智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

面向听视觉信息的多模态人格识别研究进展

赵小明，唐志伟，张石清

引用本文:
赵小明, 唐志伟, 张石清. 面向听视觉信息的多模态人格识别研究进展[J]. 智能系统学报, 2021, 16(2): 189–201.
ZHAO Xiaoming, TANG Zhiwei, ZHANG Shiqing. Research advance of multimodal personality recognition based on audio and visual cues[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(2): 189–201.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202101034

DOI: 10.11992/tis.202101034

面向听视觉信息的多模态人格识别研究进展

赵小明[1,2], 唐志伟[1], 张石清[2]

(1. 浙江理工大学 机械与自动控制学院, 浙江 杭州 310018; 2. 台州学院 智能信息处理研究所, 浙江 台州 318000)

摘 要: 人格识别分析是人格计算研究中一个重要的研究内容, 在人类行为分析、人工智能、人机交互、个性化推荐等方面具有重要的应用价值, 是近年来心理学、认知学、计算机科学等领域中的一个多学科交叉的热点研究课题。本文介绍了与人格识别相关的各种人格类型表示理论和人格识别数据库, 阐述了面向听视觉信息的各种听视觉人格特征提取技术, 如手工特征和深度特征, 并在此基础上对面向听视觉信息人格识别的多模态融合方法做了详细的分类和归纳, 最后概括了面向听视觉信息的多模态人格识别发展趋势, 并进行了展望。

关键词: 人格识别; 人格计算; 人格类型; 听视觉信息; 特征提取; 手工特征; 深度特征; 多模态融合

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2021)02−0189−13

中文引用格式: 赵小明, 唐志伟, 张石清. 面向听视觉信息的多模态人格识别研究进展 [J]. 智能系统学报, 2021, 16(2): 189–201.

英文引用格式: ZHAO Xiaoming, TANG Zhiwei, ZHANG Shiqing. Research advance of multimodal personality recognition based on audio and visual cues[J]. CAAI transactions on intelligent systems, 2021, 16(2): 189–201.

Research advance of multimodal personality recognition based on audio and visual cues

ZHAO Xiaoming[1,2], TANG Zhiwei[1], ZHANG Shiqing[2]

(1. School of Faculty of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2. Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China)

Abstract: Personality recognition analysis is an important research topic in personality computing, which has important applications in human behavior analysis, artificial intelligence, human-computer interaction, and personalized recommendation. In recent years, personality recognition analysis has become an active research topic in psychology, cognition, and computer science. This study introduces different types of personality representation theories and databases related to personality recognition and presents various audio-visual cue feature extraction technologies for personality recognition, such as handcrafted and depth features. Then, multimodal fusion methods integrating audio and visual cues for personality recognition are classified and summarized in detail. Finally, the development trend of multimodal personality recognition integrating audio and visual cues is explored and summarized.

Keywords: Personality recognition; personality computing; types of personality; audio-visual cues; feature extraction; hand-crafted features; depth features; multimodal fusion

心理学家认为, 一个人的性格从出生时就已经注定的。因此, 每个人都有其特定的性格。在心理学领域, 为了研究与人的性格相关的个性, 现已提出了各种各样的理论和方法对其进行解释或测量。Vinciarelli 等[1]将人格定义为: “人格是一种心理结构, 旨在解释人类行为的多样性, 具有少数、稳定和可测量的个体特征”。Costa 等[2]提出一种人格特质理论, 即认为特质 (traits) 是决定人类个体行为的基本特性的重要因素之一, 是人格组成的关键元素, 也是用于测评人格的基本度量单位。它用来衡量人的性格特征, 即随着时间的推移相对稳定的人类行为、观念和情感的习惯模式。

近年来, 随着认知科学、计算机科学等理论的发展, 研究者开始尝试根据一个人给人第一印象 (first impression) 的行为数据 (如听觉、视觉等

收稿日期: 2021−01−28.

基金项目: 国家自然科学基金项目 (61976149); 浙江省自然科学基金项目 (LZ20F020002).

通信作者: 赵小明. E-mail: tzxyzxm@163.com.

音视频信息), 采用机器学习方法来实现人格的建模与计算, 称为人格计算 (personality computing)[3]。其中, 自动人格识别是人格计算研究中一个重要的研究内容。它是指利用计算机根据一个人第一印象的行为数据来自动识别和分析个体心理特征的过程。可见, 人格识别是利用计算机科学理论来实现认知科学中的性格预测问题的建模。如果采用的行为数据为单一模态的听觉或视觉信息, 则称为单模态人格识别。如果是融合多个模态的行为数据, 如听觉、视觉等音视频信息, 则称为多模态人格识别。该研究在人类行为分析、人工智能、人机交互、个性化推荐等方面具有重要的应用价值。例如, 企业招聘新员工时, 可根据面试人员的第一印象, 采用自动人格识别技术来识别面试人员的性格特征, 从而筛选出更合适的员工。

当前, 有关自动人格识别的研究已成为心理学、认知科学、计算机科学等领域中的一个多学科交叉的热点研究课题。2016 年, 欧洲计算机视觉大会 (ECCV) 举办了第一场全球范围的基于短视频的自动人格识别专题竞赛[4], 使得基于社交媒体内容的自动人格特质识别成为一个富有挑战性的热点研究课题。2017 年, 国际计算机视觉与模式识别大会 (CVPR) 举办了第二场采用人格特质进行求职者筛选的相关专题竞赛[5]。从那之后, 有关自动人格识别的研究开始备受计算机视觉、模式识别、人工智能等领域研究者的关注。

本文将对自动人格识别领域涉及的核心研究内容, 如人格类型表示理论、人格识别数据库的建设、人格特征提取以及面向听视觉信息融合的多模态人格识别方法等方面, 详述国内外研究进展状况, 并指出未来的发展方向。

1　人格类型表示理论

心理学中的人格被认为是个体与行为、情感、思维方式有关的特征综合, 人格特质可以被识别为对用户行为的预测。在心理学领域, 通过多个分类维度和测量问卷[6] 对人的人格状况进行建模, 形成了几种不同的人格模型, 主要有“大五”模型[7]、卡特尔 16 种人格因素 (16PF)[8]、Myers-Briggs 类型指标 (MBTI)[9]、明尼苏达多项人格调查表 (MMPI)[10] 和艾森克人格问卷 (EPQ)[11]。这几种人格模型具体表述如下。

1.1　大五模型

由美国著名心理学家 McCrae 等[8] 提出的大五类 (Big-Five) 因素模型被广泛用来描述人的人格。该模型通过以下 5 个维度描述人类的人格, 具体特征如下:

1) 开放性 (openness, O): 艺术性、好奇心、想象力、洞察力、独创性、广泛兴趣等;

2) 尽责性 (conscientiousness, C): 高效、有组织、有计划、可靠、负责任、彻底等;

3) 外向性 (extroversion, E): 积极、自信、精力充沛、外向、健谈等;

4) 宜人性 (agreeableness, A): 欣赏、善良、慷慨、宽容、富有同情心、信任他人等;

5) 神经质 (neuroticism, N): 焦虑、自怜、紧张、敏感、不稳定、令人担忧等。

“大五”模型是一种心理学中最具影响力的模型之一。五项的人格特质因素评分值进行加权求和得出五项人格因素的最终评分, 预测出测定人的五项人格因素指数值, 并将其作为测定人的最终的人格特质预测结果。目前, 大五因素模型, 被人们广泛用于人格计算, 用于发现人类 5 个人格维度方面的证据。

1.2　卡特尔 16 人格因素模型

卡特尔 16 种人格因素模型[8] 是美国伊利诺州立大学人格及能力测验研究所卡特尔教授编制的用于人格检测的一种问卷, 简称 16PF。他用因素分析法对人格特质进行了分析, 提出了一种基于人格特质的理论模型。该模型分成 4 层:个别特质和共同特质, 表面特质和根源特质, 体质特质和环境特质, 动力特质、能力特质和气质特质。

16 种人格因素的含义如下:

1) 因素 A 乐群性: 低分特征—— 内向、缄默、孤独; 高分特征——外向、热情、乐群。

2) 因素 B 聪慧性: 低分特征—— 思想迟钝, 学识浅薄; 高分特征——聪明伶俐, 富有才识。

3) 因素 C 稳定性: 低分特征——情绪激动不稳定; 高分特征——情绪稳定且成熟。

4) 因素 E 恃强性: 低分特征——谦逊、顺从; 高分特征——好强、固执。

5) 因素 F 兴奋性: 低分特征——严肃、冷静; 高分特征——轻松、兴奋。

6) 因素 G 有恒性: 低分特征—— 原则性差、做事敷衍; 高分特征——有恒心、做事尽责。

7) 因素 H 敢为性: 低分特征—— 做事畏缩、缺乏自信心; 高分特征——冒险敢为、少有顾忌。

8) 因素 I 敏感性: 低分特征—— 理智、粗心、着重现实; 高分特征——敏感、细心、易感情用事。

9) 因素 L 怀疑性: 低分特征——依赖、随和、易与人相处; 高分特征——怀疑、刚愎、固执己见。

10) 因素 M 幻想性: 低分特征——现实、合乎

成规; 高分特征——幻想、狂放不羁。

11) 因素 N 世故性: 低分特征—— 坦诚、直率、天真; 高分特征——精明、圆滑、世故。

12) 因素 O 忧虑性: 低分特征—— 安详、沉着、有自信心; 高分特征——忧虑、抑郁、缺乏自信。

13) 因素 Q1 实验性: 低分特征—— 保守、尊重传统观念; 高分特征——激进、不拘于现实。

14) 因素 Q2 独立性: 低分特征—— 依赖、随群附众; 高分特征——当机立断、自主性强。

15) 因素 Q3 自律性: 低分特征—— 不守纪律、随心所欲; 高分特征——自律谨严、知己知彼。

16) 因素 Q4 紧张性: 低分特征—— 镇定自若、心平气和; 高分特征——手足无措、心神不宁。

上述 16 种人格因素是各自独立的, 相互之间的相关度极小, 每一种因素的测量都能使被试某一方面的人格特质有清晰而独特的认识, 更能对被试人格的 16 种不同因素的组合做出综合性的了解, 从而全面评价其整个人格。

1.3　其他常用人格模型

Myers-Briggs 类型指标 (MBTI)[9]: 包含 4 个维度, 即内向–外向、感觉–直觉、思维–情感、判断–感知。4 个维度如同四把标尺, 每个人的性格都会落在标尺的某个点上, 这个点靠近哪个端点, 就意味着个体偏向哪一方面。

明尼苏达多项人格调查表 (MMPI)[10]: 通过测试人回答的问题选择是、否的个数来测试其人格; 艾森克人格问卷 (EPQ)[11] 包括 3 个维度 (E 为内外向性、N 为神经质、P 为精神质) 和一个效度量表 (L), 通过对 E、N、P 得分的不同, 综合得出测试者的人格。

因为这些人格测试模型的适用人群不同, 所以获得的效果也有所不同, 近年来, 随着研究的深入, “大五”模型被认为是最适合人类的, 被广泛用于测量人类人格。

2　人格识别数据库

开展听视觉人格识别研究, 首先要建立人格识别数据库。近年来, 研究者已经相继建立了一些人格识别数据库, 其中代表性的数据库有 8 个, 如表 1 所示。YouTube vlogs[12-14]: 该数据集介绍了 vlogs 作为一种丰富的人与人之间的交互, 这种交互方式本质上是多模态的, 适合于新的大规模行为数据分析。YouTube vlogs[13] 数据集包含 2 269 个视频, 视频长度在 1~6 min, 共计 150 h, 来自 469 个不同的 vloggers。该数据集包含 2009 年收集的视频元数据 (包括性别、年龄、笑的出现次数等) 和观众评论。录音设置是参与者正在与能够显示参与者头部和肩膀的摄像机交谈, 其中, 录制内容包含个人视频博客、影片、产品评论等各种主题内容。

表 1　代表性的人格数据库

Table 1　Representative personality database

数据库名称	年份	简要介绍	中心内容	标签	模态
YouTube vlogs[12-14]	2011	2 269个视频, 视频长度在1~6 min, 共计150 h, 来自469个vloggers	会话视频博客与表象人格特质分析	“Big-Five”印象	多模态
ELEA[15]	2012	40次会议: 每次约15 min, 时长为10 h, 148名参与者	小组互动和应急领导	“Big-Five”印象	听视觉
SEMAINE[17]	2012	959次对话: 每次约5 min, 150名参与者	与敏感的人工听众代理进行面对面对话	5个情感维度和27个相关类别	听视觉
SSPNet[1]	2012	322名发言者的640个音频片段, 从新闻中随机抽取	表象人格特质	“Big-Five”印象	听觉
ChaLearn First Impression V1[4]	2016	10K短视频: 每段约15 s收集自2 762名YouTube用户	表象人格特质分析(单人对着摄像机说话)	“Big-Five”印象	听视觉
ChaLearn First Impression V2[19]	2017	数据集[4]的扩展版	表象人格特质与可雇佣性印象	“Big-Five”印象、工作面试	多模态
Physiognomy[20]	2017	186人的面部照片	表象人格特质与人脸图像之间的相关性	卡特尔16人格因素(16PF)	视觉
MHHRI[21]	2017	12次互动会议(约4 h), 18名参与者	HHI和HCI期间的人格和参与度	自我/熟人评估“Big-Five”, 以及参与度	多模态

ELEA[15]: 该数据集由 40 个会议组成, 每次约 15 min, 时长为 10 h。它由 28 个 4 人会议和 12 个新组建团体的 3 人会议组成, 即由以前不认识的人组成。共有 148 名参与者 (48 名女性, 100 名男性), 平均年龄 25.4 岁 (标准差 5.5)。ELEA 会议的所有参与者都被要求参与冬季生存挑战任务, 但没有被分配特殊的角色[16]。使用麦克风采集音频, 音频采样率为 16 kHz。有关录像带设置有两种: 第一种是静态设置, 包括 6 个摄像头, 视频帧率为 25 f/s; 第二种是便携式设置, 包括两个摄像头, 视频帧率为 30 f/s。

SEMAINE[17]: 该数据集采用敏感人工监听 (sensitive artificial listener, SAL) 方式进行录制对话。它可以让一个人进行持续的、带有情感色彩的对话。高质量的录音由 5 个高分辨率、高帧频的摄像机和 4 个麦克风同步录制。录音包含 150 名参与者 (57 名男性和 93 名女性), 平均年龄 32.8 岁。固态 SAL(使操作员做出合适的非语言行为) 和半自动 SAL(用户的体验接近于与机器交互) 的代表性对话持续时间约为 30 min。共收集到 959 个与个人 SAL 角色的对话, 每个对话持续约 5 min。自动 SAL 对话持续近 1 h, 每 3 min 8 个角色进行交互。所有参与者与两个版本的系统进行互动, 间隔 10~15 min 完成心理测量。

SSPNet[1]: 该说话人语料库是从语音中进行人格特质评估的数据集。它包括 322 个发言者的 640 个音频片段, 是从瑞士的法国新闻公报中随机抽取。所有的音频片段都是以 8 kHz 的频率采样的, 大部分都是 10 s, 有些更短。此外, 还邀请 11 名法官 (不熟悉法语, 不受语言线索的影响) 通过填写 BFI-10 个性评价问卷[18], 对每一个片段进行注释。在调查问卷的基础上, 计算出每个大五人格特质的得分。

ChaLearn First Impression V1[4]: 该数据集是由 YouTube 视频中的 10 000 个短片组成, 每段视频分辨率为 1 280×720, 每段时长约 15 s。这些短视频是从约 2 762 个 YouTube 高清视频中收集的。视频是用英文面对和对摄像机说话的人。视频中涉及的人具有不同的性别、年龄、民族和种族。这是迄今为止可用于表象人格分析 (apparent personality recognition) 的较大规模的音视频数据集。ChaLearn First Impression V2[19] 数据集是 ChaLearn First Impression V1[4] 的扩展版, 添加了一个以前没有使用过的预测变量, 即“工作面试”场景进行预测, 并且提供了与视频相对应音频信息的手动转录。

Physiognomy[20]: 该数据集是用来研究人格特质与人脸图像之间的相关性。它包括 186 人 (94 名男性和 92 名女性) 的面部照片。参与者被要求坐在白色背景前, 用中性的面部表情拍照。此数据集是为东亚人种设计的, 不同于现有的针对白种人的研究[21]。

MHHRI[22]: 该数据集旨在同时研究人−人−机器人交互 (HHI) 和人−机器人交互 (HRI) 中的人格特质。它包含 18 名参与者 (9 名女性), 其中大部分是研究生和研究人员。包括 12 段互动对话, 时长约 4 h。每次互动对话都有 10~15 min。对话使用以自我为中心的两个静态和两个动态摄像头以及两个生物传感器记录。另外, 参与者需要佩戴一个 Q-传感器, 配有 Effectiva 设备来记录生理信号。

3　特征提取

针对听视觉信息的人格特征提取是人格识别研究的一个关键问题。目前有关人格特征提取主要有两种: 手工设计的听视觉人格特征 (语音人格特征和视觉人格特征) 和采用近年来发展起来的深度学习技术[23-27] 进行提取的深度听视觉特征, 具体表述如下。

3.1　语音人格特征提取

对于基于听觉信息的人格特质识别, 它涉及两个关键部分: 特征提取和人格特质预测的分类器, 如支持向量机 (SVM) 和线性回归器。这里我们将重点介绍低层次的手工特征提取和高层次的深度特征提取。表 2 简要总结了基于听觉信息的人格识别情况, 详细内容如下。

3.1.1　手工语音特征

语音信号主要包括语义信息和声学信息。目前常提取的低层次的音频特征是手工制作的低层描述 (Low-level descriptors, LLD) 特征, 主要有三类: 韵律特征、音质特征以及谱特征。韵律特征包括基频 (pitch)、能量等, 音质特征包括共振峰、声道参数等, 而谱特征包括梅尔频率倒谱系数 (Mel frequency cepstrum coefficient, MFCC) 等。

Mohammadi 等[1] 提取韵律特征 (如基频、能量、有声段和无声段的长度) 和音质特征 (如前两个共振峰), 对于说话人人格语料库 SSPNet 中的 322 名参与者进行了 640 多个语音片段 (10 s 内) 的实验测试。每个语音片段的评估人数为 11 人, 他们采用 Logistic 回归和 SVM 分类器来识别一个音频片段是否超过了大五人格特质中每个人的平均得分。

Mairesse 等[28] 采用了与文献 [1] 相同的韵律

特征, 使用大五人格特质用于人格识别。实验测试是在语料库 EAR[29] 中的 96 名参与者身上进行。人格评估分数是通过对个体分配的分数进行平均得到, 每份样本由 6 名独立评估员组成。实验旨在预测参与者确切的人格特质分数, 并根据预测的分数对参与者进行排序。基于听觉和文本对所有五大人格特质的识别的实验结果中, 通过返回观察到的平均得分的方法, 最好的结果 (针对外向性和神经质) 是减少大约 15% 的错误率。

Valente 等[30] 通过韵律特征 (说话速率、基频平均值、最小值、最大值、中值和标准差等) 等, 在一个包括 128 名参与者的会议场景中工作的会议语料库进行了实验, 通过 SVM 分类器对大五人格特质进行分类。Ivanov 等[31] 提取包含韵律特征、音质特征和谱特征等 6 552 个声学特征, 在包含 12 个人的 119 个语音样本的数据集进行说话人的“大五”人格的分类。Levitan 等[32] 提取包含韵律特征、音质特征和谱特征等 6 373 个声学特征, 在包含 172 个人的 1 225 个语音样本数据集进行大五人格的识别。

表 2　基于听觉的人格识别总结

Table 2　Summary of audio-based personality recognition

语音特征	时间	作者	简介
手工语音特征	2007	Mairesse等[28]	韵律特征(基频、能量等)和音质特征(共振峰)
	2011	Ivanov等[31]	韵律特征、音质特征和谱特征等6 552个声学特征
	2012	Mohammadi等[1]	韵律特征(基频、能量等)和音质特征(共振峰)
	2012	Valente等[30]	韵律特征(说话速率、基频平均值、最小值等)
	2016	Levitan等[32]	韵律特征、音质特征和谱特征等6 373个声学特征
	2020	Carbonneau等[33]	基于特征学习和谱图分析的方法
深度语音特征	2017	Su等[39]	基于小波多分辨率分析和CNN相结合的方法
	2018	Zhu等[40]	提出了一种跳帧(skip-frame)LSTM系统
	2019	Hayat等[38]	基于CNN的语音人格特征提取方法

Carbonneau 等[33] 提出了一种基于特征学习和谱图分析的方法, 在保持高精度的同时简化了特征提取过程。所提出的方法从训练语音段的谱图中提取的块中学习了一个判别式特征字典。然后使用该字典对每个语音段进行编码, 随后用 SVM 分类器对大五人格特质进行分类。

3.1.2　深度语音特征

近年来, 深度学习方法被广泛用于基于语音信号的人格识别领域。本质上, 深度学习方法的目标是通过使用多个非线性变换的层次结构来实现高层的抽象特征表示。因为低级特征数量有限并且不能完整描述语音信号, 研究者尝试利用深度学习方法从低级特征中学习提取高级的深度属性特征。常用深度学习方法有卷积神经网络 (convolutional neural network, CNN)[34]、深度信念网络 (deep believe network, DBN)[35]、循环神经网络 (recurrent neural networks, RNN)[36] 等。

CNN 最初由 LeCun 等于 1998 年提出, 在 2012 年被发展成一种深度高级版本 (AlexNet)。CNN 的基本结构包括输入层 (input layer)、卷积层 (convolutional layer)、池化层 (pooling layer)、全连接层 (full connection layer) 和输出层 (output layer)。DBN 是由 Hinton 等在 2006 年提出的一种生成模型, 其目的是捕捉输入数据的高阶分布特征表示。RNN 是一种捕捉时间信息的单一前馈神经网络, 可用来处理序列数据。RNN 包含连接相邻时间步的递归边缘, 从而提供了模型中时间的概念。长期短时记忆 (long short term memory, LSTM)[37] 是由 Hochreiter 等于 1997 年提出的一种改进的 RNN 结构。LSTM 可以缓解 RNN 训练过程中产生的梯度消失和爆炸问题。

Hayat 等[38] 提出一种基于卷积神经网络 (CNN) 的语音人格特征提取方法, 通过采用 CNN 学习音频特征来预测说话人的五大人格特质得分。他们采用一个在现有大型语料数据库上预训练好的 CNN 模型 (AudioSet), 在目标第一印象人格数据集上进行微调, 从而提取用于人格特质识别的高层次音频特征表示。实验结果表明, 采用这种 CNN 学习到的深度特征获得的性能优于手工特征方法。

Su 等[39] 研究采用一种基于小波多分辨率分析和 CNN 相结合的方法, 用于学习语音信号的人

格特征。该方法首先采用小波变换将语音信号分解为不同分辨率的信号，然后提取各分辨率下语音信号的声学特征。随后，利用 CNNs 生成 BFI-10 的轮廓进行量化测度，最后利用人工神经网络进行人格特质识别。

Zhu 等[40] 提出了一种跳帧 (skip-frame) LSTM 系统，用于实现从普通话语音中自动识别说话者的人格。所研究的人格特质从标准的大五人格特质扩展到每种人格特质的 6 个子特征，总共有 30 个人格子特征。该 LSTM 系统利用跳帧采样来增加训练数据，同时长时间保持韵律变化。LSTM 系统直接从 MFCCs 中学习人格特质信息，而不是像采用 SVM 的系统那样需要手动设计韵律特征。实验结果表明，外向性 (extraversion) 特质最容易被识别，而开放性 (openness) 特质最难被识别。

3.2　视觉人格特征提取

根据视觉输入数据的类型，基于视觉的人格特质识别可分为静态图像和动态视频序列。Junior 等[3] 研究利用静态图像进行自动人格识别实验。这类实验通常关注面部信息来驱动模型，通常是将不同层次的特征和它们之间的关系结合起来。通过手工提取的低层次视觉特征包括方向梯度直方图 (histogram of oriented gradient, HOG)[41]、局部二值模式 (local binary pattern, LBP)[42]、尺度不变特征变换 (scale-invariant feature transform, SIFT)[43] 等，而深度视觉特征是采用深度学习方法从低级图像中提取的高层次视觉属性特征。表 3 简要总结了基于视觉信息的人格识别情况，详细内容如下。

表 3　基于视觉的人格识别

Table 3　Summary of visual-based personality recognition

视觉特征	时间	作者	简介
静态图像	2015	Guntuku等[47]	颜色直方图, 局部二值模式(LBP)
	2016	Dhall等[44]	梯度直方图金字塔(PHOG), 局部相位量化(LPQ)
	2016	Yan等[48]	方向梯度直方图(HOG), 局部二元模式(LBP)
	2016	聂婕等[49]	颜色、纹理、形状等特征
	2017	Zhang等[20]	微调预训练好的VGG模型
	2019	Tareaf等[50]	从个人资料图像中提取50个独特的面部特征
动态视频序列	2012	Biel等[51]	基于帧的面部活动统计
	2014	Teijeiro等[53]	提取视频行为信息
	2016	Gürpnar等[56]	微调预训练的VGG-19模型
	2017	Ventura等[54]	一种描述符聚合网络(DAN)
	2018	Gatica等[52]	为每个视频收集了21个印象变量
	2019	Beyan等[57]	从关键动态图像中提取基于深度视觉活动的特征

3.2.1　手工视觉特征

1) 静态图像：在基于视频的人格特征提取中，主要侧重于人格相关的视觉人脸图像特征的提取。静态图像是既不使用音频信息也不使用时间信息的静止图片，在经过一系列的预处理 (如旋转校正、人脸定位等) 之后作为后续处理过程的对象。手工从图像中提取低层特征主要有方向梯度直方图 (HOG)、局部二值模式 (LBP)、尺度不变特征变换 (SIFT) 等。

Dhall 等[44] 提出了一种采用计算机视觉方法实现从用户的 Twitter 资料图片中推断用户的人格特质。这类似于人类仅通过查看另一个人的资料图片就会对其产生心理印象。该方法首先采用了梯度直方图金字塔 (pyramid of histogram of gradients, PHOG)[45]、局部相位量化 (local phase quantisation, LPQ)[46] 等手工制作特征描述符，这些描述符是在用户资料图片上计算出来的。然后采用核偏最小二乘法 (kernel partial least square, KPLS) 回归，最后预测大五人格特质。

Guntuku 等[47] 提出采用人脸图像的颜色直方图、局部二值模式等低层次的视觉特征，用于检测性别、年龄等线索，在包含 123 个人的 123 幅图像样本进行大五人格的分类。Yan 等[48] 研究了外貌与人格印象之间的关系。他们从不同的人脸区

域以及区域之间的关系中提取不同的低层特征。例如, 方向梯度直方图 (HOG) 用于描述眉毛的形状, 而欧式距离用于描述眼睛的宽度。为了缓解低层特征和高层特征之间的语义差异, 通过聚类的方法从低层次的方向梯度直方图 (HOG)、局部二元模式 (LBP)、尺度不变特征变换 (SIFT) 等视觉特征中提取中间层次的特征, 然后利用支持向量机 (SVM) 实现包含 250 个人的 2 010 幅图像样本的大五人格的识别。

聂婕等[49] 提出采用 5 种视觉特征, 包括颜色、纹理、形状、伊顿对比 (Itten contrast) 和表情特征, 来构建“大五”人格模型, 在包含 64 个人的 2 000 幅图像样本中实现人格的自动预测。Tareaf 等[50] 研究了在社交媒体上发布的个人资料图片与用户人格的关系。他们使用来自推特平台的个人资料图像, 根据 170 万个数据点预测了他们的人格。他们通过提取 50 个独特的面部特征对用户的面部进行分析, 以检验人格和个人资料图片之间的关系。实验结果表明不同人格之间在个人资料图片选择上的显著差异。

2) 动态视频序列: 动态视频序列由一系列视频图像帧组成, 从而提供时间信息和场景动态。基于动态视频序列的特征提取方法与静态图像的不同之处在于动态序列图像反映了人脸表情运动的连续过程, 因此动态视频序列的表情特征主要由人脸的形变及面部各个区域的肌肉运动上体现出来。

Biel 等[51] 利用 Youtube vlog 数据集的一个子集, 研究了对话视频 (vlogs) 中的人格印象, 重点是面部表情分析。他们在逐帧估计的基础上, 通过结合面部活动统计, 再利用 SVM 分类器实现了自动对人格特质进行预测。实验结果表明, 外向性是活动特征线索利用率最高的特征。Gatica 等[52] 提出了一项使用 vloggers 的行为数据的研究方法, 即使用同一用户的多个视频有助于在表象人格预测中达到更好的效果。这些 vloggers 在 YouTube 上发布 vlogs 长达 3~6 年。他们为每个视频收集了 21 个印象变量, 包括感知的人格、情绪、技能和专业知识。

Teijeiro 等[53] 研究了一项关于自动提取面部表情与 YouTube vlogs 中大五人格特质印象之间的联系。他们使用计算机表情识别工具箱 (computer expression recognition toolbox, CERT) 系统来描述对话式 vlogs 的用户特征。从即时识别的面部表情类别的 CERT 时间信号中, 他们提出采用 4 组行为特征线索来描述人脸的统计和动态特征。这些特征线索首先被用于相关性分析, 以评估每个面部表情与从观看 vlogs 的人群参与者获得的大五人格特质印象的相关性, 同时也作为自动人格特质预测的特征。最后, 利用 SVM 分类器来对人格特质进行预测。实验结果表明, 当多个面部表情特征线索与一定数量的大五人格特质显著相关时, 它们只能明显地预测外向性的特定特征。

3.2.2　深度视觉特征

目前, 深度学习方法广泛用于人格识别领域, 用以提取高层次的视觉特征, 常用的有 CNN、RNN、LSTM 等方法。

1) 静态图像: 对于静态图像的深度特征提取, Zhang 等[20] 提出采用 CNN 用于实现含有 186 个人的 186 幅图像样本的人格自动识别, 用人脸来评估一个人的人格特质和智力。他们最初建立了一个由面部照片、人格测量和智力测量组成的数据集, 然后提出了一种端到端的 CNN 模型, 通过微调预先训练好的 VGG-face 模型来共同预测人格特质和智力。他们旨在探讨自我报告的人格特质和智力是否可以从面部图像中共同测量。实验结果表明, CNN 特征在预测人格特质方面优于传统的人工特征。

Ventura 等[54] 使用一种描述符聚合网络 (descriptor aggregation networks, DAN)[55] 来学习面部特征。DAN 是改进了的传统 CNN 模型 (首先删除了全连接层, 然后对最后一个卷积层的深度描述符通过平均池化和最大池化进行聚合, 最后级联到最终的图像表示中进行回归)。他们利用类激活图 (class activation map, CAM) 进行可视化, 从而为理解 CNN 模型成功学习与用户人格特质相关的面部特征 (如眼睛、鼻子、眉毛和嘴巴) 提供了可能的解释。

2) 动态视频序列: 对于动态视频序列的深度特征提取, Gürpnar 等[56] 通过微调一个预先训练好的 VGG-19 网络来提取人脸面部表情以及周围场景的深层特征, 以便进行人格特质分析。然后, 将表示面部表情和场景的视觉特征进行合并, 输入到核极限学习机 (kernel extreme learning machine, ELM) 回归器中, 并在 ChaLearn First Impression V1 数据集上实现大五人格特质的预测。

Ventura 等[54] 从模型可解释性的角度研究了 CNN 模型在人格自动预测方面表现很好的原因。他们利用当前关于 CNN 模型可解释性的技术 (如可视化), 结合人脸检测和动作单元 (action unit, AUs) 识别系统进行定量研究。实验结果表

明：1) 人脸为人格特质预测提供了大部分的判别特征信息；2) CNNs 表征主要分析关键的人脸区域，如眼睛、鼻子和嘴巴等；3) 部分动作单元对面部特征的预测有一些影响。

Beyan 等[57] 通过从来自视频中的关键动态图像 (主要用于动作、活动和手势识别) 中提取的基于深度视觉活动 (visual activity, VA) 的特征来感知人格特质。由于关键动态图像带有更多的判别信息，因此他们构建多个动态图像，通过采用 CNN+LSTM 学习长期视觉活动和检测时空显著性来确定关键动态图像，一旦提取了基于视觉活动的非语言特征，就使用基于协方差的特征编码方法，最后得到的特征向量利用 SVM 分类器来对人格特质进行预测。

4　融合听视觉的多模态人格识别

近年来，尽管针对单一模态信息 (听觉或视觉) 的自动人格识别方面的研究取得了一些研究成果，但在实际生活中人类自身的人格判别或分析往往都是多模态的。因此，仅仅通过单一模态信息来分析人格存在诸多的局限性，如识别性能还不尽如人意，识别结果的鲁棒性得不到保证。从 2016 年开始，研究者开始尝试在融合听觉、视觉等信息的视频序列中实现多模态人格识别，研究重点侧重于听觉和视觉方面的人格特征提取，以及多模态信息融合方法。这部分将重点介绍人格识别中的多模态信息融合方法。

在人格识别任务中，面向听视觉信息的多模态融合方法一般分为 3 种类型[58-59]：特征层 (feature-level) 融合、决策层 (decision-level) 融合、模型层 (model-level) 融合。

1) 特征层融合：该方法是一种比较简单的方法，计算复杂程度相对较低。它只需要将提取的听觉特征和视觉特征直接串联起来构成一个总的特征向量，所以特征层融合也称为早期融合 (early fusion, EF)。不过，特征层融合可能会显著增加级联特征向量的维数，容易出现维数灾难问题。

Güçlütürk 等[60] 提出采用一种端到端的深度残差网络来学习高层次的听觉语音特征和视觉人脸图像特征，然后在特征层 (feature-level) 实现多模态的大五人格识别。该方法在包含 10 000 个视频样本的 ChaLearn First Impression V1 挑战数据集上取得了较好的效果。Subramaniam 等[61] 采用两种端到端的深度学习模型 (3D CNN 和 LSTM) 进行视听第一印象分析。(他们利用 3D CNN 对人脸对齐图像进行视觉特征提取。而对于听觉信息，他们提取过零率、能量、MFCCs 等手工特征的均值和标准差等统计量作为听觉特征参数。然后，他们将提取的听视觉特征在特征层进行级联，然后使用 LSTM 网络进行时间建模，完成最终的人格特质预测任务。

Wei 等[62] 提出一种从短视频序列中进行深度特征学习的解决方案。为了从听视觉模态中提取丰富的信息，采用深度双峰回归 (deep double peak regression, DBR) 方法来完成听视觉特征提取任务。在 DBR 中，对于视觉模态，他们使用 DAN 模型，用于提取重要的视觉特征信息，然后通过端到端训练获得五大人格特质预测。对于听觉模态，他们从每个原始的以人为中心的视频中提取 MFCC 和对数滤波器组 (log filter bank, logfbank) 等听觉特征，基于这些提取的听觉特征，采用线性回归方法实现人格预测。为了更好地结合两种模态的互补信息，他们采用特征层融合方法来整合这些预测的回归分数，以便获得最佳人格预测性能。

Güçlütürk 等[63] 使用一种深度残差网络提取视听特征，采用 skip-thought 的向量模型提取文本 (音频转录) 特征。然后将提取的音频、视觉和文本特征在特征层面进行多模态大五人格特质分析和工作面试推荐。Escalante 等[65] 提出从视频序列中融合听觉、视觉和文本 3 个模态信息用于研究第一印象分析的可解释性。该方法采用 ResNet18 模型用于提取听视觉特征，而采用 skip-thought 向量模型提取文本特征，最后，在特征层上融合 3 个模态提取的特征用于多模态第一印象的分析。

2) 决策层融合：该方法首先对每个模态先独立建模，然后采用某种决策融合规则将单模态得到的结果进行组合，并得到最终的融合结果。因此，决策层融合也称为后期融合 (Late Fusion, LF)。决策层融合认为不同的模态是相互独立的，但它未能利用特征层模式之间的相关性。分数层 (score-level) 融合是决策层融合的一种变体。它通过组合各个类别分数来实现的，这些分数代表了一个样本属于各种类别的概率，可以用于人格识别。而决策层融合则是通过结合若干预测性类别标签来实现。

Celiktutan 等[64] 探讨了人格印象如何随时间和情境环境的变化而波动。首先提取听视觉特征 (如面部、头部、身体运动等)，然后采用双向 LSTM 网络对连续生成的注释与提取特征之间的时间关系进行建模。最后在决策层上将听觉和视觉回归预测模型的输出结果相结合，从而实现人

格的预测。Gorbova 等[65-66] 提出一种基于短视频的视觉、听觉和文本 (词汇) 信息的自动人格预测方法。该方法提取的手工特征包括声学 LLD 特征 (MFCCs、说话速率等)、面部动作单元特征、以及消极和积极的词汇得分。该系统采用加权平均策略, 将 3 种模态取得的结果在决策层加以融合, 获取最后的人格预测结果。

Zhang 等[67-68] 提出采用语音信号的 MFCCs 特征作为听觉特征, 而视觉特征采用 DAN 模型提取, 然后在特征层 (feature-level) 和决策层 (decision-level) 上融合听视觉两个模态。在包含 3 000 个人的 10 000 个短视频的 ChaLearn First Impression V1 挑战数据集上的实验结果表明, 该方法能够取得较好的多模态人格识别结果。Sarkar 等[71] 使用 logistic 回归模型, 结合听视觉、语言内容和情绪特征等进行大五人格特质分类。实验结果表明, 使用不同的特征组合可以更好地预测不同的人格特质。

Gürpinar 等[69] 使用预先训练的 VGG 模型从视觉图像中提取面部表情和场景信息, 而从听觉语音信号中提取 INTERSPEECH-2009 特征, 然后分别采用核极限学习机 (Kernel ELM) 实现人格预测, 最后采用分数层 (score-level) 方法融合这些不同模态信息的人格预测结果。

3) 模型层融合: 该方法作为特征层融合和决策层融合的一种折中方案, 近年来也被用于人格识别。这种方法的目的是在考虑模态间相关性的同时, 分别实现对每个模态的建模。它可以考虑不同模态之间的相互关联性。

Principi 等[70] 研究了影响人格感知不同可能因素源的影响, 包括来自面部表情、吸引力、年龄、性别和种族等因素。他们提出了一种多模态深度神经网络模型的听视觉人格识别方法。该方法将原始的听觉和视觉信息相结合, 用于测试特定属性模型 (attribute-specific models), 在大五人格特质预测方面的性能。对于视觉特征提取, 他们采用了在 ImageNet 数据上预先训练好的 ResNet-50 网络, 在每个视频帧上获取高层次的视觉特征表示。对于听觉特征提取, 采用类似 ResNet-18 网络这样的 14 层一维卷积神经网络 (1D CNN) 从原始语音频谱信号学习高层次的听觉特征表示。对于视频级 (video-level) 属性特征的提取, 采用 VGG-16 网络从视频图像中学习出面部表情、吸引力、年龄、性别和种族等因素各自对应的特征向量。为了有效融合各种提取的特征, 采用两步来实现: 1) 采用一个全连接层在模型层上学习所有提取的视频级属性特征串联之后的联合特征表示, 同时降低其特征维数; 2) 将学习到的联合特征表示与之前提取的听觉、视觉特征相串联, 输入到一个全连接层实现大五人格特质预测。

Kampman 等[71] 提出了一种端到端的融合听觉、视觉和文本三模态深度学习模型来预测大五人格特质。对于听觉通道, 为将幅值平方的原始音频波形及其能量分量输入到一个包含 4 个卷积层和一个全局平均池化层的 CNN 网络用于音频特征提取。对于视觉通道, 选取视频随机帧图像微调预训练好的 VGG-16 模型用于视频特征提取。对于文本通道, 采用“Word2vec”字嵌入 (Word embedding) 模型的输出作为文本 CNN 网络的输入, 用于文本特征提取。最后, 在决策层和模型层上实现听觉、视觉和文本模式的融合。决策层融合采用了投票方法, 而模型层融合是通过串联每个模态 CNN 的输出特征输入到一个包含两个全连接层的网络, 用于学习三模态输入数据的联合特征表示。

5　结束语

人格识别是一个涉及多学科交叉的研究课题, 其中基于听视觉信息的人格识别近年来成为了计算机视觉领域的一个研究热点。本文详细介绍了国内外现有的近年来用于人格识别的人格类型表示理论和相关数据集, 并重点阐述了近年来新发展起来的深度学习技术在人格特征提取方面的应用。同时, 也对人格识别中的多模态信息融合方法做了整理和归纳, 并给出了该领域未来的发展趋势。虽然, 人格识别研究已经取得了一些成果, 但在许多方面还存在一些挑战。例如, 用于人格识别的现有建设的数据集规模都不大, 还不能很好地满足现有依靠大数据驱动的深度学习技术的训练需求, 未来可针对多模态人格识别方法在跨数据集环境下的使用进行研究。另外, 人格识别是多模态的, 目前很少有研究者关注生理信号与现有听觉、视觉等模态信号相结合的多模态人格识别方法, 未来如何将生理信号和其他模态结合是一个新的研究方向。此外, 现有研究也很少考虑采用被观察者与人格分析更多的信息来进行自动人格识别, 如考虑不同目标人群的文化相似性或者差异性等背景信息, 以改善人格识别模型的性能。

参考文献:

[1] VINCIARELLI A, MOHAMMADI G. A survey of person-

ality computing[J]. IEEE transactions on affective computing, 2014, 5(3): 273–291.

[2] COSTA P T, MCCRAE R R. Trait theories of personality[M]//BARONE D F, HERSEN M, VAN HASSELT V B. Advanced Personality. Boston: Springer, 1998: 103−121.

[3] JUNIOR J C S J, GÜÇLÜTÜRK Y, PÉREZ M, et al. First impressions: a survey on vision-based apparent personality trait analysis[J]. IEEE transactions on affective computing, 2019: 1−1.

[4] PONCE-LÓPEZ V, CHEN Baiyu, OLIU M, et al. ChaLearn LAP 2016: first round challenge on first impressions-dataset and results[C]//Proceedings of European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 400−418.

[5] ESCALANTE H J, KAYA H, SALAH A A, et al. Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos[J]. IEEE transcations on affective computing, 2020: 1−1.

[6] MATTHEWS G, DEARY I J, WHITEMAN M C. Personality traits[M]. 2nd ed. Cambridge: Cambridge University Press, 2003.

[7] MCCRAE R R, JOHN O P. An introduction to the five-factor model and its applications[J]. Journal of personality, 1992, 60(2): 175–215.

[8] KARSON S, O'DELL J W. A guide to the clinical use of the 16 PF[M]. Champaign, IL: Institute for Personality & Ability Testing, 1976.

[9] FURNHAM A. The big five versus the big four: the relationship between the Myers-Briggs Type Indicator (MBTI) and NEO-PI five factor model of personality[J]. Personality and individual differences, 1996, 21(2): 303–307.

[10] GREENE R L. The MMPI: an interpretive manual[M]. New York: Grune & Stratton Inc, 1980.

[11] EYSENCK H J, EYSENCK S B G, EYSENCK H J, et al. Manual of the eysenck personality questionnaire[J]. Journal of cardiac failure, 1975, 20(5): S67.

[12] BIEL J I, GATICA-PEREZ D. Voices of vlogging[C]// Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media. Washington, USA, 2010.

[13] BIEL J I, GATICA-PEREZ D. The youtube lens: crowdsourced personality impressions and audiovisual analysis of vlogs[J]. IEEE transactions on multimedia, 2013, 15(1): 41–55.

[14] BIEL J I, GATICA-PEREZ D. Vlogcast yourself: nonverbal behavior and attention in social media[C]//Proceedings of the International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction. Beijing, China, 2010: 1−4.

[15] SANCHEZ-CORTES D, ARAN O, MAST M S, et al. A nonverbal behavior approach to identify emergent leaders in small groups[J]. IEEE transactions on multimedia, 2012, 14(3): 816–832.

[16] KICKUL J, NEUMAN G. Emergent leadership behaviors: the function of personality and cognitive ability in determining teamwork performance and KSAs[J]. Journal of business and psychology, 2000, 15(1): 27–51.

[17] MCKEOWN G, VALSTAR M, COWIE R, et al. The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent[J]. IEEE transactions on affective computing, 2012, 3(1): 5–17.

[18] RAMMSTEDT B, JOHN O P. Measuring personality in one minute or less: a 10-item short version of the Big Five Inventory in English and German[J]. Journal of research in personality, 2007, 41(1): 203–212.

[19] ESCALANTE H J, GUYON I, ESCALERA S, et al. Design of an explainable machine learning challenge for video interviews[C]//Proceedings of 2017 International Joint Conference on Neural Networks. Anchorage, USA, 2017: 3688−3695.

[20] ZHANG Ting, QIN Rizhen, DONG Qiulei, et al. Physiognomy: personality traits prediction by learning[J]. International journal of automation and computing, 2017, 14(4): 386–395.

[21] CELIKTUTAN O, SKORDOS E, GUNES H. Multimodal human-human-robot interactions (MHHRI) dataset for studying personality and engagement[J]. IEEE transactions on affective computing, 2019, 10(4): 484–497.

[22] OOSTERHOF N N, TODOROV A. The functional basis of face evaluation[J]. Proceedings of the national academy of sciences of the United States of America, 2008, 105(32): 11087–11092.

[23] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436–444.

[24] WANG Xizhao, ZHAO Yanxia, POURPANAH F. Recent advances in deep learning[J]. International journal of machine learning and cybernetics, 2020, 11(4): 747–750.

[25] GAO Jing, LI Peng, CHEN Zhikui, et al. A survey on deep learning for multimodal data fusion[J]. Neural computation, 2020, 32(5): 829–864.

[26] CHOI Y, EL-KHAMY M, LEE J. Universal deep neural network compression[J]. IEEE journal of selected topics in signal processing, 2020, 14(4): 715–726.

[27] ANGELOV P, SOARES E. Towards explainable deep neural networks (xDNN)[J]. Neural networks, 2020, 130: 185–194.

[28] MAIRESSE F, WALKER M A, MEHL M R, et al. Using linguistic cues for the automatic recognition of personal-

ity in conversation and text[J]. Journal of artificial intelligence research, 2007, 30(1): 457–500.

[29] MEHL M R, GOSLING S D, PENNEBAKER J W. Personality in its natural habitat: manifestations and implicit folk theories of personality in daily life[J]. Journal of personality and social psychology, 2006, 90(5): 862–877.

[30] VALENTE F, KIM S, MOTLICEK P. Annotation and recognition of personality traits in spoken conversations from the ami meetings corpus[C]//Proceedings of Interspeech 2012. Portland, USA, 2012.

[31] IVANOV A V, RICCARDI G, SPORKA A J, et al. Recognition of personality traits from human spoken conversations[C]//Proceedings of Interspeech 2011. Florence, Italy, 2011: 1549−1552.

[32] AN G, LEVITAN S I, LEVITAN R, et al. Automatically classifying self-rated personality scores from speech[C]//Proceedings of Interspeech 2016. San Francisco, USA, 2016: 1412−1416.

[33] CARBONNEAU M A, GRANGER E, ATTABI Y, et al. Feature learning from spectrograms for assessment of personality traits[J]. IEEE transactions on affective computing, 2020, 11(1): 25–31.

[34] LECUN Y, BOTTOU L, BENGIO Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278–2324.

[35] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006, 18(7): 1527–1554.

[36] ELMAN J L. Finding structure in time[J]. Cognitive science, 1990, 14(2): 179–211.

[37] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[38] HAYAT H, VENTURA C, LAPEDRIZA À. On the use of interpretable CNN for personality trait recognition from audio[C]//Proceedings of CCIA. Mallorca, Spain, 2019: 135−144.

[39] SU M H, WU C H, HUANG Kunyi, et al. Personality trait perception from speech signals using multiresolution analysis and convolutional neural networks[C]//Proceedings of 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Kuala Lumpur, Malaysia, 2017: 1532−1536.

[40] ZHU Minxian, XIE Xiang, ZHANG Liqiang, et al. Automatic personality perception from speech in mandarin[C]//Proceedings of 2018 11th International Symposium on Chinese Spoken Language Processing. Taipei, China, 2018: 309−313.

[41] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886−893.

[42] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(7): 971–987.

[43] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91–110.

[44] DHALL A, HOEY J. First impressions-predicting user personality from twitter profile images[C]//Proceedings of the 7th International Workshop on Human Behavior Understanding. Amsterdam, The Netherlands, 2016: 148−158.

[45] BOSCH A, ZISSERMAN A, MUNOZ X. Representing shape with a spatial pyramid kernel[C]//Proceedings of the 6th ACM International Conference on Image and Video Retrieval. Amsterdam, The Netherlands, 2007: 401−408.

[46] OJANSIVU V, HEIKKILÄ J. Blur insensitive texture classification using local phase quantization[C]//Proceedings of the 3rd International Conference on Image and Signal Processing. Cherbourg-Octeville, France, 2008: 236−243.

[47] GUNTUKU S C, QIU Lin, ROY S, et al. Do others perceive you as you want them to?: modeling personality based on selfies[C]//Proceedings of the 1st International Workshop on Affect & Sentiment in Multimedia. Brisbane, Australia, 2015: 21−26.

[48] YAN Yan, NIE Jie, HUANG Lei, et al. Exploring relationship between face and trustworthy impression using mid-level facial features[C]//Proceedings of 22nd International Conference on Multimedia Modeling. Miami, USA, 2016: 540−549.

[49] 聂婕, 黄磊, 李臻, 等. 基于人物图像视觉特征的人物性格隐私分析 [J]. 通信学报, 2016, 37(11): 129–136.
NIE Jie, HUANG Lei, LI Zhen, et al. Human personality privacy analysis based on visual features[J]. Journal on communications, 2016, 37(11): 129–136.

[50] TAREAF R B, ALHOSSEINI S A, MEINEL C. Facial-based personality prediction models for estimating individuals private traits[C]//Proceedings of 2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom). Xiamen, China, 2019: 1586−1594.

[51] BIEL J I, TEIJEIRO-MOSQUERA L, GATICA-PEREZ D. FaceTube: predicting personality from facial expressions of emotion in online conversational video[C]//Pro-

ceedings of the 14th ACM International Conference on Multimodal Interaction. Santa Monica, California, USA, 2012: 53−56.

[52] GATICA-PEREZ D, SANCHEZ-CORTES D, DO T M T, et al. Vlogging over time: longitudinal impressions and behavior in youtube[C]//Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia. Cairo, Egypt, 2018: 37−46.

[53] TEIJEIRO-MOSQUERA L, BIEL J I, ALBA-CASTRO J L, et al. What your face vlogs about: expressions of emotion and big-five traits impressions in YouTube[J]. IEEE transactions on affective computing, 2015, 6(2): 193–205.

[54] VENTURA C, MASIP D, LAPEDRIZA A. Interpreting CNN models for apparent personality trait regression[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 55−63.

[55] WEI Xiushen, LUO Jianhao, WU Jianxin, et al. Selective convolutional descriptor aggregation for fine-grained image retrieval[J]. IEEE transactions on image processing, 2017, 26(6): 2868–2881.

[56] GÜRPINAR F, KAYA H, SALAH A A. Combining deep facial and ambient features for first impression estimation[C]//Proceedings of the European conference on computer vision. Amsterdam, The Netherlands, 2016: 372−385.

[57] BEYAN C, ZUNINO A, SHAHID M, et al. Personality traits classification using deep visual activity-based nonverbal features of key-dynamic images[J]. IEEE transactions on affective computing, 2019: 1−1.

[58] ATREY P K, HOSSAIN M A, EL SADDIK A, et al. Multimodal fusion for multimedia analysis: a survey[J]. Multimedia systems, 2010, 16(6): 345–379.

[59] ZENG Zhihong, PANTIC M, ROISMAN G I, et al. A survey of affect recognition methods: audio, visual, and spontaneous expressions[J]. IEEE transactions on pattern analysis and machine intelligence, 2009, 31(1): 39–58.

[60] GÜÇLÜTÜRK Y, GÜÇLÜ U, VAN GERVEN M A, et al. Deep impression: audiovisual deep residual networks for multimodal apparent personality trait recognition [C]//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 349−358.

[61] SUBRAMANIAM A, PATEL V, MISHRA A, et al. Bimodal first impressions recognition using temporally ordered deep audio and stochastic visual features[C]//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 337−348.

[62] WEI Xiushen, ZHANG Chenlin, ZHANG Hao, et al. Deep bimodal regression of apparent personality traits from short video sequences[J]. IEEE transactions on affective computing, 2018, 9(3): 303–315.

[63] GÜÇLÜTÜRK Y, GÜÇLÜ U, BARÓ X, et al. Multimodal first impression analysis with deep residual networks[J]. IEEE transactions on affective computing, 2018, 9(3): 316–329.

[64] ÇELIKTUTAN O, GUNES H. Automatic prediction of impressions in time and across varying context: personality, attractiveness and likeability[J]. IEEE transactions on affective computing, 2017, 8(1): 29–42.

[65] GORBOVA J, LÜSI I, LITVIN A, et al. Automated screening of job candidate based on multimodal video processing[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 29−35.

[66] GORBOVA J, AVOTS E, LÜSI I, et al. Integrating vision and language for first-impression personality analysis[J]. IEEE MultiMedia, 2018, 25(2): 24–33.

[67] ZHANG Chenlin, ZHANG Hao, WEI Xiushen, et al. Deep bimodal regression for apparent personality analysis[C]//Proceedings of European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 311−324.

[68] SARKAR C, BHATIA S, AGARWAL A, et al. Feature analysis for computational personality recognition using youtube personality data set[C]//Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition. Orlando, USA, 2014: 11−14.

[69] GÜRPINAR F, KAYA H, SALAH A A. Multimodal fusion of audio, scene, and face features for first impression estimation[C]//Proceedings of 2016 23rd International Conference on Pattern Recognition. Cancun, Mexico, 2016: 43−48.

[70] PRINCIPI R D P, PALMERO C, JUNIOR J C, et al. On the effect of observed subject biases in apparent personality analysis from audio-visual signals[J]. IEEE transactions on affective computing, 2019:1−1.

[71] KAMPMAN O, BAREZI E J, BERTERO D, et al. Investigating audio, visual, and text fusion methods for end-to-end automatic personality prediction[C]//Proceedings of 56th Annual Meeting of the Association for Computational Linguistics. Melbourne,Australia, 2018: 606−611.

作者简介：

赵小明，教授，主要研究方向为音频和图像处理、机器学习和模式识别。

唐志伟, 硕士研究生, 主要研究方向为人格计算和模式识别。

张石清, 教授, 博士, 主要研究方向为情感计算和模式识别。发表学术论文 40 余篇。

CAAI 国际人工智能会议 (CICAI 2021)
CAAI International Conference on Artificial Intelligence (CICAI 2021)

The CAAI International Conference on Artificial Intelligence (CICAI 2021) will be held at Hangzhou, China on June 5th-6th. CICAI is organized by Chinese Association for Artificial Intelligence (CAAI). The aim of CICAI is to promote advanced research in AI, and foster scientific exchange between researchers, practitioners, scientists, students, and engineers in AI and its affiliated disciplines.

CICAI 2021 will be a hybrid conference with both online and in-person presentations.

The program committee of CICAI 2021 invites the submission of papers for the technical program of the conference. High-quality original submissions are welcome from research results and applications of all areas of AI including but not limited to the following areas:

- Brain Inspired AI
- Optimization
- Machine Learning
- Multi-agent Systems
- Computer Vision
- Humans and AI
- Natural Language Processing
- AI Ethics, Privacy, Fairness and Security
- Knowledge Representation and Reasoning
- Explainability, Understandability, and Verifiability of AI
- Data Mining
- Multidisciplinary Research with AI
- Robotics
- Applications of Artificial Intelligence
- AI Ethics, Privacy, Fairness and Security
- Other AI related topics