮肤弹性或另设回拉线, 在眼口等空间受限部位不易隐藏.

连杆式传动系统通过刚性连接结构直接作用于皮肤控制点, 可实现双向或复杂曲线运动(图3(c)). 其优点包括刚性强、运动轨迹可控、无须依赖皮肤弹性, 以及模块化设计便于装配. 例如, Asheber等人^[42]使用仅8个伺服电机与简单连杆机构, 开发出低成本社交机器人, 实现了基本的眉、眼、唇表情; Hu等人^[43]开发的模块化EMO机器人结合连杆与磁力结构, 通过五杆机构精确驱动嘴角垂直运动, 位移精度达 ± 0.1 mm. 然而, 要实现复杂运动所需的多重连杆结构往往导致整体系统体积庞大、头部重量增加, 且难以适配人脸曲面, 限制了其对人类柔软表情的拟真能力.

混合式传动系统综合应用连杆、拉线、同步带与齿轮等多种机构, 兼顾各自优点, 以提升面部表情的真实度. 例如, Berns和Hirth^[44]在ROMAN机器人的颈部采用齿轮驱动头部俯仰, 配合拉线控制眉毛与嘴唇, 实现了 $\pm 30^\circ$ 转动范围; Affetto机器人^[45]综合使用连杆、

表 2 不同的面部机器人对比
Table 2 Comparison of different face robots

机器人名称	年份	所属机构	面部总自由度	详细自由度分布	驱动方式	细节描述	文献
SAYA	2006	东京理科大学	24	—	气动	聚氨酯树脂皮肤; McKibben执行器	[7]
ROMAN	2006	卡尔斯鲁厄工业大学	17	4 眉毛, 6 眼睛, 2 鼻子, 4 嘴巴, 1 下巴	电机	硅胶皮肤; 混合传动 (齿轮+线缆)	[44]
Affetto	2011	大阪大学	10	2 眉毛, 2 眼睑, 2 眼睛, 2 嘴巴, 2 下巴	电机	硅胶皮肤; 混合传动 (皮带、连杆)	[45]
EveR-4 H33	2012	韩国生产技术研究院	30	6 眉毛, 4 眼睑, 4 眼睛, 14 嘴巴, 2 舌头	电机	硅胶皮肤; 支持眉毛、眼睑和嘴唇的协调	[41]
PKD	2014	汉森机器人公司	24	5 眉毛, 6 眼睛, 2 脸颊, 9 嘴唇, 1 下巴, 1 鼻子	电机	硅胶皮肤; 拉线传动	[47]
ERICA	2016	京都大学	13	2 眉毛, 2 眼睛, 3 眼睑, 5 嘴巴	电机	嵌入式摄像头; 气动伺服阀	[28]
NTUST Robot	2016	台湾科技大学	6	3 脸, 2 眼睛, 1 下巴	气动	模块化设计支持基本面部表情	[54]
Abel	2021	比萨大学	16	4 眉毛, 8 眼睛, 1 下巴, 8 嘴巴	电机	仿生头部和上半身; 混合传动	[48]
Ameca	2021	工程艺术公司	25	4 眉毛, 4 眼睑, 4 眼睛, 12 嘴唇, 2 下巴	电机	模块化设计; 混合传动	[9]
Nikola	2022	理化学研究所	32	6 眉毛, 9 眼睛, 1 鼻子, 2 脸颊, 2 下巴, 10 嘴唇, 3 舌头	电机	硅胶皮肤; ROS控制系统	[30]
HHU Robot	2023	河海大学	9	4 眼睑, 2 脸颊, 1 嘴唇, 2 眉毛	电机	硅胶皮肤; 模块化设计; 拉线传动	[49]
BIGAI Robot	2024	北京通用人工智能研究院	24	4 眉毛, 8 眼睛, 2 鼻子, 2 脸颊, 8 嘴巴	气动	硅胶面罩; 拉线传动肌腱驱动执行器	[16]
EMO	2024	哥伦比亚大学	23	4 眉毛, 4 眼睑, 4 眼睛, 10 嘴巴, 1 下巴	电机	带磁铁的硅胶皮肤 连杆驱动	[43]

同步带与气动驱动, 其中同步带驱动眼球水平转动, 连杆控制下颌动作; Engineered Arts开发的Ameca机器人^[9]更是融合齿轮、连杆与同步带三者, 赋予其27个自由度, 覆盖面部多个区域, 支持12种复杂表情。

直接驱动方式(人工肌肉或嵌入式驱动)通过将驱动器直接集成于人工皮肤内部, 最大限度模拟人脸自然肌肉运动机制。例如, SAYA^[7]与Face Robot^[39]均采用McKibben气动肌肉实现直接牵引, 精确模拟肌肉收缩; Almubarak等人^[46]提出将TCP人工肌肉嵌入硅胶仿生皮肤中, 该肌肉由镀银尼龙通过扭转、缠绕等工艺制成, 具备结构紧凑、响应快速的特点, 极大提高了面部表情的真实感与动态响应性。但该方法控制复杂、制造工艺要求高, 尚未大规模应用。

综上所述, 面部机器人传动技术正在向更逼真、更精确、更紧凑的方向演进。拉线传动机制尽管成熟, 但维护成本高^[47], 连杆系统精度优越, 却结构复杂^[48,49]; 直接驱动方式具备仿生优势与结构紧凑性, 但仍需在稳定性与解剖结构匹配性方面进一步优化^[50]。

2.4 面部机器人皮肤制作技术

类人面部皮肤技术对于实现自然且富有情感的人机交互至关重要。作为非语言交流的核心媒介, 机器人皮肤不仅需要呈现逼真的外观, 还应具备与人类皮肤相似的弹性、柔软度与耐久性, 并集成感知外部刺激的能力。具备真实触感与响应的皮肤可显著增强机器人传递情绪的能力, 从而提升用户的情感认同与接受度^[51]。皮肤材料的弹性、厚度及其固定方式共同决定了面部表情的灵活性与响应速度。这些物理特性从根本上制约着AI算法将控制信号转化为自然、连贯表情动作的有效性。理想情况下, 机器人面部皮肤应具备多模态感知能力, 能够识别触觉、温度与压力等多种刺激, 以支持细腻的情绪表达, 并在交互过程中实现情境感知与自适应响应。这种感知能力不仅有助于机器人更自然地表达情绪, 也为实现具备共情能力的高阶情感交互奠定了基础。

2.4.1 面部皮肤材料选择

在选择机器人面部皮肤材料时, 必须确保其具备高延展性、柔软性与生物相容性, 以支持逼真的面部变形与表情生成。常用材料主要包括硅胶、聚氨酯弹性体以及Frubber等仿生材料。例如, 由东京理科大学

开发的SAYA机器人^[7]采用柔性聚氨酯树脂作为面部皮肤材料, 该材料具有优异的弹性与柔韧性, 可实现大幅度的面部变形, 支持丰富的表情表现。大阪大学研制的Affetto机器人^[45]则采用一种特殊的聚氨酯弹性凝胶, 能够高度模拟婴儿皮肤的质感与弹性, 使机器人具备更真实的面部触感。Hanson Robotics公司使用的Frubber材料是一种轻质、多孔、高弹性的仿生皮肤, 专为复制人类皮肤的力学特性而设计。该材料可在施加极小力的情况下产生大幅变形, 从而显著减轻驱动器负载, 同时保持高度的表情保真度^[52]。此外, Park等人^[53]提出了一种仿生皮肤复合结构, 由水凝胶与硅胶橡胶组成, 并嵌入电阻器与振动传感器, 从而实现高精度的触觉感知, 为面部机器人提供了更丰富的人机交互能力。

2.4.2 面部皮肤浇筑技术

面部机器人的皮肤制造主要依赖于模具成型技术。目前常用的方法包括包覆成型、3D打印辅助模具制造以及嵌入式制造。Lin等人^[54]采用包覆成型技术制备聚氨酯橡胶皮肤: 首先使用藻酸盐材料制作面部印模, 进而生成石膏模具以捕捉面部几何特征; 随后形成硅胶模与黏土模型, 构建注射腔体, 最终注入并固化聚氨酯橡胶, 经拆模后得到最终的面部皮肤(图4(a)-1)。Hu等人^[43]在EMO机器人开发中应用了3D打印辅助模具制造技术, 选用EcoFlex 00-30铂金硅胶作为皮肤材料。通过3D打印正模与负模, 成功实现了高精度面部皮肤的铸造(图4(a)-2)。Almubarak等人^[46]进一步提出嵌入式制造方案, 将TCP人工肌肉直接嵌入硅胶皮肤中, 提升面部动态响应能力(图4(a)-3)。Tomar等人^[55]则在此基础上开发出多层复合皮肤结构, 将传感器、驱动器与保护层集成于统一的硅胶体内, 显著提升皮肤系统的紧凑性与多功能性。此外, Zhong等人^[56]引入一种创新的“化学流体皮肤”, 通过流体控制实现皮肤的形变与变色效果, 极大增强了机器人面部的动态表现力与仿生逼真度(图4(a)-4)。

2.4.3 面部皮肤固定方法

将面部皮肤固定于3D打印的机器人颅骨结构的方式, 对于实现自然、逼真的面部表情至关重要^[57]。目前常用的固定方法包括黏合固定、卡扣式结构和磁吸式连接。黏合固定法操作简便, 具有良好的结构稳定



图 4 (网络版彩图)面部机器人皮肤的制造技术。(a) 机器人面部皮肤制造。(1) 包覆成型: 通过成型、注射和固化聚氨酯橡胶来制造面部皮肤^[54]。(2) 3D打印辅助成型: 使用3D打印机打印内外模具, 然后将两个模具组合以获得机器人面部皮肤^[43]。(3) 嵌入式制造: 将TCP人工肌肉嵌入硅胶橡胶皮肤^[46]。(4) 嵌入式制造: 将化学流体结合到面部皮肤中^[56]。(b) 皮肤粘合剂黏合法^[25]。(2) 卡扣连接法^[58]。(3) 磁铁法^[43]

Figure 4 (Color online) Fabrication techniques of face robot skin. (a) Robot face skin fabrication. (1) Overmolding: a process of making face skin by molding, injecting, and curing urethane rubber [54]. Copyright©2014, Cambridge University Press. (2) Three-dimensional (3D) printing-assisted molding: using FDM 3D printer to print inner and outer molds and then combining two molds to get robot face skin [43]. Copyright©2024, AAAS. (3) Embedded manufacturing: embedded TCP artificial muscles into silicone rubber skin [46]. Copyright©2018, SPIE. (4) Embedded manufacturing: combine chemical fluid into face skin [56]. Copyright©2025, AAAS. (b) Skin attachment methods. (1) Adhesive bonding method [25]. Copyright©2020, The Author(s). (2) Snap-fit method [58]. Copyright©2012, Springer. (3) Magnets methods [43]. Copyright©2024, AAAS.

性。例如, Faraj等人^[25]开发的EVA机器人采用胶水与纤维布结合的方式, 将拉线锚点牢固地附着在硅胶皮肤内侧, 有效保证了皮肤在变形过程中的控制点定位稳定(图4(b)-1)。尽管该方法稳定性强, 但在后期维护与皮肤更换过程中操作较为繁琐。为提升可维护性, 研究者引入了卡扣式结构, 通过机械连接件将皮肤固定于颅骨结构^[58]。此类设计在提高连接精度与便捷性的同时, 增强了系统的可重复使用性。卡扣结构通常带有一定弹性, 可有效缓解局部应力, 减少皮肤撕裂的风险(图4(b)-2)。然而, 长时间使用后, 由于反复受力与材料老化, 连接部位可能发生松动或失效, 导致表情区域的控制能力下降。为克服上述问题, 磁吸式连接逐渐成为更为可靠且广泛应用的解决方案。例如, EMO机

器人^[43]采用嵌入式磁铁将面部皮肤与颅骨固定连接。这种方法不仅便于皮肤的快速拆卸与更换, 还能维持皮肤在变形过程中的自然形态(图4(b)-3), 在实现结构灵活性的同时兼顾表情的动态稳定性。

3 面部机器人表情交互的传统AI实现路径

本节基于“感知-推理-生成”框架, 阐述面部机器人表情交互的传统AI实现路径。该框架为理解面部机器人如何感知人类情绪、进行推理并生成情境适宜的表情的提供了基础。首先介绍单模态感知方法, 随后详细阐述基于规则的情绪建模方法, 最后论述如何将计算所得的情绪状态应用于面部表情生成、视线交互以及语

音与表情的同步. 尽管该流程能有效实现表情交互, 但其模块化设计在情境适应性方面存在固有限制, 因此在第4节引入LLM驱动的先方法以解决此类限制, 两种交互方法对比如表3所示.

3.1 情感识别

人类情感识别在实现人机表情交互中扮演着关键角色, 涵盖了心理与生理两个维度. 研究表明, 情感信息主要通过面部表情(55%)、语音语调(38%)以及语言内容(7%)进行传达^[6]. 尽管脑电(electroencephalogram, EEG)、心电(electrocardiogram, ECG)等生理信号亦能反映个体的情绪状态, 但因其依赖专业设备, 实际应用受限. 近年来, 情感计算技术已从单一模态识别向多模态融合方向发展, 通过整合视觉、语音与文本等多种数据源, 提升情绪识别的鲁棒性与准确性^[59]. 第4节将详细探讨多模态情绪识别技术与基于LLM的情感推理方法. 本节将聚焦于人机表情交互中常用的视觉与语音情感识别技术, 作为后续高层推理与表情生成的感知基础.

3.1.1 视觉情感识别

(1) 面部表情分析

面部表情分析在人机表情交互中至关重要, 能够帮助机器人理解并响应人类情绪, 从而实现更自然的交流体验. 然而, 由于面部图像的复杂性、环境干扰以及表情本身的微妙性, 准确解读面部表情仍面临诸多挑战. 为了解决以上问题, 表情识别过程通常被划分为3个阶段: 图像预处理、面部特征提取与面部情绪识别(facial emotion recognition, FER)(图5(a)-1). 其中, 特征提取包括面部关键点检测与AU检测及其强度估计. 预处理可以提升数据质量, 面部特征检测可以建

立面部结构, AU分析可以捕捉肌肉动态, FER可以对情绪进行分类, 所有这些技术共同实现了从“像素”到“情绪”的完整映射. 本节将详细探讨这些技术.

图像预处理在FER中至关重要, 因为原始图像常受到光照条件、姿态变化以及背景干扰等因素影响, 从而降低模型的识别性能. 常见预处理方法包括: 人脸检测、对齐、光照归一化与数据增强. 其中, 人脸检测用于剔除无关背景信息; 对齐通过眼角、嘴角等关键点调整人脸位置与角度^[60,61]; 光照归一化与对比度增强有助于改善面部特征清晰度^[62]; 数据增强(如旋转、翻转)提升训练数据多样性, 增强模型泛化能力^[63]. 上述方法可有效降低后续识别难度, 并提升整体识别精度^[64,65].

面部关键点检测用于定位眼睛、鼻子、嘴巴等核心部位, 是面部表情分析与情绪识别的关键(图5(a)-2)^[66]. 目前主要方法包括坐标回归法与热图回归法. 坐标回归计算效率高, 但对姿态变化与遮挡敏感^[67,68]; 热图回归通过预测关键点的概率分布, 提升了鲁棒性, 但计算开销大, 不利于实时应用^[69,70]. 为解决这些问题, 研究提出了混合方法与基于三维模型的方法. 前者结合两种回归方法的优势^[71,72], 后者利用3D人脸结构信息, 增强了对复杂环境下姿态变化与遮挡的适应能力, 提升了特征点定位精度^[73,74].

AU检测用于识别由面部肌肉运动并通过强度估计量化其运动幅度, 以更准确地判断情绪状态(图5(a)-3)^[75,76]. 当前方法主要包括3类: 监督学习方法、领域迁移方法与多任务学习方法. 监督学习细分为自监督、弱监督与半监督方法: 自监督方法从无标签图像中提取AU特征; 弱监督依赖不完全标注提升准确率^[77,78]; 半监督结合标签与无标签数据, 通过协同正则或对抗学习提升模型性能^[79,80]. 领域迁移方法

表 3 AI驱动的面部机器人表情交互方法对比
Table 3 Comparison of AI driven face robot human-robot expression interaction method

方法类别	输入模态	情感建模方式	生成机制	优劣势对比	代表性研究
传统方法	视觉、语音	离散情感模型、维度情感模型 (如VA, PAD)	预定义表情映射、基于规则或模板库	优点: 结构简单、实时性较好; 缺点: 情感理解浅层、缺乏上下文适应性、难以处理混合情绪	Han等人 ^[148] Ishihara等人 ^[133]
先进方法	视觉、语音、文本、生理信号、触觉等	大语言模型、多模态大语言模型驱动的上下文推理	端到端神经网络、利用Agent生成动作代码	优点: 表达丰富、泛化能力强; 缺点: 算力消耗大、实时性瓶颈、训练数据依赖强、存在幻觉现象、伦理与隐私风险	Yoshida等人 ^[19] , Kang等人 ^[20]

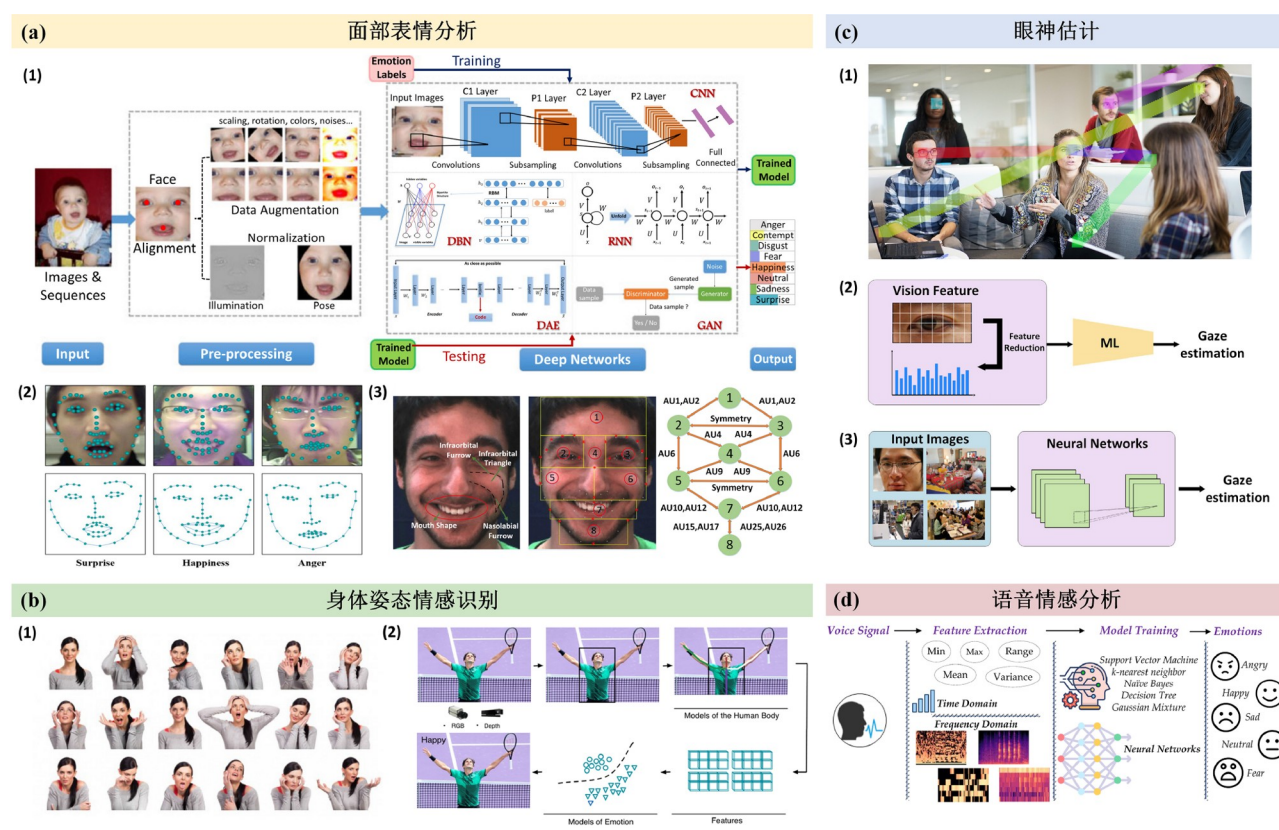


图5 (网络版彩图)视觉情感识别((a)~(c))和语音情感识别(d). (a) 面部情感识别: (1) 深度面部表情识别系统的通用流程^[62]; (2) 从面部图像中提取的面部关键点^[65]; (3) 基于知识驱动的自监督表征学习的面部动作单元识别^[77]. (b) 身体姿态情感识别: (1) 表达情绪的肢体语言示例^[103]; (2) 情绪肢体语言识别系统概述^[103]. (c) 注视估计: (1) 多人交互过程中注视估计的示例^[109]; (2) 传统的基于外观的注视估计方法^[112]; (3) 基于深度学习的注视估计方法^[112]. (d) 语音情感识别的处理流程^[116]

Figure 5 (Color online) Visual emotion recognition ((a)~(c)) and speech emotion recognition (d). (a) Facial emotion recognition: (1) The general pipeline of deep facial expression recognition systems ^[62]. Copyright©2022, IEEE. (2) Facial landmarks extracted from facial images ^[65]. Copyright©2023, IEEE. (3) Facial action unit recognition using knowledge-driven self-supervised representation learning ^[77]. Copyright©2022, IEEE. (b) Emotional body gesture recognition: (1) Examples of body language indicative of emotion ^[103]. Copyright©2021, IEEE. (2) General overview of an emotional body gesture recognition system ^[103]. Copyright©2021, IEEE. (c) Gaze estimation: (1) Example of gaze estimation during multi-person interaction ^[109]. Copyright©2020, The Author(s). (2) Traditional appearance-based gaze estimation method ^[112]. Copyright©2022, Elsevier. (3) Deep learning-based gaze estimation method ^[112]. Copyright©2022, Elsevier. (d) The processing flow of speech emotion recognition ^[116]. Copyright©2023, The Author(s).

通过缩小数据集间的域差异, 提高模型在不同应用场景下的泛化能力^[81,82]. 多任务学习方法通过同时引入面部关键点检测等辅助任务增强AU检测效果^[83,84]. 在强度估计方面, 常用方法包括概率建模方法^[85,86]、深度神经网络^[87]以及上下文感知方法, 后者可更好地捕捉AU在时序中的动态变化^[88,89].

通过面部关键点与AU检测所获取的肌肉运动模式, 为FER提供了关键支持, 广泛应用于人机交互与情感计算等领域^[90]. 当前FER主要研究方向包括3类: 离散情绪识别、连续情绪识别与微表情识别. 离散情绪识别聚焦于识别“高兴”“愤怒”等基本情绪, 通常基于

静态图像或视频, 通过卷积神经网络(convolutional neural network, CNN)、生成对抗网络(generative adversarial network, GAN)与注意力机制等算法不断提升其准确率^[91~93]. 动态视频情绪识别引入时间信息, 采用CNN、循环神经网络(recurrent neural network, RNN)与注意力机制建模情绪的时序演化^[94~97]. 连续情绪识别基于情绪维度(如愉悦度与唤醒度), 采用时空融合与三维人脸建模, 追踪情绪在连续时间上的变化趋势^[98~101]. 微表情识别针对持续时间短、强度低的情绪信号, 依赖高精度视频分析技术识别细微表情变化^[102].

(2) 身体姿态情感识别

身体语言作为非语言交流的重要组成部分, 包含手势、头部动作、姿态与步态等, 能够传达丰富的情绪信息(图5(b)-1). 通过识别这些身体信号, 机器人系统可推断用户的情绪状态, 从而增强人机交互的自然性与情感深度. 作为情感计算的重要研究方向之一, 身体姿态情感识别旨在通过图像、视频或深度图等视觉输入, 分析人体动作与姿态, 并将其映射至情绪空间, 用于识别“高兴”“愤怒”“悲伤”或“惊讶”等离散情绪, 或预测如“愉悦度(valence)”与“唤醒度(arousal)”等连续情绪维度^[103]. 该识别过程一般包括4个核心步骤: 目标检测、身体姿态估计、特征提取与情绪识别(图5(b)-2). 第一阶段目标检测通过算法识别图像中的人体区域. 传统方法如基于梯度方向直方图(histogram of oriented gradient, HOG)的方法, 或更先进的深度学习算法如基于快速区域的卷积神经网络(fast region-based convolutional neural network, faster R-CNN)被广泛用于此阶段^[104]. 姿态估计阶段通过基于模型的方法或运动学骨架模型, 定位身体各关节位置, 从而建立人体结构模型^[105,106]. 特征提取阶段将捕捉到的身体动作数据转化为反映情绪状态的特征表示. 可采用人工设计的统计特征, 或利用CNN、长短时记忆网络(long short-term memory, LSTM)等深度学习技术提取时空特征. 情绪识别阶段将提取的特征映射至情绪模型. 该过程可采用传统机器学习分类或回归方法, 也可采用多任务学习与多模态融合的深度神经网络模型, 以提升识别准确性与情境适应性^[107,108].

(3) 眼神估计

眼神估计是一种通过分析眼部或面部图像来判断人类注视方向的技术(图5(c)-1)^[109]. 该技术在人机情感交互中具有重要作用, 能够揭示人的注意力焦点、情绪状态与参与程度. 注视不仅指示关注的目标对象, 还反映了注意力强度及潜在的情绪反应, 如压力或疲劳. 当机器人能够准确感知人的注视方向时, 便可做出更加自然与具同理心的响应, 从而提升交互的亲合力与情感适应能力. 根据实现方式与应用场景, 注视估计方法可大致分为五类. 第一类是基于附着式传感器的方法, 如在眼部周围贴附电极, 可实现高精度测量, 但会对使用者造成不便, 通常用于医疗或实验环境中^[110]. 第二类是基于三维眼球模型重建的方法, 依赖角膜反射等光学原理实现注视方向计算, 需个性

化校准并依赖专用硬件支持. 第三类方法通过检测二维眼部特征, 如瞳孔中心或高亮点位置, 配合红外摄像头实现注视估计, 在研究领域较为常见. 第四类是传统的基于外观的图像分析方法, 利用眼周或面部图像进行估计, 但在复杂环境中的鲁棒性较弱(图5(c)-2)^[111,112]. 第五类则是当前主流的基于深度学习的方法, 采用CNN或变压器(transformer)结构, 从眼部或面部图像中自动学习多尺度特征, 显著增强了模型在外貌差异、光照变化与姿态变化等复杂条件下的适应能力(图5(c)-3)^[113~115].

3.1.2 语音情感识别

语音情感识别(speech emotion recognition, SER)是一种通过分析语音信号中的语言内容及副语言特征, 自动识别说话者情绪状态的技术. 在人机交互中, 语音不仅承载语义信息, 还包含诸如音高、强度、节奏、音色与相位等声学特征, 这些变化往往能够反映出情绪的动态状态^[116]. SER的任务通常包括将语音情绪归类为离散标签(如高兴、悲伤、愤怒等)或预测情绪维度(如唤醒度、愉悦度)上的连续值. 输入数据可为原始音频、频谱图或预处理后的低层特征描述符, 输出则表现为情绪标签或在情感空间中的坐标点. 整体处理流程如图5(d)所示.

特征提取是SER的关键环节, 决定了系统是否能捕捉语音中的情感细节. 所提取的声学特征大致可分为韵律特征、谱特征、音质特征与相位特征. 韵律特征如基频(F0)、能量与持续时间可揭示情绪激活程度^[117]; 谱特征包括梅尔频率倒谱系数(Mel-frequency cepstral coefficients, MFCC)、线性预测倒谱系数(linear prediction cepstral coefficients, LPCC)与对数频率功率系数(log-frequency power coefficients, LFPC), 反映了声道结构变化^[118,119]; 音质特征如抖动(jitter)、闪烁(shimmer)与谐噪比(harmonic-to-noise ratio, HNR)则揭示了声门运动状态及其在情绪表达中的作用^[120]. 近年来, 长期被忽视的相位特征逐渐引起关注, 其作为振幅信息的补充, 可提供更丰富的情绪表征. 例如, 群时延、修正群时延倒谱系数(modified group delay cepstral coefficients, MGDCC)与动态相对相位(dynamic relative phase, DRP)等特征, 与传统振幅特征结合后, 可有效提升情绪识别性能^[121].

传统SER方法依赖人工提取的声学特征, 并将其

输入至支持向量机(support vector machine, SVM)、K近邻(K nearest neighbor, KNN)或隐马尔可夫模型(hidden Markov models, HMM)等机器学习模型中进行识别^[122]。这类方法直观且易于解释,但难以捕捉情绪的复杂变化模式,且高度依赖专家特征设计。随着深度学习的发展,端到端的SER模型逐渐成为主流。CNN能够提取局部时频特征^[123],RNN与LSTM擅长建模语音的时间序列特征^[124],而Transformer与图卷积网络(graph convolutional networks, GCN)则通过建模长距离依赖关系与语音结构之间的连接性,进一步提升情绪感知的敏感性与系统鲁棒性^[125-127]。这些深度结构不仅提升了识别精度,也减少了对人工特征工程的依赖。

此外,连续情绪识别作为近年来的研究趋势之一,关注愉悦度、唤醒度与支配度等维度的回归建模,借助大规模情感数据库与深度模型,对情绪变化实现更细腻的表达。未来,随着多模态输入融合、注意力机制与自监督学习等新技术的发展,SER系统将更具准确性、实时性与鲁棒性,进一步推动机器人实现更具情绪响应性与自然性的交互能力。

总体来看,传统基于单模态信号的情绪识别方法在小规模、受控环境下能够达到较高准确率,但在复杂场景下鲁棒性显著下降。其优势在于实现简便、计算成本低,但在实时性和跨场景适应能力上存在明显瓶颈。随着多模态融合方法的发展,传统感知手段逐渐难以满足自然交互需求,这也为第4节所讨论的先进感知方法奠定了研究背景。

3.2 面部机器人情绪生成

3.2.1 情感模型

情感模型是用于结构化描述和分类人类情感状态的理论框架,其核心目标在于将复杂而动态的情绪体验转化为可量化或可计算的参数,从而使机器人能够理解、解释并回应人类情绪。在表情交互的语境下,情感模型为机器人识别、模拟与表达情绪提供了理论基础,支持更自然、更具人类特征的交流方式^[13,128,129]。借助情感模型,机器人能够对用户的语音、面部表情与身体语言等情绪线索进行综合分析,并据此调整自身行为与反馈,从而优化用户体验^[12,130]。本节将系统探讨三类主要的情感模型:离散情感模型、维度情感模型与其他补充性情感模型。通过对这些模型的系统回顾,本文旨在为情感交互型机器人系

统的设计提供明确的理论依据与实践指导。

(1) 离散情感模型

离散情感模型认为人类情绪可划分为若干固定且相互独立的类别,每一类情绪都对应特定的生理反应与行为特征。这类模型的提出基于跨文化研究发现:人们在面对相同情绪刺激时,往往会表现出一致的情绪反应。Ekman^[131]提出的基本情感模型识别出6种跨文化通用的基本情绪——愤怒、恐惧、厌恶、高兴、悲伤与惊讶,这些情绪可通过一致性的面部表情加以识别(图6(a)-1)。Plutchik和Kellerman^[132]进一步发展了这一理念,提出情感轮模型,在其中引入8种基本情绪,并以图形方式展现其强度变化与相互关系,从而帮助机器人实现更具层次与细腻度的情感表达(图6(a)-2)。在机器人系统中,离散情感模型凭借其结构清晰与实现简便的优势,广泛用于情绪识别与表情生成,通过预设的情绪标签实现快速响应。然而,此类模型难以刻画情绪体验的连续性与多层次性,尤其是在识别混合情绪与文化差异方面存在一定局限。这些不足促使研究者转向维度情感模型的探索,即将情绪表示为连续情感空间中的坐标点,以更真实地反映情绪状态的复杂性。

(2) 维度情感模型

维度情感模型通过在连续的多维情感空间中描述情绪状态,使情绪可以根据若干关键情感维度映射至特定坐标。与离散情感模型不同,该类模型认为情绪并非相互独立的分类,而是以连续的方式存在,支持情绪的平滑过渡。这一结构不仅使机器人能够解析情绪的强度变化,还能识别相似情绪状态之间的细微差异,从而实现更自然、更具情境感知能力的交互体验。

常见的维度情感模型包括多个变体。最早由Johnston^[133]提出的一维情感模型使用单一轴线将情绪划分为正向与负向,尽管结构简单,但在表达情绪细节方面存在局限性。Russell^[134]提出的二维情感模型(效价-唤醒, valence-arousal, VA)根据情绪的愉悦程度与激活水平进行映射,在语音、面部与身体情绪识别中得到广泛应用(图6(b)-1)。Mehrabian和Russell^[135]进一步提出的三维情感模型(愉悦-唤醒-支配, pleasure-arousal-dominance, PAD)在VA模型基础上引入“支配度”(dominance)维度,用以区分具有相似愉悦度与唤醒度的情绪状态(图6(b)-2)。在此基础上,一些研究者提出了面向特定交互场景的扩展模型。例如, Breazeal

和Scassellati^[136]提出的AVS模型引入姿态(stance)维度, 而Yamazaki等人^[137]则提出APA模型(affinity-pleasure-arousal), 强化了人际亲和性的映射能力. 此外, Izard^[138]提出的四维情感模型(pleasantness-tension-excitement-certainty, PTEC)使用“愉悦度-紧张度-兴奋度-确定性”4个维度细化情绪描述, Cambria等人^[139]提出的情绪沙漏模型则引入“注意、敏感、确定”等更细腻的情绪维度, 支持在多模态人机交互中的实时细粒度情绪追踪.

(3) 其他情感模型

除了离散情感模型与维度情感模型之外, 近年来还涌现出多种其他类型的情感模型, 它们从不同理论视角出发, 为理解与生成情感提供了补充路径, 进一步丰富了人机情感交互的建模基础. 基于认知的情感模型, 如Ortony-Clore-Collins (OCC)模型^[140,141], 强调情绪生成的逻辑链条, 认为情绪源自个体对事件、行为或对象的认知评估. 该模型定义了22种情绪类型,

能够支持机器人根据交互情境推理并生成具有上下文关联的情感反应(图6(c)-1). 概率情感模型则利用HMM等工具, 刻画情绪状态之间的概率转换, 适用于捕捉“爱恨交织”等动态变化或混合情绪的识别任务^[142](图6(c)-2). 分布式事件驱动模型基于神经网络架构, 同时处理外部环境输入与内部状态信息, 从而实现复杂或不确定交互情境的自适应响应, 具有较强的动态性与鲁棒性^[143](图6(c)-3). 量子情感模型尚处于探索阶段, 尝试借鉴量子叠加与纠缠等原理, 通过布洛赫球(Bloch sphere)等形式对复杂情绪状态进行建模, 旨在突破传统线性情感模型的表达边界^[144].

3.2.2 情感表情的评价指标

面部机器人的情感表情不仅需具备生理层面的运动能力, 更应在心理感知层面被用户认为自然、准确并具有情感价值. 因此, 情感表情的评估不应仅限于

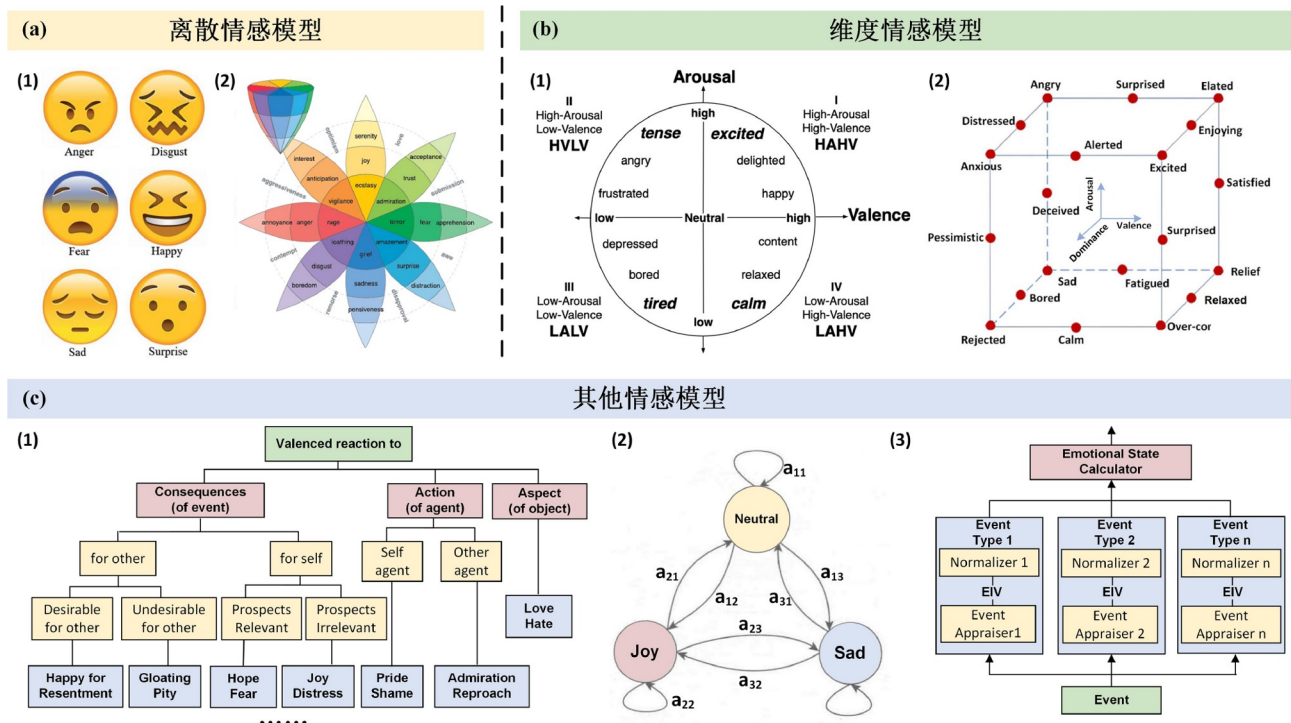


图6 (网络版彩图)不同类型情感模型. (a) 离散情感模型: (1) Ekman基本情感模型^[14]; (2) Plutchik情感轮模型^[14]. (b) 维度情感模型: (1) 二维效价-唤醒(VA)情感模型^[130]; (2) 三维愉悦-唤醒-支配(PAD)情感模型^[12]. (c) 其他情感模型: (1) 基于认知的情感模型——OCC模型^[141]; (2) 概率情感模型——HMM^[142]; (3) 分布式事件相关情感模型^[143]

Figure 6 (Color online) Different types of emotion models. (a) Discrete emotion models: (1) Ekman's basic emotion model [14]. Copyright©2022, Elsevier. (2) Plutchik's emotion wheel model [14]. Copyright©2022, Elsevier. (b) Dimensional emotion model: (1) The 2D VA emotion model [130]. Copyright©2024, The Author(s). (2) The 3D PAD emotion model [12]. Copyright©2023, The Author(s). (c) Other emotion models: (1) Cognitive-based emotion model, OCC model [141]. Copyright©2020, Springer. (2) Probabilistic emotion model, HMM [142]. Copyright©2006, IEEE. (3) Distributed event-related emotion model [143]. Copyright©2000, The Author(s).

识别结果, 而需借助系统化的指标体系对其表达效果进行全面衡量. 具体而言, 评估可从表达准确度、表达自然度、情境适合度与用户满意度4个维度展开.

表达准确度关注生成表情在情绪类别与强度上的准确性. 类别准确性方面, 可采用离散情绪识别常用指标, 如精确率、召回率、F1 分数及混淆矩阵进行评估; 强度准确性方面, 则可引入连续情绪维度(如 AU 强度或情绪在 VA 空间中的坐标), 并使用平均绝对误差(mean absolute error, MAE)、均方根误差(root mean square error, RMSE)与一致性相关系数(concordance correlation coefficient, CCC)来衡量预测强度与真实值之间的接近程度.

表达自然度衡量表情是否具有人类般的动态特征. 在主观层面, 可借鉴视频质量评估方法, 通过平均主观评分(mean opinion score, MOS)或基于成对比较的偏好测试对表情的逼真度、流畅度和生动性进行评分. 在客观层面, 则可利用运动学与生理学指标, 如面部关键点轨迹的速度、加速度与抖动, 评价表情运动的平滑性, 结合AU与面部肌肉的生物力学约束检验其合理性, 再通过眨眼频率、微表情持续时长等节律性指标, 与人类自然统计特征的吻合程度, 进一步量化自然度. 此外, 需评估唇部与下颌运动在几何形态和时间序列上与语音的一致性, 以验证发音器官之间的协调性.

情境适合度侧重考察表情在当前语境中的合理性. 从语义与表情的一致性角度, 可比较语音或文本中情绪在VA空间的位置与表情情绪的匹配度, 常用指标包括余弦相似度: 从场景与表情一致性角度, 可通过设置多种交互脚本, 由人工标注员对表情的恰当性进行评估, 并以合适率、不当表现率与纠偏成功率等指标加以量化. 在多模态一致性方面, 表情应与语音韵律相协调, 如语音高亢快速应伴随积极表情, 低沉缓慢则应体现消极表情. 可通过AU强度与语音基频、能量、语速等参数的相关性分析, 评估机器人在“声-脸同频”方面的表现.

用户满意度则从使用体验出发, 评估用户的接受度与持续使用意愿. 主观层面可采用常见量表, 如信任感、拟人化程度、社交临场感与同理感等进行评分; 客观层面则可结合会话时长、交互轮次、复访率、主动中断率与纠错请求率等行为数据反映用户偏好. 在任务型场景中, 还可引入任务完成率、任务耗时

及错误恢复效率等作为补充. 进一步地, 在伦理允许的前提下, 心率变异性、皮电反应、注视停留等生理与环境信号也可作为用户体验的补充证据.

综上所述, 构建涵盖准确性、自然性、适应性与用户反馈的多维度评价框架, 有助于全面衡量面部机器人情感表情的表达效果. 这一体系不仅能够指导情感生成算法的技术迭代, 更为人机情感交互的可解释性、有效性和人本化发展提供了坚实支撑, 推动其在各类实际应用场景中实现更高水平的智能表现与社会融合.

3.2.3 机器人情感系统

面部机器人的情感系统旨在模拟人类的情绪机制, 从而实现更自然、更具同理心的人机表情交互. 该系统在理论上融合了情绪、情绪状态(心境)、人格与记忆4个关键心理成分. 具体而言, 情绪代表一种短期、情境驱动的反应; 情绪状态则反映出持续性更强的行为倾向; 人格决定情绪状态的处理方式与表达风格; 记忆则通过回溯以往的交互与情绪经历, 为当前反应提供情境依据^[128,145].

在实际应用中, 面部机器人通常通过传感器与算法对用户的面部表情、语音语调或手势动作进行感知, 然后结合自身的内部情绪状态、人格特征与情感记忆, 生成符合当前情境的应对策略. 这一过程构成了一个闭环调控系统, 其中情绪的累积状态影响心境, 而心境又在个性与记忆反馈机制的调节下不断调整^[146]. 许多情感建模方法通过将情绪状态映射至多维情感空间, 使系统能够基于坐标位置进行动态行为选择.

为更直观地展示情感系统的实现机制, 已有多个代表性模型与案例提供了有价值的参考. Itoh等人^[147]为WE-4RII机器人设计了一个三维情感模型, 将愉悦度、激活度与确定性3个维度与协同记忆机制相结合, 以实现动态情感表达. Han等人^[148]在Russell的二维情感空间基础上, 融入“五大人格模型”中的性格特质, 使机器人能够依据不同气质类型展现出差异化的情绪行为. Ficocelli等人^[149]开发的社交机器人Brian则通过马尔可夫链建模情绪转换过程, 实现如安慰、提醒等行为的实时适应性调节. Zhang等人^[150]提出的可定制人格情感模型使用碗状情感空间模拟情绪波动, 并依据个性特征调整行为响应策略. 其他系统还包括:

Hong等人^[151]提出的双层情感模型, 上层基于隐马尔可夫模型进行决策, 下层基于规则进行快速反应; Liu等人^[152]基于同理心的建模方法, 则引入了认知重评策略来提升机器人对人类情绪的理解与共情能力。

综上所述, 这些系统展现了心理理论与计算建模的深度融合, 传统情感推理方法主要依赖离散或维度模型, 虽具逻辑清晰与计算开销小的优势, 但表达能力有限, 难以应对复杂语境与个体差异, 结果往往停留在“粗粒度”情感分类, 缺乏动态调整与个性化表达。随着应用场景的不断拓展, 这些局限日益显现, 亟需更强大的推理机制, 而这正是近年来LLM与MLLM所推动的关键变革方向。赋予面部机器人更加细腻且具适应性的情感行为。随着技术不断成熟, 具备高级情商的机器人情感系统有望在教育、医疗和社交陪伴等领域实现广泛应用, 推动人机表情交互向更深层次的情感智能演化。

3.3 面部机器人表情生成

面部表情生成作为面部机器人表情交互环路中的核心输出环节, 其主要任务是将机器人的内部情绪状态转化为可感知的外部表情行为, 以实现双向同理与自然对话。高效的表情生成不仅要准确传达情绪意图(如通过微笑表达友好), 还需协调多种输出通道, 如面部运动、眼神与唇部动作, 以避免在人机交互过程中产生认知不协调, 从而增强用户的信任感与沉浸感。当前支撑该过程的关键技术包括面部表情合成、眼神交流与面部表情-语音同步等。接下来将对这些核心技术进行详细介绍与分析。

3.3.1 面部表情合成

面部表情合成技术是实现自然人机交互的关键组成部分。通过生成恰当的面部表情, 机器人不仅能够表达自身的情绪状态, 还能增强用户的情感联结与互动参与感。当前的表情合成方法大致可分为两类: 一类基于机器人内部情感状态进行表情生成, 另一类则不依赖情感推理, 通过模仿或调用预设模板生成表情。前者可实现情感一致性响应, 适用于高情感交互需求场景; 后者则适合无需复杂情绪逻辑的应用, 如表情模仿或基础表情展示。

基于情感的表情合成方法通常将机器人内部的情绪参数(如情绪、心境、个性特质)映射为相应的面部

表情。例如, Han等人^[148]基于二维情感模型, 根据机器人内部心境与人格特征动态调整面部表情。Ishihara等人^[153]提出“动态衰减波合成”技术, 将激活水平转化为如扬眉、瞳孔扩张等动态面部动作, 以模拟惊讶情绪(图7(a)-1)。Liu等人^[152]构建了融合三维情绪空间(激活度、愉悦度、姿态)与认知重评策略的模型, 用于生成更具同理心的表情响应。这类方法依赖情感系统的实时输出, 并通过特定算法将情绪参数转化为驱动面部肌肉的具体控制指令。

在不需要情感推理的情境中, 机器人可通过模仿用户表情或调用预设表情模板生成动作。表情模仿方式依赖于对用户面部动作的捕捉与映射。Huang等人^[154]提出了结合平滑约束与LSTM的反向机械模型, 可将三维面部特征序列映射为电机控制信号, 实现85%的时空相似度。Hu等人^[43]开发的共情表达系统可提前839 ms预测并模仿用户的微笑行为(图7(a)-2)。Liu等人^[49]则设计了轻量化神经网络, 实现实时检测68个面部关键点, 进而驱动表情与头部姿态的同步控制。

相较之下, 预设表情模板系统通过代码或软件指令控制面部动作生成。例如, Mazzei等人^[155]开发的面部表情混合引擎, 允许用户基于FACS中的AU手动构建表情。Ghorbandaei Pour等人^[156]在孤独症干预中应用此类预设表达, 帮助儿童识别情绪线索。Yang等人^[30]则通过贝叶斯优化方法调节仿人机器人表情参数, 使生成结果更加贴近人类感知标准。

近年来, 为提升表情生成的适应性, 研究逐渐趋向混合式方法。例如, Hong等人^[151]提出的双层模型结合了基于HMM的情感驱动表情生成与基于规则的模仿策略, 有效应对用户的即时交互需求。展望未来, 随着MLLM的发展, 机器人有望在综合内部情绪状态、用户反馈与环境上下文的基础上, 实现更具细腻性与交互感的面部表情生成, 为人机交互注入更深层次的情感智能。

3.3.2 眼神交流

眼神交流在人机交互中扮演着重要角色。通过眼神交流, 个体可以传递情绪态度、注意力焦点以及对话反馈。在交流过程中, 注视不仅有助于加强语言信息的传递, 还可以通过目光的转移与凝视表达专注、尊重等情绪。若机器人能够有效模拟这种注视行为, 将显著提升互动的自然性与情感深度, 进而促进人与

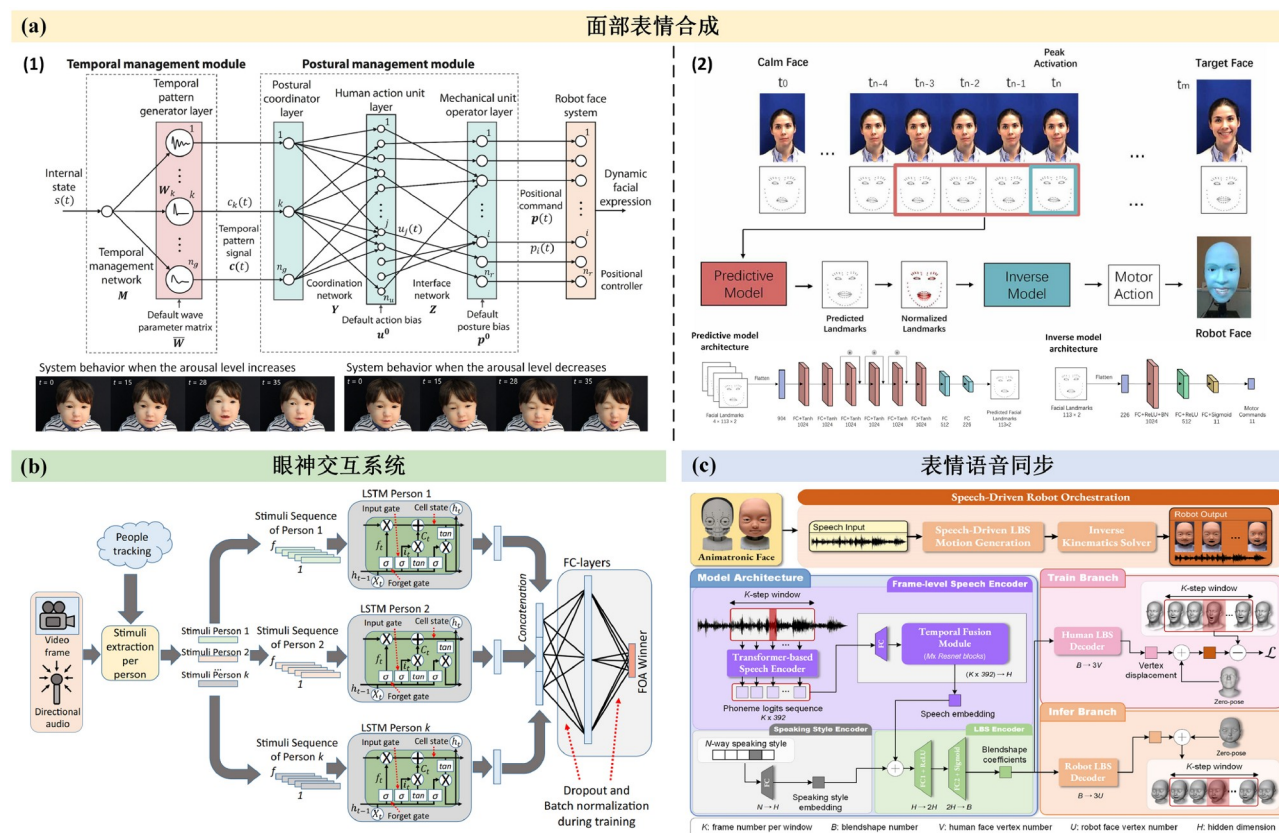


图 7 (网络版彩图)面部机器人表情生成的三大关键技术. (a) 面部表情合成: (1) 基于情感的面部表情合成, 如Ishihara^[153]提出的动态衰减波合成技术; (2) 与情感无关的面部表情合成, 包括表情模仿和预定义表情, 以Hu^[43]开发的人机面部共表情系统为代表. (b) 眼神交互系统. 一个代表性的例子是应用LSTM网络来优化和改进机器人视线控制, 如Domingo^[15]所展示的. (c) 表情语音同步. 涵盖音唇同步和表情-语音同步, 代表性案例是Li^[16]提出的语音驱动的面部表情系统

Figure 7 (Color online) Three key technologies for face robot expression generation. (a) Facial expression synthesis. (1) Emotion-based facial expression synthesis, exemplified by the dynamic decaying wave synthesis technique proposed by Ishihara^[153]. Copyright©2024, The Author(s). (2) Emotion-independent facial expression synthesis, including expression mimicry and predefined expressions, represented by the human-robot facial co-expression system developed by Hu^[43]. Copyright©2024, AAAS. (b) Eye gaze control system. A representative example is the application of LSTM networks to optimize and improve robotic gaze control, as demonstrated by Domingo^[15]. Copyright©2021, The Author(s). (c) Facial expression-speech synchronization. It covers both lip-speech and expression-speech synchronization, with a representative case being the speech-driven animatronic facial expression system proposed by Li^[16]. Copyright©2024, IEEE.

机器人之间更紧密的连接. 因此, 注视协调已成为面部机器人表情交互中的关键技术之一^[157]. 本节将从技术实现角度出发, 分析其核心方法.

注视行为首先体现为注意力焦点的表达. 例如, 在多方对话情境中, 机器人需通过注视发言者来展示其注意力, 这一功能可通过基于神经网络的注视控制系统实现. Domingo等人^[15]提出了一种基于LSTM的眼动控制系统, 通过分析声音、头部姿态等多模态刺激, 机器人能够平滑地切换注意力焦点, 优先关注主要发言者, 从而增强互动效果(图7(b)). 类似地, Yun^[158]提出一种混合感知模型, 使机器人能够根据会话角色与

空间位置选择合适的注视对象. 该方法借助多模态传感器(如Kinect骨骼跟踪与麦克风阵列)获取会话角色与空间信息, 使机器人在动态环境中表现出更具人性化的注视行为.

在协作任务场景中, 机器人的注视不仅需表达自身意图, 还需理解人类的动作信号. Raković等人^[159]提出了注视对话模型, 通过分析用户注视点的时间序列(如频繁注视工具A可能暗示下一步操作), 推断用户任务意图并调整自身注视策略. 例如, 在装配任务中, 当用户注视螺丝刀时, 机器人同步转向相应部位并点头提示, 从而降低操作出错率. Palinko等人^[160]进一步证

实,与头部姿态跟踪相比,眼动追踪可更准确地区分用户的注意对象(如区分注视屏幕或键盘),提升了27%的协作效率。

注视不仅是信息传递的工具,也可作为情绪表达的载体。Hashimoto等人^[161]开发的情感注视引导系统,将面部表情与头部动作结合,机器人通过“惊讶”表情引导用户注视特定目标,其反应速度比无表情状态提高了40%。Iimori等人^[162]研究了眨眼在注视引导中的作用,并设计了眨眼同步机制,机器人在对话停顿期间以每分钟15~20次的频率模拟人类眨眼节奏,显著提升了用户对语音内容的理解与亲和感知。这些研究表明,情绪表达不应仅停留于面部肌肉变化,还应包括眼部注视的微妙调控。

此外,人类的注视行为还涉及头部与眼球的协调运动,如扫视(saccades)与前庭-眼反射(vestibulo-ocular reflexes, VOR)。Pan等人^[163]提出的仿生注视架构模拟此机制,通过分层行为控制与运动带宽差异(眼动速度快于颈部)实现类人的动态响应。Roncone等人^[164]进一步设计了笛卡尔6自由度控制器,融合惯性传感器与视觉数据,使机器人头部与眼球能在三维空间中高精度追踪注视点,误差控制在 0.5° 以内。该技术结合仿生算法与高精度驱动器,有效缓解了“恐怖谷”效应,使注视动作更自然真实。

总体而言,注视协调技术涵盖从基础注视控制到意图推理再到复杂情绪表达的多个层面。通过对这些技术的不断优化,面部机器人将逐步接近人类的交流方式,显著提升交互的自然性与有效性。随着相关技术持续进步,未来的机器人将在注视行为上更加灵活、智能与人性化,进一步推动人机情感交互的深入发展。

3.3.3 面部表情-语音同步

面部表情与语音的同步是面部机器人中的核心技术之一,其目标在于通过将语音信号与相应面部动作进行精确对齐,消除“声画不同步”带来的认知失调,从而实现自然可信的情感交互。“McGurk效应”正是这一问题的经典体现:当个体听到音节“ba”,却看到嘴型为“ga”时,大脑会整合这两种冲突信号,最终产生“da”的错觉感知^[165]。这一现象凸显了面部动作,尤其是唇形与语音对齐的重要性。本节将分别探讨音唇同步与表情-语音同步两项关键技术。

在机器人唇语同步中,最常见的方法基于发音规

则映射机制,即将语音中的音素映射至相应的唇形(vi-seme),再驱动机器人执行预设的唇部动作。该过程将语音解构为基础发音单元,并通过指令控制唇部肌肉的位移。该方法结构简单、易于实现,且具备良好的实时性。例如, Lin等人^[166]提出了一种简化的唇语同步系统,将音素与预定义唇形进行匹配,大幅降低了机械控制复杂度。为提升自然性,Strathearn和Ma^[167]提出了基于语音频谱特征的机器学习唇部运动模型,协调下颚、嘴唇与舌头动作,实现了79.4%的音节同步精度。Madumal等人^[168]则使用唇形库,构建低延迟同步系统。尽管此类方法在控制精度上具有优势,但在适应多语言环境、表达复杂情绪与应对语调连续变化及个性化发音方面存在局限,影响了其灵活性与整体表达准确度。

随着深度学习的发展,研究者逐渐借助数字人建模技术,实现从语音驱动的面部表情生成,突破了传统唇语同步的限制,朝向完整语音-表情同步迈进。此类方法通常通过wav2vec2等语音识别模型提取音频特征,再结合逆运动学计算生成相应的面部动作。例如,Heisler等人^[169]提出基于网格(mesh)的音频-面部映射方法,将音频输入映射至三维点云的顶点坐标,再将其转换为机器人面部肌肉控制信号。Li等人^[16]提出了“以蒙皮为核心”的驱动框架,采用线性混合蒙皮技术将语音信号转化为面部皮肤实时变形参数,显著提升了表情的准确性与真实性(图7(c))。Yu等人^[170]则利用GAN合成与语音同步的三维面部动作序列,并通过逆运动学方法将其映射至物理机器人。这类数据驱动方法依赖端到端学习,能捕捉语音与面部运动之间的非线性关系,具备出色的多语言适应性与个性化表达能力。

总体而言,基于发音规则的唇语同步方法具有良好的实时性能与实现简便性,而基于数字人技术的表情-语音同步方法则在表达自然性与精度方面更具优势,尤其适用于情绪表达与复杂语音内容的同步处理。展望未来,随着多模态感知与语音建模技术的持续发展,机器人有望实现语音、上下文与用户表情的实时整合与自适应同步,有效突破“恐怖谷”效应,进一步增强人机交互的真实感与自然性。在表情生成方面,传统合成方法虽然具备稳定性高、实现难度低等优势,但在自然性、流畅度和细腻度方面仍存在显著不足,尤其在实时驱动和多表情切换场景中常表现出僵

硬的问题. 尽管其计算和控制成本较低, 但难以满足高拟真交互的需求, 因此, 如何突破传统生成方式的局限, 成为先进技术发展的关键方向.

4 面部机器人表情交互的先进LLM驱动路径

在第3节所建立的框架基础上, 本章提出了以LLM驱动的方法, 实现了对传统方法的突破. 首先, 通过多模态融合整合视听及触觉输入, 提升感知能力, 提供更丰富的上下文信息. 关键在于, LLM通过直接管理情感推理, 革新了后续处理流程, 避免了对显式情感计算模型的依赖. 同时, 基于智能体的生成采用语义规划, 实现动态且具记忆增强的表情合成. 该整体集成方案在促进人机交互的自然性方面达到了前所未有的高度. 两种方法的交互框架细节对比如表4所示.

4.1 多模态情感识别

多模态情感识别(multimodal emotion recognition,

MER)是实现人机表情交互的关键技术之一. 通过整合来自语音、视觉、文本与生理信号等多种通道的数据, 面部机器人能够更细致地理解人类情绪, 从而弥补单一模态系统在处理如面部遮挡或音频噪声等情境中的局限性, 显著提升情绪分类的准确性与鲁棒性^[171], 使机器人做出更具同理心的回应, 提升交互自然性并建立用户信任^[172].

MER的实现过程通常包括4个步骤(图8(a)): 首先, 系统通过摄像头、麦克风及生理传感器采集面部表情、语音、文本等多模态数据; 其次, 从每种模态中提取与情绪相关的特征; 然后, 对这些特征进行融合, 构建统一的情感表示, 以弥补单一模态中存在的噪声或歧义; 最后, 使用分类器或预测模型对融合后的特征进行情绪判别, 并根据识别结果驱动机器人的行为响应. 常用的输入信号包括视觉线索(如面部表情、身体姿势)、音频特征(如音调、语调)、文本内容(如对话语义)与生理数据(如脑电、皮肤电), 而输出形式包括离散情绪标签(如愉快、愤怒)或连续情绪维度(如愉悦

表 4 面部机器人表情交互感知-推理-生成框架对比

Table 4 Comparison of face robot expression interaction in the “perception-inference-generation” stage

环节	对比维度	传统方法	先进方法
感知识别	技术核心	单模态或早期/后期多模态融合. 视觉(CNN/OpenFace)、语音(MFCC/LSTM)分别处理, 在特征或决策层简单融合	跨模态对齐与融合. 利用MLLM的嵌入空间, 将不同模态信息映射到统一语义空间, 实现深层互补
	代表性工作	视听融合 ^[59] ; 面部表情识别 ^[62]	MEmoBERT: 基于Transformer的多模态情绪识别 ^[175] GPT-4V: 通过视觉-语言上下文提示理解情绪 ^[185]
	核心挑战	模态异构性: 不同模态在时间、空间上的对齐难题 信息冲突: 当语音和表情传递矛盾情绪时, 难以可靠裁决	幻觉与偏见: 模型可能生成与输入不符的“合理”推断 计算开销: 多模态编码与推理对算力要求极高, 影响实时性
情感推理	技术核心	依赖预定义的情感模型(如VA、PAD), 通过规则或HMM进行状态推理	基于LLM的语义理解. 将多模态输入作为上下文, 利用LLM的常识、因果推理和情境建模能力进行深层次情感归因
	关键能力	情绪状态分类; 基础的情绪强度估计	情感记忆: 记忆用户历史情绪状态, 实现个性化交互 上下文理解: 理解对话背景, 区分复杂情境 实时对话规划: 生成共情回应并规划后续对话策略
	代表性工作	OCC情感模型 ^[140] ; HMM概率情感模型 ^[142]	CoT: 逐步推理用户情绪原因 ^[180] SoR-ReAct: 将推理与行动结合, 调用工具API ^[20]
	核心挑战	僵化与脆弱: 无法处理预定义模型之外的复杂或混合情绪 缺乏上下文: 难以利用长时对话历史	可解释性: LLM的“黑箱”决策过程难以追溯和验证 文化敏感性: 可能放大训练数据中的文化偏见
表情生成	技术核心	直接映射. 将推理出的情绪标签或维度值, 通过查找表或函数映射转换为预定义的表情参数(AU组合)	语义规划. LLM Agent输出高层次行为指令, 再由底层控制器解析为具体的电机控制指令
	代表性工作	面部表情混合引擎: 手动设计AU组合 ^[155] , 贝叶斯优化参数 ^[29]	Alter3: LLM直接生成Python代码控制机器人动作 ^[19] Nadine: LLM驱动生成表情、注视点和语言 ^[20]
	核心挑战	不连贯与不自然: 表情之间切换生硬, 缺乏细腻的过渡 “恐怖谷”效应: 因硬件限制导致的轻微不自然会被放大	指令-执行差距: LLM规划的高层指令可能与底层硬件的实际能力不匹配, 需进行可行性验证

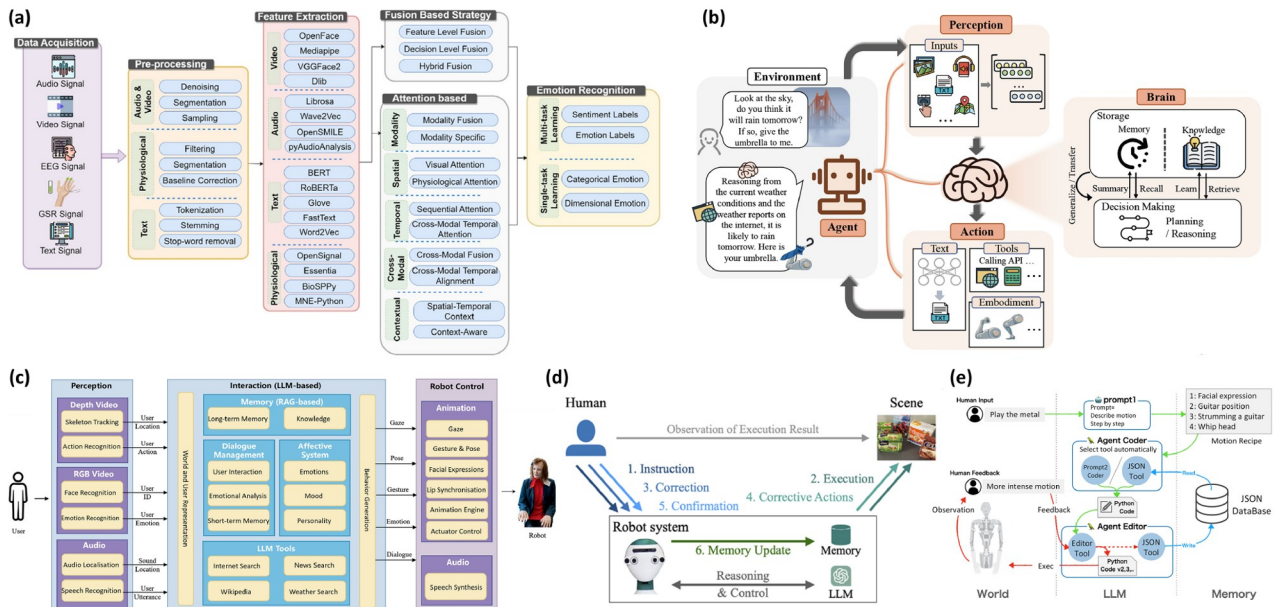


图 8 (网络版彩图) AI驱动的高级表情交互技术。(a) 基于深度学习的多模态情感识别通用流程^[173]。(b) LLM代理概念框架, 由大脑、感知和行动3个部分组成^[197]。(c) 集成了LLM代理的Nadine社交机器人系统^[20]。(d) ARMAR-6人形机器人通过交互体验逐步学习其行为^[202]。(e) 人形机器人Alter3文本转动作生成系统^[19]

Figure 8 (Color online) AI-driven advanced expression interaction technologies. (a) A generic deep learning-based pipeline of multimodal emotion recognition^[173]. Copyright©2023, Elsevier. (b) A conceptual framework of an LLM-based agent consisting of three components: brain, perception, and action^[197]. Copyright©2025, Science China Press. (c) The Nadine social robot system integrated with an LLM-powered agent^[20]. Copyright©2024, The Author(s). (d) ARMAR-6 humanoid robot incrementally learns its behaviors through interactive experiences^[202]. Copyright©2024, The Author(s). (e) A text-to-motion generation system for the humanoid robot Alter3^[19]. Copyright©2025, The Author(s).

度、唤醒度)^[173]。

多模态情绪识别系统通常按融合模态数量划分为双模态与三模态系统。双模态系统中最常见的是视听融合, 例如, 使用CNN提取面部特征、LSTM处理语音模式, 再通过决策级融合策略提升“愤怒”与“悲伤”等相似情绪的区分度。语音-文本组合则结合音高、能量等语音特征与语义信息, 用于检测讽刺、隐性情感等细微表达。视觉-文本融合也具有较高的应用价值, 如将面部AUs与字幕中的情感词汇对齐, 可提升视频情绪分析的精度^[174]。近年来, 三模态情绪识别(视听文本融合)因其对用户情绪具有更全面的解析能力而受到广泛关注。尽管该方法在系统设计与计算开销上更为复杂, 但其在识别准确性与交互自然性方面具备显著优势。已有研究表明, 如双向门控循环单元(bidirectional gated recurrent unit, Bi-GRU)、基于变压器的多模态情感识别双向编码表达(multimodal emotion recognition bidirectional encoder representations from transformers, MEmoBERT)与跨模态翻译(translation

between modalities, TransModality)等模型, 通过整合模态间的互补情绪信息, 可获得更细腻的情感评估结果^[175]。

针对不同模态在数据结构与时间序列方面存在差异的问题, 研究者提出了多种融合策略。特征级融合(早期融合)通过将原始或低层特征(如音频的MFCC、图像的HOG描述)拼接为联合特征向量进行情绪分类。该方法能保留模态间的相关性, 但存在维度爆炸与噪声敏感等问题。决策级融合(后期融合)则先独立处理各模态, 再通过平均、加权或投票等规则整合结果, 具有良好的模块化与鲁棒性, 但可能忽略模态间深层次关联。模型级融合利用注意力机制、混合网络等结构构建联合表示空间, 能够刻画复杂的跨模态依赖关系, 但对模型设计与训练数据需求较高。混合融合策略结合早期与后期融合的优势, 在保留模态间细粒度关联的同时增强系统鲁棒性, 但系统复杂性较高, 易对实时性构成挑战^[14,152,176]。

尽管MER技术已取得重要进展, 但仍面临一些挑

战,如高质量数据集匮乏、模态间时序对齐困难与计算资源消耗大等问题。新兴解决方案包括:利用GAN与联邦学习缓解数据短缺;采用如移动网络MobileNet的轻量级模型实现低延迟部署;针对跨文化情绪差异,引入元学习等动态融合机制^[177]。随着视觉、语音与文本的深度融合,多模态情绪识别正成为提升人机表情交互精度与适应性的关键途径。相较传统方法,多模态感知在情绪识别的准确性、鲁棒性及情境理解能力方面表现更优,能够在复杂环境中保持稳定输出。然而,其对大规模训练数据与高算力的依赖显著提升了资源成本。未来研究需在保障性能的同时,探索轻量化建模、知识蒸馏等手段,以降低资源消耗,推动该技术更广泛地落地应用。

4.2 LLM驱动的情感推理

基于LLM的情绪推理近年来已成为情感计算领域的变革性范式,凭借其在情绪识别与理解方面的高级能力,正逐步被集成于面部机器人与各类人机交互系统之中。与传统多模态情绪识别方法不同,后者主要通过融合语音、文本与视觉数据来进行情绪分类,而基于LLM的情绪推理则通过语言逻辑推理能力,将多模态输入与上下文语义融合,从表层识别向深层理解转变,显著提升了机器人在情感交互中的深度与适应性^[178]。

通过融合LLM的语言推理优势与MLLM在视觉与听觉感知上的能力,机器人不仅能够更准确地解释复杂的情绪状态与动态情境,还能够借助上下文建模、因果推理等机制,模拟类人情绪认知过程。这种能力使机器人在交互中表现出更高水平的个性化、自然性与同理心^[179]。

基于LLM的情绪推理流程主要可划分为3个核心阶段:情境理解、情绪识别与情绪生成。首先,模型通过分析对话历史与多模态上下文线索,构建对当前交互场景的全面认知;接着,结合情绪识别算法与链式思维提示(chain-of-thought, CoT)机制^[180],逐步推理用户的情绪状态及其潜在原因,展现出基本的“换位思考”能力;最后,综合内部情绪建模、用户情境与提示引导,生成富有共情与情感智能的回应^[181,182]。例如,Mishra等人^[183]展示了GPT-3.5能够通过分析对话历史实时预测机器人的情绪反馈,显著提升用户交互体验。

从技术路径来看,基于LMM的情绪推理可分为3

类:纯语言推理、视觉-语言联合推理与全多模态动态推理,分别对应传统文本LLM、视觉语言模型(vision large model, VLM)与MLLM的应用框架。

第一类方法仅依赖文本输入,利用生成式预训练变压器(generative pre-trained transformer, GPT)等LLM的强大语义理解能力,通过上下文建模与CoT提示完成情绪推理。该方法擅长捕捉语言互动中的细微情绪变化,但在处理如面部微表情或语调等非语言线索方面存在局限^[184]。

第二类方法融合视觉与语言模态,如GPT4代-视觉模型(GPT-4Vision, GPT-4V)等模型采用上下文提示策略,将语义生成与视觉输入动态匹配^[185],从而实现面部表情、身体姿态与语言信息的联合分析。尽管该类模型在静态图像情绪分类中接近人类水平,但在处理情绪时序变化方面仍具挑战^[186]。

第三类方法则整合视觉、听觉、文本乃至生理信号,构建端到端情绪推理系统。Arjmand等人^[187]提出的共情基础模型在分析语音与面部表情的基础上实时生成共情回应,情绪感知评分显著高于传统方法。R1-Omni模型进一步结合强化学习与可验证奖励机制,在推理过程中不仅实现因果链条建模与情绪解释,还能输出结构化推理路径,表现出高度的泛化能力与可解释性^[188]。

相较传统情绪识别系统,基于大模型的情绪推理在泛化能力、情境理解与交互自然性方面展现出明显优势。例如,使用GPT-4V等大型预训练模型无需微调即可获得接近顶尖水平(state of the art, SOTA)的性能^[189],并支持因果推理、微情绪识别与情境适配反馈等复杂能力,显著丰富了人机情感交互维度^[190,191]。

然而,该技术仍面临一些挑战,包括非语言信息整合困难、模型解释性不足与计算资源消耗过大等^[192,193]。例如,GPT-4V对微表情识别能力有限,其黑箱式推理过程也可能掩盖潜在的文化偏见^[194]。尽管由LLM或MLLM驱动的推理方法突破了传统模型在表达能力方面的局限,在上下文理解、个性化回应与生成灵活性等方面展现出显著优势,但其庞大的模型规模与较高的推理延迟,仍对训练与部署提出了极高的资源要求,同时存在可解释性不足等问题。未来研究应在提升模型性能的同时,聚焦于数据去偏、模型压缩与蒸馏、可解释性增强等关键方向^[195],以在性能与

效率之间寻求平衡, 并加强对系统安全性与透明性的保障。

4.3 智能体驱动的情绪到行动框架

尽管多模态情绪识别与基于LLM的情绪推理为情感感知交互奠定了基础, 但若要实现真正“类人”的面部机器人, 还必须进一步突破“理解”层面, 迈向“表达”与“行为”层面的联动。面部表情并非单纯的肌肉运动, 而是一种具备情境感知与认知驱动的响应行为, 因此, 如何将内部情绪状态有效转化为外部行为, 是实现人机同理互动的关键。然而, 现有的LLM在情绪行为生成方面仍面临诸多限制, 如容易生成幻觉(hallucination)、知识更新滞后, 以及难以维持复杂多轮交互中的行为连贯性^[196]。这些问题促使研究者引入智能体(agent)技术, 通过集成控制流程、记忆机制与工具接口, 支持面部机器人实现一致、具目标导向的表情行为。因此, 智能体架构为“从情绪到行动”的全流程建模提供了可行路径。

在实际系统中, 实现“情绪-认知-行为”转换的机器人通常采用基于智能体的系统架构。该架构整合情感理解、计划、记忆与目标导向行为, 将高层次的心理状态(如信念与意图)与底层的运动执行连接, 形成面部机器人中的“心智-行为”闭环^[197]。典型智能体系统包含3个核心模块: 感知模块(用于解读多模态输入如表情、语音与手势)、认知控制中心(用于记忆管理与决策推理)以及行为执行模块(包括面部肌肉控制、语言响应或工具和应用程序接口(application programming interface, API)调用等)(图8(b))。这三大模块共同赋予了智能体计划能力、记忆机制与工具使用能力, 构成情绪行为生成的核心基础。以下将逐一分析这3项能力的应用与发展。

首先, 计划能力使智能体能够将复杂任务分解为可执行的子任务, 并生成与长期目标相一致的行为序列。其中, “CoT提示”是一种常用的任务分解技术, 鼓励模型通过逐步推理解决复杂问题。LLM+P框架进一步结合LLM与符号规划器, 借助规划域定义语言将自然语言任务转化为可执行计划^[198]。在更复杂的交互中, 自我反思能力成为智能体改进行为策略的重要机制。例如, 思考和行动(reasoning and acting, ReAct)框架^[199]通过交替进行语言推理与动作执行, 使智能体可调用外部工具并记录自然语言推理轨迹。Nadine机器

人采用的社交机器人ReAct(social robotics reasoning and acting, SoR-ReAct)系统即集成了RAG记忆、工具调用与情感模块, 能生成自然的注视、姿态、手势、情绪与对话行为(图8(c))^[200]。在此基础上, 反思(reflexion)框架引入“反思-修正-迭代”机制, 记录成功与失败的行为轨迹并生成自我反馈, 从而实现类似人类的经验学习与行为优化^[200]。

其次, 记忆机制是支撑人机交互连续性与个性化的关键。短期记忆使智能体能即时处理上下文, 实现对话连贯与快速场景切换^[201]; 长期记忆则支持跨会话的用户识别与情感一致性回应。例如, Nadine系统中的情节记忆模块可识别回访用户, 并根据过往情绪交流调整当前行为^[201]。Bärmann等人^[202]提出的一种增量学习架构使LLM可通过用户反馈不断优化行为策略, 并将交互历史存储为可复用内容, 模拟人类对社交经验的长期保留与泛化(图8(d))。相关研究指出, 具备记忆能力的智能体是实现情绪智能系统的关键组成^[203]。

最后, 工具使用能力显著扩展了智能体的边界, 使其超越预训练知识, 在交互中访问外部数据库、执行代码、控制设备或获取实时信息, 从而缓解幻觉现象与知识老化问题。例如, ToolLLM框架允许智能体根据任务需要, 动态调用超过16000个真实API^[204], 实现高度灵活的任务响应。Cai等人^[205]更进一步提出LLM可自主生成并缓存可复用工具, 提高任务效率。在面部机器人中, 工具调用包括访问表情数据库、调度表情驱动指令或调用生成模型生成特定情绪表达。代表性案例Alter3机器人展示了如何通过LLM Agent联合提示生成、代码执行与记忆编辑, 将用户输入实时转化为多模态行为反馈, 实现复杂人机互动(图8(e))^[19]。

总体而言, 计划、记忆与工具使用三大核心能力的融合使智能体得以完成“情绪到行为”的完整链路, 使面部机器人能够感知、推理并以社会化、情感化方式进行表达与响应。展望未来, 智能体系统正朝向更分布式、协作型结构演进。例如, Busoniu等人^[206]提出的多智能体系统实现了任务角色分工、记忆共享与联合决策。MetaGPT框架^[207]进一步将专业分工引入智能体结构, 而模型上下文协议^[208]提供了高效的信息交换机制。随着这些技术与情绪推理能力的融合发展, 下一代面部机器人有望实现具有情绪理解力、表达能力与个性化适应性的“共情型人工智能”。

智能体式的表情生成方法通过引入规划、记忆和

工具使用,极大增强了交互的自主性与适应性。但其系统复杂度高,控制开销大,稳定性与可预测性不足。未来的发展趋势是将智能体的灵活性与传统方法的稳定性结合,构建兼具自然性、可控性与资源效率的混合式生成框架。

综上所述,与传统方法相比,LLM驱动路径的先进性并非简单地体现在感知精度或生成速度的提升,而是实现了一次范式转换:从基于规则的、模块化的情感计算转向了基于理解的、端到端的社交智能。其核心优势在于赋予了机器人情境化的理解力、个性化的记忆能力和主动的交互规划能力,从而显著提升了交互的适应性与真实感。然而,这一路径也带来了可解释性、计算效率与伦理安全等方面的新挑战。具体的表情交互关键技术总结见表5。

5 AI驱动的面部机器人应用前景

随着AI驱动的面部机器人技术快速发展,其在各行业中的应用潜力正在逐步显现,亟需系统评估其现

状与未来前景。尽管该类技术仍处于早期阶段,目前主要应用于酒店接待、展览展示、特殊医疗与教育等特定场景,但其在重塑社会结构、重定义劳动力关系,乃至日常生活深度嵌入方面的潜力,已引起广泛关注。面部机器人凭借其通过表情传达情绪与社交线索的能力,在建立亲和感与信任方面具备独特优势。本节将围绕4个关键领域-教育、医疗、陪伴与服务-探讨AI驱动的面部机器人的应用场景与市场前景(图9)。

(1) 教育领域

在教育场景中,面部机器人有望深刻变革传统教学方式。目前它们已被广泛应用于科研实验、幼儿早教与特殊教育等领域。例如,在人机交互研究中,具备情感表达能力的“小瑶”机器人被用于研究人类在面对机器表情时的沟通反应;在幼教场景中,Kaspar机器人通过具情绪响应的交互机制提升幼儿语言与读写能力^[209];远程教育方面,AI教师机器人(T-bots)可依据学生的实时参与度动态调整教学内容^[210];特殊教育中,Kiddo机器人通过模拟人类动作辅助儿童进行书写与动作训练^[211],而社交机器人则在孤独症干预中改善儿

表 5 面部机器人表情交互关键技术总结

Table 5 Summary of key technologies for face robot expression interaction

技术环节	关键技术	代表性机器人与案例	实现水平/特点	现有局限
多模态感知	视觉情感识别	Sophia(Hanson Robotics ^[8]) Ameca(Engineered Arts ^[9])	集成OpenFace等视觉工具箱,可实时检测用户面部表情、注视方向,识别基本情绪类别	依赖良好光照条件;对微表情、复杂混合情绪识别能力有限
	语音情感识别	ERICA(Glas等人 ^[28])	能够从语音中提取韵律特征,识别说话者的情绪状态(如高兴、悲伤),并作为对话回应的输入	多语言、多方言的适应性不足;在嘈杂环境下的鲁棒性较差
	多模态融合	Nadine(Kang等人 ^[20]) Alter3(Yoshida等人 ^[19])	结合视觉、听觉和上下文信息,利用MLLM进行高级语义融合,实现对用户情绪更稳健的理解	系统延迟相对较高;多模态数据的时间同步仍是一个工程挑战
情感推理	上下文情感推理	Nadine(Kang等人 ^[20])	具备长期情感记忆模块,能记住过往互动中的用户情绪状态,实现跨会话的个性化情感交互	推理过程不可控,可能存在“幻觉”;计算资源消耗大,影响实时性
	LLM驱动决策	Alter3(Yoshida等人 ^[19])	利用GPT-4直接将自然语言指令转化为机器人控制代码,生成复杂行为序列	生成的动作可能不符合物理约束,需安全校验;缺乏真正的情感理解,是语义层面的模仿
表情生成	逼真表情合成	Geminoid(Nishio等人 ^[27]) Sophia(Hanson Robotics ^[8])	采用气动或电机驱动,配备高弹性仿生皮肤,能流畅呈现6种以上基本表情,逼真度高	表情库有限,难以生成预设之外的复杂表情;机械噪声和功耗问题依然存在
	共情与模仿	EMO(Hu等人 ^[43]) HHU Robot(Liu等人 ^[49])	可提前839 ms预测用户的微笑并同步模仿,实现高度自然的人际表情协调	模仿的主动性不足,主要是对清晰外部信号的响应;无法模仿复杂或负面情绪
	注视与唇语同步	Ameca(Engineered Arts ^[9]) BIGAI Robot(Li等人 ^[16])	具备高精度眼球运动机构,实现逼真注视行为;唇形与语音同步精度高,有效避免“McGurk效应”	注视行为有时显得机械;快速唇动时可能暴露机械结构的局限性

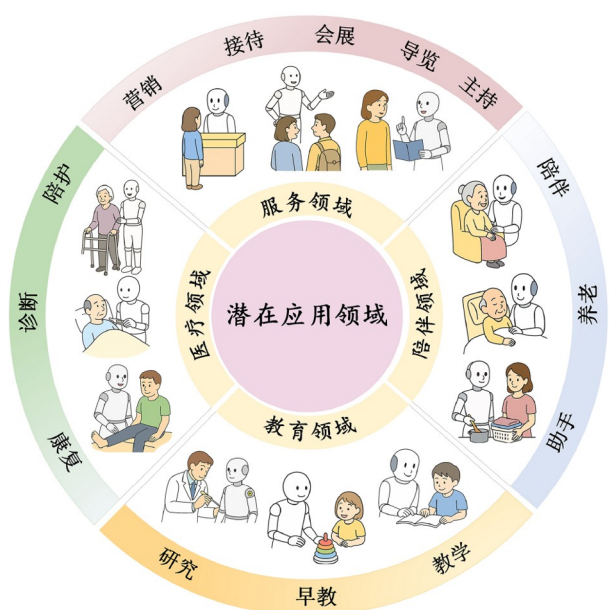


图 9 (网络版彩图) AI驱动的面部机器人的潜在应用领域
Figure 9 (Color online) Potential application areas of face robots.

童的感官处理能力与社交表达^[212]。这些案例表明, 面部机器人正推动教育的情境感知化、情感响应化与个性化发展。

(2) 医疗领域

医疗行业是面部机器人应用的另一重要领域, 尤其在康复治疗、情绪干预与陪护诊疗方面前景广阔。具有表情能力的机器人可作为看护者, 辅助孤独症、痴呆、社交焦虑、抑郁症与智力障碍等患者。例如, Chen等人^[213]的研究表明, 面部机器人通过情感共鸣交互可有效缓解老年人的抑郁情绪。在康复场景中, 机器人通过微表情与语音反馈引导患者完成训练过程, 提升治疗顺应性; 在辅助诊断中, 机器人识别患者情绪反应可为医生提供补充性评估依据。此外, 在长期护理机构中, 面部机器人不仅能提供基础照护, 还可承担情绪陪伴者角色, 显著提升患者的心理健康水平^[214]。在心理康复治疗中, 面部机器人展现出独特的潜力。由于其具备非评判性与高可控性的交互特征, 机器人能够在患者不愿与人类咨询师直接交流时, 提供一个安全的情感宣泄与沟通渠道。例如, 带有温和表情与情绪反馈的机器人可在抑郁、焦虑、孤独症谱系障碍等干预中发挥辅助作用, 帮助患者更易表达情绪并建立交流信任。此外, 机器人通过表情与语音语调

的结合, 可在心理康复训练中起到正向引导作用, 提升患者的情绪调节能力。与传统心理健康服务相比, 表情机器人具有可扩展性与可持续陪伴优势, 尤其在资源有限的地区, 可为基层人群提供普惠化的心理健康支持。

(3) 陪伴领域

机器人陪伴是近年来发展迅猛的方向, 涵盖家庭助理、老人陪护, 甚至扩展至情感与亲密关系领域。在家庭场景中, 面部机器人可作为个性化助理或情绪支持者, 缓解孤独感并增强情感联系。例如, PARO机器人通过表情与触觉反馈实现情绪疗愈, 被用于安抚老年人情绪与减轻焦虑^[215,216]。值得注意的是, 尽管陪伴机器人的研究多聚焦于老年照护与家庭辅助, 近年来其在亲密关系与性别角色模拟方面的应用也逐渐受到关注。Belk^[217]指出, “人工情感伴侣”的兴起已进入亲密交往领域, 具AI情绪算法的仿人机器人被作为模拟“爱与共情”的伙伴投入市场, 挑战了人机关系的传统边界与社会伦理。

(4) 服务领域

在服务行业中, 面部机器人正逐步成为提升顾客体验与运营效率的重要工具。它们广泛应用于客户接待、导览讲解、智能客服与娱乐互动等场景。面部机器人通过类人的面部表情与手势动作创造亲和氛围, 增强用户沉浸感。

在智能客服场景中, 相比传统客服系统, 面部机器人可借助微笑、点头等非语言行为表达理解与共情, 有效缓解用户焦虑, 营造更友好的交流氛围。在投诉处理和危机应对中表现出更高的稳定性与耐心, 有助于维系客户信任、避免服务冲突升级。例如, 在银行、保险、零售等行业, 面部机器人不仅具备业务引导功能, 还能够感知并响应用户情绪, 进一步提升服务的个性化与亲和力。

在娱乐领域, 具备表情生成与情感互动能力的机器人则推动了沉浸式游戏、虚拟偶像和舞台表演的创新。例如, 它们可在主题乐园中担当表演者, 成为虚拟角色具象化身, 或在直播平台中作为虚拟主播, 通过动态表情调动观众情绪, 从而实现更高的观众黏性和商业价值。

此外, 在如主持播报、展会导览、市场促销^[218]与表演艺术(如唱歌)^[219]等细分领域, 机器人亦展现出独特优势。在文娱传媒领域, 随着AI机器人逐渐涉足营

销传播, 其在内容生成与人机交互营销中的角色也日益凸显^[220].

尽管AI驱动的面部机器人在教育、医疗、陪伴、服务等领域展现出广阔前景, 但其大规模推广仍面临诸多社会性挑战. 首先, 在用户接受度方面, 不同年龄、文化和社会背景的群体对表情机器人的态度差异显著, 部分用户可能因“恐怖谷效应”而产生排斥情绪, 这对其社会普及构成阻碍. 其次, 在伦理风险与隐私保护层面, 面部机器人在情绪识别和数据采集过程中不可避免地涉及个人敏感信息, 如果缺乏透明的算法机制和严格的监管框架, 可能导致用户隐私泄露与伦理困境. 对此, 应通过跨学科研究推动政策、伦理与技术的协同发展, 确保机器人应用既能促进人类福祉, 又不损害社会价值.

6 AI挑战和未来发展方向

6.1 面部机器人的核心挑战

为了系统识别面部机器人的核心挑战, 必须首先明确其主要功能: 促进人机交互. 该交互过程既包括单个面部模块内的互动, 也涵盖整个机器人系统的协调. 因此, 挑战被划分为5个相互关联的层面“感知、推理、表达、反馈、系统集成”, 既涵盖局部过程, 也涉及更广泛的系统动态. 这种双重视角不仅揭示了独立的技术难题, 也揭示了各组件间复杂的相互作用, 为未来研究提供了全面的指导框架.

6.1.1 感知层

当前, 面部机器人的感知系统存在严重限制, 削弱了其交互能力, 主要体现在以下4个方面.

(1) 感官模态缺失: 缺乏与生物相似的触觉、嗅觉和热感知, 使机器人难以捕捉细微的人类触摸手势, 这对丰富的社会交流至关重要.

(2) 跨模态处理分散: 视觉、听觉和力反馈数据各自独立处理, 缺乏有效整合, 导致环境理解割裂.

(3) 瞬态信号解析不足: 无法动态识别如微表情等短暂信号, 且难以区分真实情感与社交动机驱动的表情(如礼貌或讽刺的微笑).

(4) 语义接口缺失: 缺少与高阶认知的标准接口, 原始传感器数据被视为简单的物理变化, 未能转化为有意义的信号, 形成语义瓶颈.

这种瓶颈削弱了后续的推理与表达层, 致使整个交互环节基础薄弱.

6.1.2 推理层

推理层负责将感知数据转化为符合社会文化的情感表达, 但存在多项关键限制.

(1) 情感理解浅显: 现有系统依赖离散的情绪类别, 如6种基本情绪, 限制了对复杂混合情感(如哭笑交加)的理解和表达.

(2) 文化背景忽视: 系统常忽略特定文化的表达规范, 产生不合适或令人不安的表情, 降低用户舒适度和接受度.

(3) 时间预测失败: 推理模块难以预测互动时机, 无法提前生成与对话趋势匹配的微表情, 导致表情延迟或不匹配.

(4) 认知框架孤立: 缺乏融合个体特征、情境因素和社会规范的综合认知模型, 且跨学科(社会学、人类学、文化心理学)视角整合不足.

这些问题使推理层脆弱, 限制了系统生成自然且符合语境的情感表达能力.

6.1.3 表达层

表达层因机械、材料及控制方面的限制, 难以将推理转化为真实的面部表现.

(1) 驱动器尺寸限制: 当前基于电机的驱动器体积大、重量重且空间利用率低, 难以准确复制如口轮匝肌等关键肌肉动作.

(2) 驱动器精度限制: 在实际应用中机器人表情效果往往受到硬件与系统集成的制约. 例如, 在唇语同步中, 伺服器的精度直接影响嘴型与语音的一致性. 当伺服器存在微秒级延迟时, 用户会明显察觉到不协调感, 从而降低沉浸体验.

(3) 皮肤真实感缺失: 现有合成面部皮肤缺乏关键功能, 如触觉反馈、温度调节和适应性色彩变化(如脸红、出汗及弹性变化), 导致表情不自然.

(4) 动态协调失调: 驱动器响应延迟、驱动频率低及协调不足, 影响微表情的流畅执行.

6.1.4 反馈层

反馈层通过基于用户反馈实时调整机器人行为, 完成了人机交互闭环, 但存在以下关键挑战.

(1) 动态校准不足: 无法及时根据细微用户反馈调整表情, 且缺乏长期交互记忆, 导致行为重复且空洞, 降低用户参与感。

(2) 评估限制: 缺乏客观、量化的指标或标准, 难以准确评估表情的清晰度、流畅度和拟人度, 过度依赖用户问卷, 阻碍严谨的基准测试和持续改进。

(3) 伦理风险与隐私问题: 情感真实性可能引发有害的情感依赖, 尤其是对孤独老年人等脆弱群体; 用户面部、声音及行为等敏感数据在传输、存储和处理过程中存在泄露和滥用风险。

这些技术与伦理难题使反馈层成为亟需多学科合作以确保安全可信社会机器人的关键领域。

6.1.5 系统集成层

系统集成层超越单一面部系统, 涵盖更高维度的协作, 如人形机器人与面部机器人的无缝集成。

(1) 拟人不协调: 高度逼真的面部特征与机械化人形机器人身体结合, 容易引发“恐怖谷”效应, 降低用户接受度。

(2) 运动与表情冲突: 头部组件的重量和惯性限制了运动稳定性和面部表情的流畅性, 迫使整体动作做出妥协。

(3) 多模态不同步: 缺乏协调框架导致面部表情与手势、姿态等身体语言不协调, 阻碍流畅且具语境感知的交流。

(4) 接口不一致: 面部与身体之间缺乏标准化硬件、软件接口及开放系统架构, 限制了实时传感数据交换、协调感知和自适应调整。

(5) 实时性较差: 表情渲染与生成延迟可能导致机器人“表情滞后”, 使其在快速互动中缺乏自然性。例如, 在对话过程中, 若微笑表情在语音结束后才出现, 用户会感受到明显的割裂感。

(6) 算力瓶颈: 在实际部署中, 多模态大模型在机器人平台上的算力消耗是重大挑战。尤其在边缘设备环境下, GPU/CPU算力不足导致表情生成与语音响应的延迟, 削弱了机器人的共情能力。

6.2 针对性解决方案

为克服感知层的局限, 解决方案集中于弥补感官缺失和实现多模态数据融合。例如, 使用先进的电子皮肤配备压力和温度传感器, 提供必要的触觉、温度

及化学感知能力, 以捕捉如触摸等细微的社交线索。多模态融合技术将视觉、听觉及新型传感输入整合为统一的环境理解。随后使用标准化的语义接口将原始传感数据转化为有意义的社交信号, 消除语义瓶颈, 实现更丰富的上下文感知。

为解决推理层的不足, 需加强情感与语境智能。基于LLM和情感计算的AI系统, 超越基础情绪分类, 通过分析多模态输入结合多元文化数据集, 精准理解复杂情绪状态, 确保文化适应性^[221]。受神经处理启发的预测模型将预判对话进程及用户状态, 及时生成微表情。集成认知模型则结合个体特征、情境因素和社会规范, 并融合社会学、人类学等跨学科洞见^[222]。

表达层的挑战通过仿生硬件和智能控制予以解决。笨重的电机将被体积小、静音的人工肌肉替代, 实现唇角上扬和细微眉毛动作等微表情, 同时减轻头部重量^[223,224]。多功能电子皮肤和生物混合材料提供动态变色、温度调节、触觉反馈及自愈功能, 提升真实感。由智能体驱动的系统实时协调驱动器和皮肤, 生成流畅、真实的微表情, 解决响应延迟和不同步问题。在此过程中, “恐怖谷”效应成为表达层的关键挑战之一。当机器人表情接近人类但仍存在细微差异时, 容易引发用户的不适。为缓解这一问题, 可从两方面入手: 一方面, 通过植入假发、仿真皮肤与毛发等“拟人化化妆”手段提升外观逼真度, 从而跨越“恐怖谷”; 另一方面, 借助人工智能模型调控眨眼频率、微表情持续时间等动态参数, 使其更接近人类自然节律, 并有选择性地减少如嘴部等易引发不适区域的运动幅度, 以降低违和感的产生。

在反馈层, 适应性学习结合多模态传感器与长期交互记忆实现动态校准, 避免行为重复, 提升互动参与度。客观评价替代主观调查, 利用计算机视觉与深度学习量化表情清晰度、流畅度和拟人性。隐私通过设备端处理、加密及访问控制得到保障, 伦理风险则通过用户教育、透明度和设计提示防止过度依赖^[225]。

系统集成挑战通过整体设计与标准化予以解决。通过半拟人化方法, 如将关键面部特征结合电子显示屏, 缓解拟人不协调现象, 避免恐怖谷效应^[226]。采用轻质材料和柔性执行器解决重量与惯性问题, 确保稳定运动与表情流畅。系统集成层还需在软硬件一体化过程中引入用户感知实验的反馈回路, 通过迭代优化来识别并抑制可能触发恐怖谷的因素, 从而在不同应

用场景中实现动态平衡. 此外, 低功耗嵌入式大模型的落地将成为关键, 它能够在有限算力下支持表情生成与情感计算的实时运行. 还应构建标准化的人机共情评价体系, 从跨文化接受度、伦理风险到用户体验开展多维度评估, 为不同研究提供可比较的基准. 最后, 未来的探索应更加跨学科, 融合认知科学、社会学与人机交互学, 从而推动面部机器人在社会性和实践性方面的进一步深化.

6.3 未来研究方向

面部机器人的未来发展可依据技术成熟度与社会融合水平, 大致划分为3个渐进阶段, 如图10所示. 该划分综合考量了技术发展的内在逻辑与权威机构的预测路线图. 初期阶段, 人工智能的进步将驱动表情表现力的革新; 中期阶段, 硬件仿生的突破将实现物理外观的

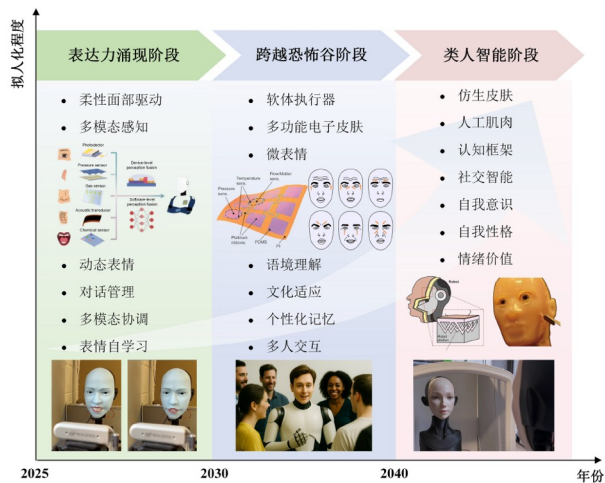


图 10 (网络版彩图)面部机器人未来研究的三个阶段. (1) 表达力涌现阶段. 面部机器人显著增强了它们的表达能力. 多模态感知, 引自文献[230]; 表情自学习, 引自文献[43]. (2) 跨越恐怖谷阶段. 面部机器人实现了高级物理真实感和细致入微的情感互动. 多功能电子皮肤, 引自文献[229]; 微表情, 从2023年维基共享资源获取; 多人交互, 引自Qincheng Sheng生成图片. (3) 类人智能阶段. 面部机器人实现完全的社会融合. 仿生皮肤, 引自文献[231]; 自我意识, 引自Yuhang Hu生成图片

Figure 10 (Color online) Three stages of future research roadmap in face robots. (1) Emergent expressivity stage. Face robots markedly enhance their expressive capabilities. Multimodal sensing [230]. Copyright©2023, The Author(s). Self-expression learning [43]. Copyright©2024, AAAS. (2) Biomimetic realism stage. Face robots achieve advanced physical realism and nuanced emotional engagement. Multifunctional e-skin [229]. Copyright©2017, Wiley. Micro-expression, retrieved 2023 from Wikimedia Commons. Multi-person interaction. Copyright©2025, Qincheng Sheng. (3) Near-human parity stage. Face robots attain full social integration. Biohybrid skin [231]. Copyright©2024, The Author(s). Self-awareness. Copyright©2025, Yuhang Hu.

超高真实感与社会接纳; 长期阶段, 研究的焦点将转向类人智能与社会身份的构建, 最终实现深度的情感协作与社会共生.

6.3.1 表达力涌现阶段(2025~2030年)

近期, 面部机器人将在人工智能的推动下显著提升其表达与感知能力. 此阶段聚焦于通过强化认知智能, 提升机器人作为有效社交交流者的能力, 更好地感知和理解人类互动. 机器人将采用柔性面部驱动器和多模态传感技术, 构建复杂的感知融合系统[227]. 这些提升将支持动态、实时生成面部表情, 突破预设情绪显示的局限. 先进的对话管理和多模态协调使机器人能够主动参与交流, 实现面部表情与视线、语音的同步, 保障流畅互动. 此外, 表情自学习将使机器人根据情境调整和个性化表情. 因此, 面部机器人将更自然、响应更灵敏, 适用于日常社交场景, 提供细腻且具文化敏感度的交互[43].

6.3.2 跨越恐怖谷阶段(2030~2040年)

中期阶段, 面部机器人将在仿生技术的推动下实现显著的物理真实感突破, 并逐步融入社会场景. 软体执行器将带来平滑、安静的面部运动, 模拟人体肌肉活动. 多功能电子皮肤及其嵌入式传感器将提供逼真的触觉和环境感知[228,229]. 改进的传感器与执行器整合将使机器人表现出细微的微表情, 增强情感真实感. 与此同时, 机器人将提升对语境的理解, 拥有个性化记忆和文化适应能力, 能够自然地管理多人会话的注意分配. 这些硬件与软件的协同创新将共同帮助机器人克服“恐怖谷”效应, 促进更深层次的社会互动与接纳. 最终, 这些进展旨在增强社交能力, 并促使社会重新思考与面部机器人建立情感纽带的意义和价值.

6.3.3 类人智能阶段(2040年以后)

在长期未来, 面部机器人将迈向接近人类的水平, 其社会角色可能发生根本性转变. 仿生皮肤、人工肌肉和多感知系统的进步将使机器人在物理外形和情感表达上高度逼真[230]. 然而, 研究重点将转向其全面融入人类社会结构的能力. 机器人将动态构建复杂的社会身份, 表达情感价值, 形成独特个性, 并可能展现出某种程度的自我意识, 积极参与情感交流和社会融合[231]. 这一阶段将引发深刻的社会、伦理思考: 我们

究竟渴望与那些在社交和情感上与人类难以区分的实体建立什么样的关系? 回答这一问题对于引导负责任

的创新, 以及将社交智能面部机器人以合乎道德的方式融入我们的生活至关重要。

参考文献

- 1 Ishiguro H, Nishio S. Building artificial humans to understand humans. *J Artif Organs*, 2007, 10: 133–142
- 2 McGinn C. Why do robots need a head? The role of social interfaces on service robots. *Int J Soc Robot*, 2020, 12: 281–295
- 3 Sheng Q, Zhou Z, Li J, et al. A comprehensive review of humanoid robots. *SmartBot*, 2025, 1: e12008
- 4 Song Y, Luximon Y. Trust in AI agent: A systematic review of facial anthropomorphic trustworthiness for social robot design. *Sensors*, 2020, 20: 5087
- 5 Youssef K, Said S, Alkork S, et al. A survey on recent advances in social robotics. *Robotics*, 2022, 11: 75
- 6 Mehrabian A, Ferris S R. Inference of attitudes from nonverbal communication in two channels. *J Consult Psychol*, 1967, 31: 248–252
- 7 Hashimoto T, Hitramatsu S, Tsuji T, et al. Development of the face robot SAYA for rich facial expressions. In: Proceedings of the 2006 SICE-ICASE International Joint Conference. Busan: IEEE, 2006. 5423–5428
- 8 Fortunati L, Sorrentino A, Fiorini L, et al. The rise of the roboid. *Int J Soc Robot*, 2021, 13: 1457–1471
- 9 Nieto Agraz C, Hinrichs P, Eichelberg M, et al. Is the robot spying on me? A study on perceived privacy in telepresence scenarios in a care setting with mobile and humanoid robots. *Int J Soc Robot*, 2025, 17: 363–377
- 10 Tadesse Y, Hong D, Priya S. Twelve degree of freedom baby humanoid head using shape memory alloy actuators. *J Mech Robot*, 2011, 3: 011008
- 11 Wang Y, Zhu J. Artificial muscles for jaw movements. *Extreme Mech Lett*, 2016, 6: 88–95
- 12 Khare S K, Blanes-Vidal V, Nadimi E S, et al. Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations. *Inf Fusion*, 2024, 102: 102019
- 13 Yan F, Iliyasu A M, Hirota K. Emotion space modelling for social robots. *Eng Appl Artif Intell*, 2021, 100: 104178
- 14 Wang Y, Song W, Tao W, et al. A systematic review on affective computing: Emotion models, databases, and recent advances. *Inf Fusion*, 2022, 83–84: 19–52
- 15 Domingo J D, Gómez-García-Bermejo J, Zalama E. Optimization and improvement of a robotics gaze control system using LSTM networks. *Multimed Tools Appl*, 2022, 81: 3351–3368
- 16 Li B, Li H, Liu H X. Driving animatronic robot facial expression from speech. In: Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, 2024. 7012–7019
- 17 Rojas-Quintero J A, Rodríguez-Liñán M C. A literature review of sensor heads for humanoid robots. *Robot Autonom Syst*, 2021, 143: 103834
- 18 Li Y, Zhu L X, Zhang Z Q, et al. Humanoid robot heads for human-robot interaction: A review. *Sci China Tech Sci*, 2024, 67: 357–379
- 19 Yoshida T, Masumori A, Ikegami T. From text to motion: Grounding GPT-4 in a humanoid robot “Alter3”. *Front Robot AI*, 2025, 12: 1581110
- 20 Kang H, Ben Moussa M, Thalmann N M. Nadine: A large language model-driven intelligent social robot with affective capabilities and human-like memory. *Comput Animat Virtual*, 2024, 35: e2290
- 21 Geller T. Overcoming the uncanny valley. *IEEE Comput Grap Appl*, 2008, 28: 11–17
- 22 Ekman P, Friesen W V. Facial action coding system (FACS): A technique for the measurement of facial actions. *Riv Psichiatri*, 1978, 47: 126–138
- 23 Marur T, Tuna Y, Demirci S. Facial anatomy. *Clin Dermatol*, 2014, 32: 14–23
- 24 Parent R. Computer Animation. 3rd ed. San Francisco: Morgan Kaufmann, 2012. 317–337
- 25 Faraj Z, Selamet M, Morales C, et al. Facially expressive humanoid robotic face. *HardwareX*, 2021, 9: e00117
- 26 Hashimoto T, Hiramatsu S, Kobayashi H. Development of face robot for emotional communication between human and robot. In: Proceedings of the 2006 International Conference on Mechatronics and Automation. Luoyang, 2006. 25–30
- 27 Nishio S, Ishiguro H, Hagita N. Geminoid: Teleoperated android of an existing person. In: de Pina Filho A C, ed. Humanoid Robots: New Developments. Vienna: IntechOpen, 2007. 343–352
- 28 Glas D F, Minato T, Ishi C T, et al. ERICA: The ERATO intelligent conversational android. In: Proceedings of the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). New York, 2016. 22–29

- 29 Diel A, Sato W, Hsu C T, et al. An android can show the facial expressions of complex emotions. *Sci Rep*, 2025, 15: 2433
- 30 Yang D S, Sato W, Liu Q Y, et al. Optimizing facial expressions of an android robot effectively: A bayesian optimization approach. In: Proceedings of the 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids). Ginowan, 2022. 542–549
- 31 Hara F, Akazawa H, Kobayashi H. Realistic facial expressions by SMA driven face robot. In: Proceedings of the 10th IEEE International Workshop on Robot and Human Interactive Communication. Bordeaux, 2001. 504–511
- 32 Pan F, Ruan J P, Tang W, et al. Research on crawling-jumping robot driven by dielectric elastomers (in Chinese). *Chin J Eng Des*, 2025, 32: 316–325 [潘锋, 阮佳平, 唐威, 等. 介电弹性体驱动的爬行-跳跃机器人研究. *工程设计学报*, 2025, 32: 316–325]
- 33 Kwak J W, Chi H J, Jung K M, et al. A face robot actuated with artificial muscle based on dielectric elastomer. *J Mech Sci Technol*, 2005, 19: 578–588
- 34 Li L, Godaba H, Ren H, et al. Bioinspired soft actuators for eyeball motions in humanoid robots. *IEEE ASME Trans Mechatron*, 2019, 24: 100–108
- 35 Hanson D, White V. Converging the capabilities of EAP artificial muscles and the requirements of bio-inspired robotics. In: Proceedings of the Smart Structures and Materials 2004: Electroactive Polymer Actuators and Devices (EAPAD). San Diego: SPIE, 2004. 29–40
- 36 Almubarak Y, Tadesse Y. Design and motion control of bioinspired humanoid robot head from servo motors toward artificial muscles. In: Proceedings of the Electroactive Polymer Actuators and Devices (EAPAD). Portland: SPIE, 2017. 101631U
- 37 Zhang T, Wang K S, Tang W, et al. Design and analysis of flexible bending actuator driven by electrohydrodynamic pumps (in Chinese). *Chin J Eng Des*, 2023, 30: 467–475 [张涛, 王开松, 唐威, 等. 电流体泵驱动的柔性弯曲执行器的设计及分析. *工程设计学报*, 2023, 30: 467–475]
- 38 Li Y, Zhu L, Zhang Z, et al. Modular multilayered PVC gel soft actuators for novel lightweight humanoid facial robot. *Sens Actuat A-Phys*, 2024, 378: 115825
- 39 Hashimoto T, Hiramatsu S, Kobayashi H. Dynamic display of facial expressions on the face robot made by using a life mask. In: Proceedings of the 2008 8th IEEE-RAS International Conference on Humanoid Robots. Daejeon, 2008. 521–526
- 40 Oh J H, Hanson D, Kim W S, et al. Design of android type humanoid robot Albert HUBO. In: Proceedings of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. Beijing, 2006. 1428–1433
- 41 Ahn H S, Lee D W, Choi D, et al. Designing of android head system by applying facial muscle mechanism of humans. In: Proceedings of the 12th IEEE-RAS International Conference on Humanoid Robots. Osaka, 2012. 799–804
- 42 Asheber W T, Lin C Y, Yen S H. Humanoid head face mechanism with expandable facial expressions. *Int J Adv Robot Syst*, 2016, 13: 29
- 43 Hu Y, Chen B, Lin J, et al. Human-robot facial coexpression. *Sci Robot*, 2024, 9: eadi4724
- 44 Berns K, Hirth J. Control of facial expressions of the humanoid robot head ROMAN. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems. Beijing, 2006. 3119–3124
- 45 Ishihara H, Yoshikawa Y, Asada M. Realistic child robot “Affetto” for understanding the caregiver-child attachment relationship that guides the child development. In: Proceedings of the 2011 IEEE International Conference on Development and Learning (ICDL). Frankfurt, 2011. 1–5
- 46 Almubarak Y, Maly N X, Tadesse Y. Fully embedded actuators in elastomeric skin for use in humanoid robots. In: Proceedings of the Electroactive Polymer Actuators and Devices (EAPAD). Denver: SPIE, 2018. 1059416
- 47 Habib A, Das S K, Bogdan I C, et al. Learning human-like facial expressions for Android Phillip K. Dick. In: Proceedings of the 2014 IEEE International Conference on Automation Science and Engineering (CASE). New Taipei, 2014. 1159–1165
- 48 Cominelli L, Hoegen G, De Rossi D. Abel: Integrating humanoid body, emotions, and time perception to investigate social interaction and human cognition. *Appl Sci*, 2021, 11: 1070
- 49 Liu X, Chen Y, Li J, et al. Real-time robotic mirrored behavior of facial expressions and head motions based on lightweight networks. *IEEE Internet Things J*, 2023, 10: 1401–1413
- 50 Lu Y X. On the evolution of design (in Chinese). *Chin J Eng Des*, 2025, 32: 427–428 [路甬祥. 论设计的进化. *工程设计学报*, 2025, 32: 427–428]
- 51 Ye Z, Pang G, Xu K, et al. Soft robot skin with conformal adaptability for on-body tactile perception of collaborative robots. *IEEE Robot Autom Lett*, 2022, 7: 5127–5134
- 52 Hanson D. Progress toward EAP actuators for biomimetic social robots. In: Proceedings of the Electroactive Polymer Actuators and Devices (EAPAD). San Diego: SPIE, 2013. 86870F
- 53 Park K, Yuk H, Yang M, et al. A biomimetic elastomeric robot skin using electrical impedance and acoustic tomography for tactile sensing. *Sci*

- Robot*, 2022, 7: eabm7187
- 54 Lin C Y, Huang C C, Cheng L C. An expressional simplified mechanism in anthropomorphic face robot design. *Robotica*, 2016, 34: 652–670
 - 55 Tomar A, Tadesse Y. Multi-layer robot skin with embedded sensors and muscles. In: *Proceedings of the Electroactive Polymer Actuators and Devices (EAPAD)*. Las Vegas: SPIE, 2016. 979809
 - 56 Zhong Y, Tang W, Gui H, et al. Human camouflage and expression via soft mask from reprogrammable chemical fluid skin. *Sci Adv*, 2025, 11: eadq6141
 - 57 Zong H, Lou B, Yuan H, et al. Integrating kinematic and dynamic factors with generative design for high-performance additive manufacturing structures. *Virtual Phys Prototyp*, 2025, 20: e2501383
 - 58 Cheng L C, Lin C Y, Huang C C. Visualization of facial expression deformation applied to the mechanism improvement of face robot. *Int J Soc Robot*, 2013, 5: 423–439
 - 59 Tzirakis P, Trigeorgis G, Nicolaou M A, et al. End-to-end multimodal emotion recognition using deep neural networks. *IEEE J Sel Top Signal Process*, 2017, 11: 1301–1309
 - 60 Kumar A, Kaur A, Kumar M. Face detection techniques: A review. *Artif Intell Rev*, 2019, 52: 927–948
 - 61 Gogić I, Ahlberg J, Pandžić I S. Regression-based methods for face alignment: A survey. *Signal Process*, 2021, 178: 107755
 - 62 Li S, Deng W. Deep facial expression recognition: A survey. *IEEE Trans Affect Comput*, 2020, 13: 1195–1215
 - 63 Mollahosseini A, Chan D, Mahoor M H. Going deeper in facial expression recognition using deep neural networks. In: *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. New York, 2016
 - 64 Baltrušaitis T, Robinson P, Morency L P. OpenFace: An open source facial behavior analysis toolkit. In: *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. New York, 2016
 - 65 Karnati M, Seal A, Bhattacharjee D, et al. Understanding deep learning techniques for recognition of human emotions using facial expressions: A comprehensive survey. *IEEE Trans Instrum Meas*, 2023, 72: 1–31
 - 66 Khabarлак K, Koriashkina L. Fast facial landmark detection and applications: A survey. arXiv: 2101.10808
 - 67 Xu Z, Li B, Yuan Y, et al. AnchorFace: An anchor-based facial landmark detector across large poses. *AAAI*, 2021, 35: 3092–3100
 - 68 Li H, Guo Z, Rhee S M, et al. Towards accurate facial landmark detection via cascaded transformers. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, 2022. 4176–4185
 - 69 Zhao Y, Liu Y, Shen C, et al. MobileFAN: Transferring deep hidden representation for face alignment. *Pattern Recogn*, 2020, 100: 107114
 - 70 Gao P, Lu K, Xue J, et al. A facial landmark detection method based on deep knowledge transfer. *IEEE Trans Neural Netw Learn Syst*, 2021, 34: 1342–1353
 - 71 Wu W, Qian C, Yang S, et al. Look at boundary: A boundary-aware face alignment algorithm. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, 2018. 2129–2138
 - 72 Jin H, Liao S, Shao L. Pixel-in-pixel net: Towards efficient facial landmark detection in the wild. *Int J Comput Vis*, 2021, 129: 3174–3194
 - 73 Ruan Z, Zou C, Wu L, et al. SADNet: Self-aligned dual face regression networks for robust 3D dense face alignment and reconstruction. *IEEE Trans Image Process*, 2021, 30: 5793–5806
 - 74 Wu C Y, Xu Q, Neumann U. Synergy between 3DMM and 3D landmarks for accurate 3D facial geometry. In: *Proceedings of the 2021 International Conference on 3D Vision (3DV)*. London, 2021. 453–463
 - 75 Shao Z, Zhou Y, Cai J, et al. Facial action unit detection via adaptive attention and relation. *IEEE Trans Image Process*, 2023, 32: 3354–3366
 - 76 Ma B, An R, Zhang W, et al. Facial action unit detection and intensity estimation from self-supervised representation. *IEEE Trans Affect Comput*, 2024, 15: 1669–1683
 - 77 Chang Y, Wang S. Knowledge-driven self-supervised representation learning for facial action unit recognition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, 2022. 20385–20394
 - 78 Yan J, Wang J, Li Q, et al. Weakly supervised regional and temporal learning for facial action unit recognition. *IEEE Trans Multimedia*, 2022, 25: 1760–1772
 - 79 Niu X, Han H, Shan S, et al. Multi-label co-regularization for semi-supervised facial action unit recognition. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. New York: ACM, 2019. 909–919
 - 80 Peng G, Wang S. Dual semi-supervised learning for facial action unit recognition. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Hawaii, 2019. 8827–8834

- 81 Ntinou I, Sanchez E, Bulat A, et al. A transfer learning approach to heatmap regression for action unit intensity estimation. *IEEE Trans Affect Comput*, 2021, 14: 436–450
- 82 Shao Z, Cai J, Cham T J, et al. Unconstrained facial action unit detection via latent feature domain. *IEEE Trans Affect Comput*, 2022, 13: 1111–1126
- 83 Shang Z, Liu B. Facial action unit detection based on multi-task learning strategy for unlabeled facial images in the wild. *Expert Syst Appl*, 2024, 253: 124285
- 84 Tallec G, Dapogny A, Bailly K. Multi-order networks for action unit detection. *IEEE Trans Affect Comput*, 2022, 14: 2876–2888
- 85 Walecki R, Rudovic O, Pavlovic V, et al. Deep structured learning for facial action unit intensity estimation. In: *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Hawaii, 2017. 5709–5718
- 86 Sandbach G, Zafeiriou S, Pantic M. Markov random field structures for facial action unit intensity estimation. In: *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops (ICCVW)*. Sydney, 2013. 738–745
- 87 Fan Y, Lam J, Li V. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Palo Alto, 2020. 12701–12708
- 88 Zhang Y, Jiang H, Wu B, et al. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data. In: *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, 2019. 733–742
- 89 Sanchez E, Tellamekala M K, Valstar M, et al. Affective processes: Stochastic modelling of temporal context for emotion and facial expression recognition. In: *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, 2021. 9070–9080
- 90 Kopalidis T, Solachidis V, Vretos N, et al. Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. *Information*, 2024, 15: 135
- 91 Ma F. Generative technology for human emotion recognition: A scope review. arXiv: [2407.03640](https://arxiv.org/abs/2407.03640)
- 92 Khemakhem F, Ltfi H. Neural style transfer generative adversarial network (NST-GAN) for facial expression recognition. *Int J Multimed Info Retr*, 2023, 12: 26
- 93 Mao J, Xu R, Yin X, et al. POSTER++: A simpler and stronger facial expression recognition network. *Pattern Recogn*, 2025, 157: 110951
- 94 Liu X, Jin L, Han X, et al. Identity-aware facial expression recognition in compressed video. In: *Proceedings of the 25th International Conference on Pattern Recognition (ICPR)*. Milan, 2021. 7508–7514
- 95 Pei E, Oveneke M C, Zhao Y, et al. Monocular 3D facial expression features for continuous affect recognition. *IEEE Trans Multimedia*, 2021, 23: 3540–3550
- 96 Qu X, Zou Z, Su X, et al. Attend to where and when: Cascaded attention network for facial expression recognition. *IEEE Trans Emerg Top Comput Intell*, 2022, 6: 580–592
- 97 Chen W, Zhang D, Li M, et al. STCAM: Spatial-temporal and channel attention module for dynamic facial expression recognition. *IEEE Trans Affect Comput*, 2023, 14: 800–810
- 98 Toisoul A, Kossaifi J, Bulat A, et al. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nat Mach Intell*, 2021, 3: 42–50
- 99 Yu C, Zhang D, Zou W, et al. Joint training on multiple datasets with inconsistent labeling criteria for facial expression recognition. *IEEE Trans Affect Comput*, 2024, 15: 1812–1825
- 100 Tellamekala M K, Sümer Ö, Schuller B W, et al. Are 3D face shapes expressive enough for recognising continuous emotions and action unit intensities? *IEEE Trans Affect Comput*, 2024, 15: 535–548
- 101 Liu S, Agaian S, Grigoryan A. PortraitEmotion3D: A novel dataset and 3D emotion estimation method for artistic portraiture analysis. *Appl Sci*, 2024, 14: 11235
- 102 Zhao G, Li X, Li Y, et al. Facial micro-expressions: An overview. *Proc IEEE*, 2023, 111: 1215–1235
- 103 Noroozi F, Corneanu C A, Kaminska D, et al. Survey on emotional body gesture recognition. *IEEE Trans Affect Comput*, 2021, 12: 505–523
- 104 Birdwhistell R L. *Kinesics and Context: Essays on Body Motion Communication*. New York: Ballantine, 1970
- 105 Kleinsmith A, Bianchi-Berthouze N. Affective body expression perception and recognition: A survey. *IEEE Trans Affect Comput*, 2013, 4: 15–33
- 106 Karg M, Samadani A A, Gorbet R, et al. Body movements for affective expression: A survey of automatic recognition and generation. *IEEE*

[Trans Affect Comput](#), 2013, 4: 341–359

- 107 Razzaq M A, Bang J, Kang S S, et al. UnSkEm: Unobtrusive skeletal-based emotion recognition for user experience. In: Proceedings of the 2020 International Conference on Information Networking (ICOIN). Barcelona, 2020. 92–96
- 108 Wei J, Hu G, Yang X, et al. Learning facial expression and body gesture visual information for video emotion recognition. [Expert Syst Appl](#), 2024, 237: 121419
- 109 Cazzato D, Leo M, Distanto C, et al. When I look into your eyes: A survey on computer vision contributions for human gaze estimation and tracking. [Sensors](#), 2020, 20: 3739
- 110 Eggert T. Eye movement recordings: Methods. *Neuro-Ophthalmol*, 2007, 40: 15–34
- 111 Tan K H, Kriegman D J, Ahuja N. Appearance-based eye gaze estimation. In: Proceedings of the 6th IEEE Workshop on Applications of Computer Vision (WACV). Orlando, 2002. 191–195
- 112 Pathirana P, Senarath S, Meedeniya D, et al. Eye gaze estimation: A survey on deep learning-based approaches. [Expert Syst Appl](#), 2022, 199: 116894
- 113 Zhang X, Sugano Y, Fritz M, et al. Appearance-based gaze estimation in the wild. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, 2015. 4511–4520
- 114 Krafska K, Khosla A, Kellnhofer P, et al. Eye tracking for everyone. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, 2016. 2176–2184
- 115 Cheng Y, Wang H, Bao Y, et al. Appearance-based gaze estimation with deep learning: A review and benchmark. [IEEE Trans Pattern Anal Mach Intell](#), 2024, 46: 7509–7528
- 116 Alluhaidan A S, Saidani O, Jahangir R, et al. Speech emotion recognition through hybrid features and convolutional neural network. [Appl Sci](#), 2023, 13: 4750
- 117 Huang J, Tao J, Liu B, et al. Learning utterance-level representations with label smoothing for speech emotion recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech). Shanghai, 2020. 4079–4083
- 118 Juslin P N, Laukka P. Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. [Emotion](#), 2001, 1: 381–412
- 119 Mori H, Kasuya H. Voice source and vocal tract variations as cues to emotional states perceived from expressive conversational speech. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech). Antwerp, 2007. 625–628
- 120 Grimm M, Kroschel K, Narayanan S. Support vector regression for automatic recognition of spontaneous emotions in speech. In: Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Honolulu, 2007. IV-1085–IV-1088
- 121 Mowlaee P, Saeidi R, Stylianou Y. Phase importance in speech processing applications. In: Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech). Singapore, 2014. 1623–1627
- 122 El Ayadi M, Kamel M S, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. [Pattern Recognit](#), 2011, 44: 572–587
- 123 Araño K A, Gloor P, Orsenigo C, et al. When old meets new: Emotion recognition from speech signals. [Cogn Comput](#), 2021, 13: 771–783
- 124 Lian Z, Liu B, Tao J. DECN: Dialogical emotion correction network for conversational emotion recognition. [Neurocomputing](#), 2021, 454: 483–495
- 125 Tran M, Soleymani M. A pre-trained audio-visual transformer for emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore, 2022. 4698–4702
- 126 Liu Z, Kang X, Ren F. Dual-TBNet: Improving the robustness of speech features via dual-transformer-BiLSTM for speech emotion recognition. [IEEE ACM Trans Audio Speech Lang Process](#), 2023, 31: 2193–2203
- 127 Ghosal D, Majumder N, Poria S, et al. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, 2019. 154–164
- 128 Dimitrievska V, Ackovska N. Behavior models of emotion-featured robots: A survey. [J Intell Robot Syst](#), 2020, 100: 1031–1053
- 129 Shu L, Xie J, Yang M, et al. A review of emotion recognition using physiological signals. [Sensors](#), 2018, 18: 2074
- 130 Aly L, Godinho L, Bota P, et al. Acting Emotions: A comprehensive dataset of elicited emotions. [Sci Data](#), 2024, 11: 147
- 131 Ekman P. An argument for basic emotions. [Cognit Emotion](#), 1992, 6: 169–200

- 132 Plutchik R, Kellerman H. Theories of Emotion. New York: Academic Press, 1980
- 133 Johnston C H. The present state of the psychology of feeling.. *Psychol Bull*, 1905, 2: 161–171
- 134 Russell J A. A circumplex model of affect. *J Pers Soc Psychol*, 1980, 39: 1161–1178
- 135 Mehrabian A, Russell JA. An Approach to Environmental Psychology. Cambridge: MIT Press, 1974
- 136 Breazeal C, Scassellati B. How to build robots that make friends and influence people. In: Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Kyongju, 1999. 858–863
- 137 Yamazaki Y, Vu H A, Le P Q, et al. Fuzzy inference-based mentality expression for eye robot in affinity pleasure-arousal space. In: Fodor J, Kacprzyk J, eds. Aspects of Soft Computing. Intelligent Robotics and Control. Studies in Computational Intelligence, Vol. 241. Berlin: Springer-Heidelberg, 2009. 89–100
- 138 Izard C E. Human Emotions. New York: Springer Science & Business Media, 1977
- 139 Cambria E, Livingstone A, Hussain A. The hourglass of emotions. In: Esposito A, Esposito A M, Vinciarelli A, et al., eds. Cognitive Behavioural Systems. Lecture Notes in Computer Science, Vol. 7403. Berlin: Springer-Heidelberg, 2012. 144–157
- 140 Ortony A, Clore G L, Collins A. The Cognitive Structure of Emotions. Cambridge: Cambridge University Press, 1990
- 141 Wang Z, Ho S B, Cambria E. A review of emotion sensing: Categorization models and algorithms. *Multimed Tools Appl*, 2020, 79: 35553–35582
- 142 Guo Y, Gao H. Emotion recognition system in images based on fuzzy neural network and HMM. In: Proceedings of the 2006 5th IEEE International Conference on Cognitive Informatics (ICCI). Beijing, 2006. 73–78
- 143 van Kesteren A J, op den Akker H J A, Poel M. Simulation of emotions of agents in virtual environments using neural networks. In: Jokinen K, Heylen D K J, Nijholt A, eds. Learning to Behave: Internalising Knowledge, Vol. 18. Enschede: University of Twente, 2000. 137–147
- 144 Ilyasu A M. Roadmap to talking quantum movies: A contingent inquiry. *IEEE Access*, 2019, 7: 23864–23913
- 145 Masuyama N, Loo C K, Seera M. Personality affected robotic emotional model with associative memory for human-robot interaction. *Neurocomputing*, 2018, 272: 213–225
- 146 Breazeal C. Emotion and sociable humanoid robots. *Int J Hum-Comput Stud*, 2003, 59: 119–155
- 147 Itoh K, Miwa H, Takanobu H, et al. Application of neural network to humanoid robots—Development of co-associative memory model. *Neural Netws*, 2005, 18: 666–673
- 148 Han M J, Lin C H, Song K T. Robotic emotional expression generation based on mood transition and personality model. *IEEE Trans Cybern*, 2013, 43: 1290–1303
- 149 Ficocelli M, Terao J, Nejat G. Promoting interactions between humans and robots using robotic emotional behavior. *IEEE Trans Cybern*, 2016, 46: 2911–2923
- 150 Zhang Z, Yang W, Wang W. Building an affective model for social robots with customizable personality. In: Cavallo F, Cabibihan J J, Fiorini L, et al., eds. Social Robotics. Lecture Notes in Computer Science, Vol. 13817. Cham: Springer, 2022. 536–546
- 151 Hong A, Lunscher N, Hu T, et al. A multimodal emotional human-robot interaction architecture for social robots engaged in bidirectional communication. *IEEE Trans Cybern*, 2021, 51: 5954–5968
- 152 Liu X, Xie L, Wang Z. Empathizing with emotional robot based on cognition reappraisal. *China Commun*, 2017, 14: 100–113
- 153 Ishihara H, Hayashi R, Lavieille F, et al. Automatic generation of dynamic arousal expression based on decaying wave synthesis for robot faces. *J Robot Mechatron*, 2024, 36: 1481–1494
- 154 Huang Z, Ren F, Hu M, et al. Facial expression imitation method for humanoid robot based on smooth-constraint reversed mechanical model (SRMM). *IEEE Trans Hum-Mach Syst*, 2020, 50: 538–549
- 155 Mazzei D, Lazzeri N, Hanson D, et al. HEFES: An hybrid engine for facial expressions synthesis to control human-like androids and avatars. In: Proceedings of the 2012 4th IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechanics (BioRob). Rome, 2012. 195–200
- 156 Ghorbandaei Pour A, Taheri A, Alemi M, et al. Human-robot facial expression reciprocal interaction platform: Case studies on children with autism. *Int J Soc Robot*, 2018, 10: 179–198
- 157 Wykowska A. Social robots to test flexibility of human social cognition. *Int J Soc Robot*, 2020, 12: 1203–1211
- 158 Yun S S. A gaze control of socially interactive robots in multiple-person interaction. *Robotica*, 2017, 35: 2122–2138
- 159 Raković M, Ferreira Duarte N, Marques J, et al. The gaze dialogue model: Nonverbal communication in HHI and HRI. *IEEE Trans Cybern*,

2024, 54: 2026–2039

- 160 Palinko O, Rea F, Sandini G, et al. Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In: Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, 2016. 5048–5054
- 161 Hashimoto M, Kondo H, Tamatsu Y. Effect of emotional expression to gaze guidance using a face robot. In: Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). Munich, 2008. 95–100
- 162 Iimori M, Furuya Y, Takashio K. Face robot performing interaction with emphasis on eye blink entrainment. In: Proceedings of the 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). Busan, 2023. 634–639
- 163 Pan M K X J, Choi S, Kennedy J, et al. Realistic and interactive robot gaze. In: Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, 2020. 11072–11078
- 164 Roncone A, Pattacini U, Metta G, et al. A cartesian 6-DoF gaze controller for humanoid robots. In: Proceedings of Robotics: Science and Systems. Ann Arbor, 2016
- 165 McGurk H, Macdonald J. Hearing lips and seeing voices. *Nature*, 1976, 264: 746–748
- 166 Lin C Y, Cheng L C, Shen L C. Oral mechanism design on face robot for lip-synchronized speech. In: Proceedings of the 2013 IEEE International Conference on Robotics and Automation (ICRA). Karlsruhe, 2013. 4316–4321
- 167 Strathearn C, Ma E M. A novel speech to mouth articulation system for realistic humanoid robots. *J Intell Robot Syst*, 2021, 101: 54
- 168 Madumal K A D A, Jayaweera J M T C, Alahakoon P M K, et al. Design of voice synchronized robotic lips. In: Proceedings of the 2021 3rd International Conference on Electrical, Control and Instrumentation Engineering (ICECIE). Kuala Lumpur, 2021. 1–5
- 169 Heisler M, Kopp S, Becker-Asano C. Making an android robot head talk. In: Proceedings of the 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). Busan, 2023. 1837–1842
- 170 Yu C, Zhang H, Shanguan Z, et al. Speech-driven robot face action generation with deep generative model for social robots. In: Cavallo F, Cabibihan J J, Fiorini L, et al., eds. Social Robotics. Lecture Notes in Computer Science, Vol. 13817. Cham: Springer, 2022. 72–84
- 171 Al-Saadawi H F T, Das B, Das R. A systematic review of trimodal affective computing approaches: Text, audio, and visual integration in emotion recognition and sentiment analysis. *Expert Syst Appl*, 2024, 255: 124852
- 172 Tsiourti C, Weiss A, Wac K, et al. Multimodal integration of emotional signals from voice, body, and context: Effects of (in)congruence on emotion recognition and attitudes towards robots. *Int J Soc Robot*, 2019, 11: 555–573
- 173 Geetha A V, Mala T, Priyanka D, et al. Multimodal emotion recognition with deep learning: Advancements, challenges, and future directions. *Inf Fusion*, 2024, 105: 102218
- 174 Poria S, Cambria E, Bajpai R, et al. A review of affective computing: From unimodal analysis to multimodal fusion. *Inf Fusion*, 2017, 37: 98–125
- 175 Zhang S, Yang Y, Chen C, et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Syst Appl*, 2024, 237: 121692
- 176 Udaheureka G, Djouani K, Kurien A M. Multimodal emotion recognition using visual, vocal and physiological signals: A review. *Appl Sci*, 2024, 14: 8071
- 177 Zhao M, Gong L, Din A S. A review of the emotion recognition model of robots. *Appl Intell*, 2025, 55: 364
- 178 Jeong H, Lee H, Kim C, et al. A survey of robot intelligence with large language models. *Appl Sci*, 2024, 14: 8868
- 179 Cao L. AI robots and humanoid AI: Review, perspectives and directions. arXiv: [2405.15775](https://arxiv.org/abs/2405.15775)
- 180 Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, 2022. 24824–24837
- 181 Li Z, Chen G, Shao R, et al. Enhancing emotional generation capability of large language models via emotional chain-of-thought. arXiv: [2401.06836](https://arxiv.org/abs/2401.06836)
- 182 Wang X, Li X, Yin Z, et al. Emotional intelligence of large language models. *J Pac Rim Psychol*, 2023, doi: 10.1177/18344909231213958
- 183 Mishra C, Verdonshot R, Hagoort P, et al. Real-time emotion generation in human-robot dialogue using large language models. *Front Robot AI*, 2023, 10: 1271610
- 184 Tak A N, Gratch J. Is GPT a computational model of emotion? In: Proceedings of the 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII). Cambridge, 2023. 1–8
- 185 Zhou K, Yang J, Loy C C, et al. Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on

- Computer Vision and Pattern Recognition (CVPR). New Orleans, 2022. 16795–16804
- 186 Lian Z, Sun L, Sun H, et al. GPT-4V with emotion: A zero-shot benchmark for generalized emotion recognition. *Inf Fusion*, 2024, 108: 102367
- 187 Arjmand M, Nouraei F, Steenstra I, et al. Empathic grounding: Explorations using multimodal interaction and large language models with conversational agents. In: Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents. New York, 2024. 1–10
- 188 Zhao J, Wei X, Bo L, et al. R1-Omni: Explainable omni-multimodal emotion recognition with reinforcement learning. arXiv: [2503.05379](#)
- 189 Nadeem M, Sohail S S, Javed L, et al. Vision-enabled large language and deep learning models for image-based emotion recognition. *Cogn Comput*, 2024, 16: 2566–2579
- 190 Lin M Y, Lee O W, Lu C Y. Embodied ai with large language models: A survey and new HRI framework. In: Proceedings of the 2024 International Conference on Advanced Robotics and Mechatronics (ICARM). Tokyo, 2024. 978–983
- 191 Schlegel K, Sommer N R, Mortillaro M. Large language models are proficient in solving and creating emotional intelligence tests. *Commun Psychol*, 2025, 3: 80
- 192 Ojha S, Vitale J, Williams M A. Computational emotion models: A thematic review. *Int J Soc Robot*, 2021, 13: 1253–1279
- 193 Zhang C, Chen J, Li J, et al. Large language models for human-robot interaction: A review. *Biomim Intell Robot*, 2023, 3: 100131
- 194 Lu H, Niu X, Wang J, et al. GPT as psychologist? Preliminary evaluations for GPT-4V on visual affective computing. In: Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Seattle, 2024. 322–331
- 195 Ma F, Yuan Y, Xie Y, et al. Generative technology for human emotion recognition: A scoping review. *Inf Fusion*, 2025, 115: 102753
- 196 Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans Inf Syst*, 2025, 43: 1–55
- 197 Xi Z, Chen W, Guo X, et al. The rise and potential of large language model based agents: A survey. *Sci China Inf Sci*, 2025, 68: 121101
- 198 Liu B, Jiang Y, Zhang X, et al. LLM+P: Empowering large language models with optimal planning proficiency. arXiv: [2304.11477](#)
- 199 Yao S, Zhao J, Yu D, et al. React: Synergizing reasoning and acting in language models. arXiv: [2210.03629](#)
- 200 Shinn N, Cassano F, Labash B, et al. Reflexion: Language agents with verbal reinforcement learning. arXiv: [2303.11366](#)
- 201 Li D, Rawat A S, Zaheer M, et al. Large language models with controllable working memory. arXiv: [2211.05110](#)
- 202 Bärmann L, Kartmann R, Peller-Konrad F, et al. Incremental learning of humanoid robot behavior from natural interaction and large language models. *Front Robot AI*, 2024, 11: 1455375
- 203 Zhang Z, Bo X, Ma C, et al. A survey on the memory mechanism of large language model based agents. arXiv: [2404.13501](#)
- 204 Qin Y, Liang S, Ye Y, et al. Toolllm: Facilitating large language models to master 16000+ real-world apis. arXiv: [2307.16789](#)
- 205 Cai T, Wang X, Ma T, et al. Large language models as tool makers. arXiv: [2305.17126](#)
- 206 Busoniu L, Babuska R, De Schutter B. A comprehensive survey of multiagent reinforcement learning. *IEEE Trans Syst Man Cybern C*, 2008, 38: 156–172
- 207 Hong S, Zheng X, Chen J, et al. Metagpt: Meta programming for multi-agent collaborative framework. arXiv: [2308.00352](#)
- 208 Hou X, Zhao Y, Wang S, et al. Model context protocol (MCP): Landscape, security threats, and future research directions. arXiv: [2503.23278](#)
- 209 Neumann M M. Social robots and young children’s early language and literacy learning. *Early Childhood Educ J*, 2020, 48: 157–170
- 210 Pillai R, Sivathanu B, Metri B, et al. Students’ adoption of AI-based teacher-bots (T-bots) for learning in higher education. *Inf Technol People*, 2024, 37: 328–355
- 211 Aktar Mispa T, Sojib N. Educational robot Kiddo learns to draw to enhance interactive handwriting scenario for primary school children. In: Proceedings of the 2020 3rd International Conference on Intelligent Robotic and Control Engineering (IRCE). Oxford, 2020. 87–91
- 212 Sartorato F, Przybylowski L, Sarko D K. Improving therapeutic outcomes in autism spectrum disorders: Enhancing social communication and sensory processing through the use of interactive robots. *J Psychiatric Res*, 2017, 90: 1–11
- 213 Chen S, Jones C, Moyle W. Social robots for depression in older adults: A systematic review. *J Nurs Scholarship*, 2018, 50: 612–622
- 214 Pu L, Moyle W, Jones C, et al. The effectiveness of social robots for older adults: A systematic review and meta-analysis of randomized controlled studies. *Gerontologist*, 2019, 59: e37–e51
- 215 Wong K L Y, Hung L, Wong J, et al. Adoption of artificial intelligence-enabled robots in long-term care homes by health care providers: Scoping review. *JMIR Aging*, 2024, 7: e55257
- 216 Abdi J, Al-Hindawi A, Ng T, et al. Scoping review on the use of socially assistive robot technology in elderly care. *BMJ Open*, 2018, 8: e018815

- 217 Belk R. Artificial emotions and love and sex doll service workers. *J Serv Res*, 2022, 25: 521–536
- 218 Nieuwenhuisen M, Behnke S. Human-like interaction skills for the mobile communication robot robotinho. *Int J Soc Robot*, 2013, 5: 549–561
- 219 Ahn H S, Lee D W, Choi D, et al. Development of an android for singing with facial expression. In: Proceedings of the 37th Annual Conference of the IEEE Industrial Electronics Society. Melbourne, 2011. 104–109
- 220 Ahmed E, Buruk O, Hamari J. Human-robot companionship: Current trends and future agenda. *Int J Soc Robot*, 2024, 16: 1809–1860
- 221 Yew G C K. Trust in and ethical design of carebots: The case for ethics of care. *Int J Soc Robot*, 2021, 13: 629–645
- 222 Weiss A, Spiel K. Robots beyond science fiction: Mutual learning in human-robot interaction on the way to participatory approaches. *AI Soc*, 2022, 37: 501–515
- 223 Obrenovic B, Gu X, Wang G, et al. Generative AI and human-robot interaction: Implications and future agenda for business, society and ethics. *AI Soc*, 2025, 40: 677–690
- 224 Ishihara H, Iwanaga S, Asada M. Comparison between the facial flow lines of androids and humans. *Front Robot AI*, 2021, 8: 540193
- 225 Boada J P, Maestre B R, Genis C T. The ethical issues of social assistive robotics: A critical literature review. *Tech Soc*, 2021, 67: 101726
- 226 Blut M, Wang C, Wunderlich N V, et al. Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *J Acad Mark Sci*, 2021, 49: 632–658
- 227 Zhang Y, Qian Y, Wu D, et al. Emotion-aware multimedia systems security. *IEEE Trans Multimedia*, 2019, 21: 617–624
- 228 Basgol H, Ayhan I, Ugur E. Time perception: A review on psychological, computational, and robotic models. *IEEE Trans Cogn Dev Syst*, 2022, 14: 301–315
- 229 Zhao S, Zhu R. Electronic skin with multifunction sensors based on thermosensation. *Adv Mater*, 2017, 29: 1606151
- 230 Tu J, Wang M, Li W, et al. Electronic skins with multimodal sensing and perception. *Soft Sci*, 2023, 3: 25
- 231 Kawai M, Nie M, Oda H, et al. Perforation-type anchors inspired by skin ligament for robotic face covered with living skin. *Cell Rep Phys Sci*, 2024, 5: 102066

A review on AI-driven face robot for human-robot expression interaction

SHENG QinCheng, TANG Wei, QIN Hao, KONG YuJie, DAI HaoKai, ZHONG YiDing, WANG YongHao, ZOU Jun & YANG HuaYong

State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou 310058, China

As artificial intelligence (AI) extends humanoid robots into social domains like education, healthcare, and home, the need for emotional interaction is increasing. Facial expressions, conveying 55% of emotional information, are key to emotional bonding, making realistic-faced humanoid robots—face robots—increasingly essential. This article reviews AI-driven expression interaction technologies in face robots. It first examines the hardware architecture of face robots, then analyzes a “perception-reasoning-generation” framework by comparing traditional and advanced approaches. Traditional methods rely on visual and speech-based emotion recognition, along with discrete or dimensional emotion models, to drive expression generation (e.g., facial movements, eye contact, and lip synchronization) through affective computing. In contrast, advanced approaches leverage multimodal fusion, large language model or multimodal large language model based emotion reasoning, and agent-based planning, memory, and tool use to enhance adaptability, realism, and emotional intelligence. The article also discusses potential application areas, current challenges, and future research directions for face robots. By integrating progress across hardware, algorithms, applications, and open issues, this review lays a comprehensive foundation for the development of empathetic, socially adaptive face robots suitable for complex human environments.

face robot, artificial intelligence, human-robot interaction, large language model, expression generation

doi: [10.1360/SST-2025-0217](https://doi.org/10.1360/SST-2025-0217)