情感识别与教育

□文 / 余梓彤、李晓白、赵国英

余梓彤

芬兰 Oulu 大学计算机科学与工程专业在读博士。研究方向为 rPPG，人脸活体检测和视频理解。曾在深圳云天励飞技术有限公司担任算法工程师，负责活体检测与行人重识别相关项目。

李晓白

2017 年于芬兰 Oulu 大学机器视觉与信号分析中心（CMVS）获得博士学位，现在 CMVS 继续博士后研究。研究领域包括情感计算、微表情识别、远程生理信号测量等。已发表 20 余篇科研论文，Google Scholar 引用率超过 1400。

赵国英

2005 年毕业于中国科学院计算技术研究所并获得工学博士学位，现为芬兰 Oulu 大学教授。IEEE 高级会员，国际期刊 Pattern Recognition 和 Image and Vision Computing 的编委。研究领域包括计算机视觉、机器学习和情感智能。已发表 200 余篇科研论文，Google Scholar 引用率超过 9700。

近几年，人工智能作为新一代技术，在与教育行业结合的过程中已经取得了初步成效。以人脸识别、语音识别、自然语音处理和知识图谱等为代表的人工智能技术已经应用在教育考勤、机器人助教、知识梳理等教育场景中，在一定程度上实现了智能化教育。但是教育并非一个单方向的机械式知识传输过程，学生自身的情感和动机会在很大程度上影响教育的成果。目前，大部分的智能教育技术缺乏对用户的情感理解，因此在教育应用中的效果受到限制。本文将探讨情感识别和智能化教育之间的关系，并介绍几类研究前沿的情感计算方法及将其应用到智能化教育中的可行性。

一、智能化教育的本质和挑战

得益于大数据时代的海量训练样本和“云 + 端”设备的强大算力，人工智能（AI）算法模型已能精准进行人脸识别、语音识别及自然语言处理，这使得 AI 辅助教育成为可能。智能化教育（AI+ 教育）作为人工智能技术在教育行业的落地，“以‘人’为本”是其核心。这里的“人”泛指参与智能化教育的用户。在参与教育的过程中，用户的学习认知能力与其情感状态是密切相关的。 我们认为，基于情感识别的智能化教育的本质包括以下几个环节：1）实时地理解及识别用户的情感；2）建立用户个人的情感“画像”；3）进行长短期的鼓励及互动，从而最终改善学习体验以及提高学习效率。

在《计算神经科学前沿》杂志（Frontiers in Computational Neuroscience Journal）中，来自人工智能研究中心的三位专家 Luis-Eduardo、Ángeles 和 Félix[1] 提到：“智能化教育中一直被忽略的词就是 Affect（情感）——在线的教学资源非常丰富，并越来越多地在课堂中得到运用，而自适应学习技术也在寻求认知上的最佳路径，但是人们很少考虑到学生的学习能力与情绪状态密切相关。机器人助教如果能够增加情绪识别和鼓励这一维度，将更能增加学习的有效性。”同时，有实验表明，参与和疑惑是在学习过程中最重要和最常出现的情感。对其它控制条件相同的两个智能教育体系进行比较，结果显示，学生成绩在应用情感识别技术的体系中比在未应用情感识别技术的体系中高 91%。

将情感识别技术应用到教育中是一个必要的也充满挑战的任务。现有的人工智能技术在情感识别方面还存在若干难点，比如难以做到精准的情感识别及长期的情绪状态跟踪，这也是智能化教育当下亟待解决的挑战。在传统课堂场景下，由于场景复杂且学生人数众多，传统算法难以精准识别每一位学生的情感，故更难以对每位学生进行个性化教育。在教育软件及助教机器人场景下，有着更多的一对一的人机互动。面对缺乏情感表达的多媒体设备，学生容易走神以及感到学习枯燥，导致失去持续学习的兴趣与动力。如何准确地长期跟踪学生的情绪状态并打造个性化情感多媒体系统，是目前智能教育领域关注的热门课题。除了各大院校和研究机构以外，国内多家公司也正在积极致力于推动人工智能在教育行业的有效落地。其中，基于情感识别的智能化教育尤其受到关注，腾讯、百度和好未来等公司都将其作为未来发展方向之一。

二、基于情感识别的智能化教育

情感（emotion）识别是人类的基本生存能力之一，是人正常社会化交往的基础。人不仅能够识别他人的情绪，同时也能够表达自己的情绪。研究表明，非常幼小的婴儿已经会对不同表情的人脸做出相应的反应 [2]。但是用计算机识别情感只有约二十年的

[1] Luis-Eduardo Imbernón Cuadrado, Angeles Manjarrés Riesco and Félix De La Paz López, ARTIE: An Integrated Environment for the Development of Affective Robot Tutors, Frontiers in Computational Neuroscience. August 3, 2016

[2] Linda A. Camras, & Kevin Allison, Nonverbal Behav (1985), Children's understanding of emotional facial expressions and verbal labels, 9: 84. https://doi.org/10.1007/BF00987140

发展历史。从 1997 年 Picard 教授首次提出“情感计算”（affective computing）[3] 这个概念至今，计算机领域对于情感识别的研究由简单到复杂，逐渐发展为一个专有的研究领域，并且仍在不断进化。具体到智能化教育，我们致力于通过情感计算技术去理解教育场景下用户的情绪状态变化，以辅助教师进行更优的教育决策，或者智能教育系统根据情感识别的结果进行渲染和反馈，进而实现个性化教育，最终提升用户的学习体验和学习效率。

人类对情感的表达是复杂而微妙的，人对情感的识别和解读也是多通道协同完成的，包括表情、姿态、语言和声调等。对于计算机识别情感的研究，我们借鉴但并不局限于人类采用的模式和线索。下面将首先从四个模块（人脸、语音、动作和生理信号）介绍计算机情感识别的研究进展。

情感识别技术

计算机对从传感器采集来的信号进行分析和处理，从而推断出用户的情感状态，这个过程叫做情感识别。从生理心理学的观点来看，情绪是有机体的一种复合状态，既涉及体验又涉及生理反应，还包含行为。在智能化教育场景中，由于用户的人脸图像序列、人体图像序列及语音数据相对容易收集，故可通过人脸、语音、动作、生理信号等模态进行精准情感识别。

人脸表情识别

20 世纪 70 年代，美国心理学家 Ekman 和 Friesen 对现代人脸表情识别 [4] 做出了开创性的工作。Ekman 首先提出了人类的六种基本表情：高兴、生气、吃惊、恐惧、厌恶和悲伤；其次建立了面部动作编码系统（Facial Action Coding System，FACS[5]），使研究者能够按照系统划分的一系列人脸动作单元（Action Unit，AU）来精确地描述人脸面部动作。FACS 系统详细描述了人脸运动组合和几类基本表情之间的对应关系。这些工作是计算机识别人脸表情的重要理论基础。

在人脸情感中，微表情扮演着重要的角色。微表情是人们在试图隐藏或抑制真实情感时无意识流露出的面部表情，持续时间短暂，通常在 1/25 秒到 1/5 秒之间，不易被人察觉。最近有研究 [6] 系统性地阐述并提出了可靠的微表情检测与识别方法，使微表情在教育行业中的应用成为了可能。将微表情应用于课堂教学中，可以帮助教师从面部表情体察学生真实的内心状态。

人脸情感识别一般可分为三个部分（见图 1）：人脸检测及预处理、特征提取和情感识别。

[3] Rosalind Picard. Affective computing. MIT press; 2000

[4] Paul Ekman & Wallace V. Friesen, (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2), 124-129.

[5] Paul Ekman & Erika L Rosenberg (Editors). What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). Oxford University Press, USA; 1997.

[6] Xiaobai Li, Xiaopeng Hong, Antti Moilanen, Xiaohua Huang, Tomas Pfister, Guoying Zhao, Matti Pietikäinen. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Transactions on Affective Computing. 2018 Oct 1;9(4) :563-77.

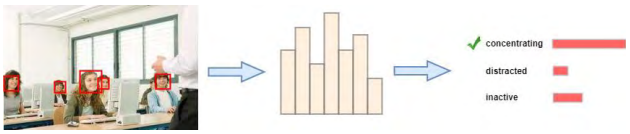


图 1：人脸情感识别流程图

摄像头采集到的图像包含复杂的背景信息，所以人脸检测会直接获取到图像中的人脸位置并进行裁剪，从而得到精细的人脸区域。通过人脸检测，无关的背景信息被去除，只保留有效的人脸信息。然而得到裁剪后的人脸图像往往会存在角度和尺寸不一致性等问题，影响后续特征提取的有效性。因此，在特征提取前，需要对人脸图像进行预处理操作来保证特征提取的一致性和鲁棒性。在课堂教育场景下，由于坐前后排、左右排的学生位置的影响，检测到的人脸区域会存在不同程度的角度偏差及分辨率不同。因此，人脸预处理的目的是人脸图像标准化，降低角度、尺寸、光照等因素对特征提取的影响。一般用到的技术包括人脸关键点提取和矫正、人脸超分辨率图像重构等。

特征提取的过程就是提取不同情感状态下的有辨识度的信息，尽可能使得相同情感状态下的信息更加紧凑，而不同情感状态下的信息更加疏远。首先，由于不同的情感状态表现在人脸外观（appearance）上具有显著差异性，多种基于外观的特征被提出，例如基于局部描述子的 Local Binary Pattern（LBP）特征 [6] 和基于卷积神经网络（CNN）的全局深度特征。前者运算量低，实时性高；后者识别准确率高。根据教育业务场景的需要，可灵活选用不同的特征。其次，不同情感状态下对应的人脸动作单元（AU）之间的关系也有显著差异，即其具有不依赖于个体的普适的几何特性（Person independent properties），因此基于几何比率（Ratio based Geometric, RbG）的特征信息也可以用来作为情感识别的依据。另外，基于差分生长模型来描述不同表情的个体人脸特征的运动，并用视频图像对齐方法来为每种表情建立显著性时间轴变化图，可以非常好地描述不同表情的外观和运动特征 [7]。

在教育场景下，可以根据具体的需要将情感识别看成是若干个主要认知情绪状态的多分类问题，例如集中的（concentrating）、分心的（distracted）、不作为的（inactive）；或者看成是传统的表情分类问题，例如高兴、吃惊、恐惧和厌恶等。这里值得指出的是，分类问题的类别选择取决于研究需要并且严格依赖于数据的有效标定，以供机器学习和训练（training）。提取人脸特征后，经典分类器，例如支持向量机（Support vector machine, SVM）和多层感知器（Multi-layer Perceptron, MLP），可以被用来进行情感识别。识别结果将会包含所有认知情绪或表情类别的预测概率，置信度最高的类别则作为最终的情感识别结果。此外，群组情感分析 [8] 也可以提供学生团队学习中的情感交互信息。

语音情感识别

语音情感识别是指由计算机自动识别输入语音的情感状态。一般来说，不同语言声

[7] Yimo Guo, Guoying Zhao and Matti Pietikäinen. Dynamic Facial Expression Recognition with Atlas Construction and Sparse Representation. IEEE Transactions on Image Processing, 25(5): 1977-1992, 2016.

[8] Xiaohua Huang, Abhinav Dhall, Roland Goecke and Matti Pietikäinen, Guoying Zhao. Multi-model Framework for Analyzing the Affect of a Group of People.IEEE Transactions on Multimedia, 2018.

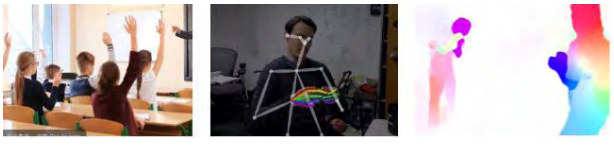
调表情的语言信号在其时间构造、振幅构造、基频构造和共振峰构造等特征方面也有着不同的构造特点和分布规律。因此，只要把各种具体情绪下的语言声调在时间构造、振幅构造、基频构造和共振峰构造等特征方面的构造特点和分布规律进行测算和分析，并以此为基础创建模板，就可以识别出测试语音中所表达的情感内容 [9]。

语音特征一般分为声学（Acoustic）特征和声韵（Prosodic）特征。前者包括波、信号、语调等，而后者包含词之间的停顿、韵律、声音大小等，后者更依赖于说话者。对于声学特征提取，Mel-Frequency Cepstral Coefficients（MFCC）[10] 特征最为经典。Mel 频率是基于人耳听觉特性提出来的，它与 Hz 频率成非线性对应关系。MFCC 则是利用它们之间的这种关系计算得到的 Hz 频谱特征。对连续语音信号进行 MFCC 参数提取的基本流程包括：预加重、分帧、加窗、FFT、三角带通滤波器、对数能量、DCT、谱加权、倒谱均值减和动态差分参数。也有研究 [11] 结合语音信号和视觉信息（Visual speech data）进行更精确的情感识别。近年来，还陆续有使用循环神经网络（RNN）和 CNN 从声谱图中提取特征用于语音情感识别的研究。和“人脸情感识别”的流程类似，当语音特征提取完毕后，会将它们输入到某个分类器进行情感识别。

动作情感识别

在日常交流中，人们的肢体动作往往包涵着不同的情感信息。动作情感识别就是通过特定的人体动作来推断其当前的情感状态。在教育场景下，主要分为宏动作和微动作两种动作类型。宏动作指的是通过识别举手、摇头、点头、趴桌子等来推断对应的积极或消极等情绪状态；微动作是指通过识别微小幅度、短暂时间片段内的动作来推理对应的情感。例如，当学生站起来回答问题时，通过其肢体动作是否微颤、不协调来推断其是否处于紧张、不自信的情感状态。

对于动作情感识别，首先需要对视频流中的关键帧进行行人检测，得到人体 RGB 图像，并生成对应的骨骼（Skeleton）图和光流（Optical flow）图（见图 2）。RGB 图像里包含不同物体（例如手部、头部、桌子等）的丰富的空间上下文外观信息；骨骼图主要包含人体关节点的空间结构性信息；而光流图包含关于人体动作的时间上下文信息。基于人体的三种模态图像集合（RGB 图、骨骼图和光流图）来提取外观及动作特征，进而对包含不同情感的动作类型进行表征。关于特征提取和建模过程，经典的稠密光流、密集轨迹、隐马尔科夫模型以及近年来火热的 RNN 和 CNN 关系网络，都可根据不同的场景被选择使用。当动作特征提取完毕，将可以使用类似“人脸情感识别”的分类器进行情感识别。



(a)RGB 图示例　(b) 骨骼图示例　(c) 光流图示例

图 2：人体模态图像示例

[9] 赵腊生，张强，魏小鹏 . 语音情感识别研究进展 [J]. 计算机应用研究，2009, 26(2):34-38.

[10] Logan B. Mel Frequency Cepstral Coefficients for Music Modeling. InISMIR 2000 Oct 23 (Vol. 270, pp. 1-11).

[11] Ziheng Zhou, Xiaopeng Hong, Guoying Zhao, Matti Pietikäinen. A compact representation of visual speech data using latent variables. IEEE transactions on pattern analysis and machine intelligence. 2013 Sep 17;36 (1):181-187.

生理信号情感识别

人类的情感变化会导致一系列生理信号的变化，其中包括心电信号 ECG、脑电信号 EEG、 皮电信号 EMG 等。虽然人类无法直接感知或读取他人的生理信号，但是通过特殊传感器测量生理信号来分析人的生理和心理状态，一直是心理学、认知神经科学和生物医学领域中主要的研究课题之一。相比于易受外界因素和主观动机干扰的外显的行为变化（例如面部表情、肢体动作等），内在的生理信号更加稳定并难以主观操控，对于测量和理解情感状态是一个十分有力的补充。生理信号分析的一个主要的限制是，由于需要特殊的测量设备，难以在实际应用中推广使用。在以往的计算机视觉和机器学习领域中，生理信号分析原本是领域外的课题。但是近几年的研究 [12] 发现，可以在环境光条件下用普通摄像机从人脸皮肤颜色的细微变化中分析、提取出心率、呼吸等生理信号。远程生理信号分析成为计算机视觉和人工智能领域的热门关注课题。

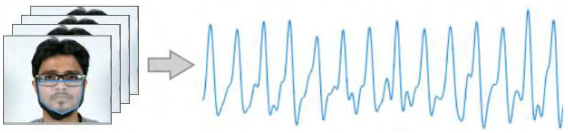


图 3：远程心跳信号提取

得益于远程光学体积描记术（remote photoplethysmography, rPPG）技术 [13] 的成熟，从普通商业摄像头捕捉的人脸视频中实时重构其心跳信号成为了可能（见图 3）。由于心脏跳动及血液流动会导致皮肤颜色有细微的变化，rPPG 技术的原理是利用摄像机获取的皮肤反射光来测量皮肤的细微亮度变化。在教育软件及助教机器人场景下，更多的是一对一浸入式学习，此时摄像头捕获的视频中人脸感兴趣区域（Region of interest, ROI）质量较优，适于提取 rPPG 信号。首先，跟踪视频帧中的人脸 ROI，并对 ROI 中所有像素亮度值进行均值融合。接着，对提取的均值信号进行盲源分解或滤波操作，去除环境噪声的影响，从而得到稳定的远程信号。若要实现通过远程生理信号测量来进行情感识别，首先需要获得鲁棒的特征。通过对 rPPG 信号里每个有效心跳的峰值检测，可以得到信号中每个峰值的时间位置，从而计算出相邻心跳间隔（Inter-beat-intervals, IBIs）曲线，再进行时频分析，得到心率变异性（heart rate variability, HRV）的声谱图作为特征。最后，可构造类似“人脸情感识别”的分类器进行情感识别，例如中立情绪（Neutral）和压力情绪（Stressful）。传统生理信号测量（如 ECG 等）需要接触式的复杂仪器，不可能实现长期测量。由于 rPPG 技术仅需普通摄像头作为信号采集设备，可以方便而持续地获得信号，为长期追踪记录情感状态、创建“用户画像”提供了可能。

rPPG 技术目前仍在起步阶段，主要的关注点仍是基本的平均心率测量。未来，随着技术的提升，预期可以通过精确测量出的 rPPG 心理信号来分析获得更多其他的生理信号，包括呼吸、血氧饱和度、血压等。这些生理信号不仅可以应用于医疗健康领域，也使通过无接触式远程测量生理信号分析情感变化成为可能。在未来智能化教育场景中，

[12] Wim Verkruysse, Lars O. Svaasand, J Stuart Nelson. Remote plethysmographic imaging using ambient light. Optics express. 2008 Dec 22;16(26): 21434-45.

[13] Xiaobai Li, Jie Chen, Guoying Zhao, Matti Pietikäinen. Remote heart rate measurement from face videos under realistic situations. In Proceedings of the IEEE conference on computer vision and pattern recognition 2014 (pp. 4264-4271).

远程生理信号测量可以和上述面部表情、身体动作，以及语音信号结合起来，实现多模态的综合情感分析，以辅助智能化教育系统更准确地分析学习中的情感动机因素。

基于情感识别的辅助教育

根据不同的教育应用场景，可以得到不同模态的特征及对应的情感识别结果。可以使用多模态融合（multimodal fusion）技术输出更为精确的情感识别结果。在得到不同用户的情感识别结果后，部分中间特征、生理信号及最后的识别结果将会绑定用户ID，建设其短期和长期的情感画像。为每个用户建立长短期情感画像是更好地实现智能化教育的关键点之一（见图 4）。对于不同的教育应用场景，长短期画像都可以起到有效的辅助作用，进而改善学习感受，提高学习效率。对于传统的一对多教师授课场景，教师通过了解用户长短期情感状态可以更有针对性地进行情感互动；对于一对一的助教场景，助教机器人及学习软件可以针对当前使用的画像记录选择相应激励和情感互动模式，以鼓励促进用户更高效地学习。

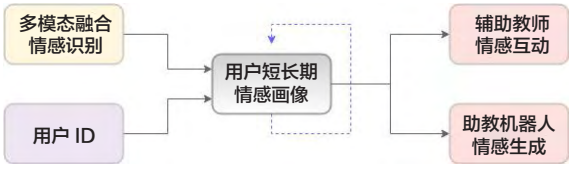


图 4：基于情感识别的辅助教育

多模态融合及辅助教育

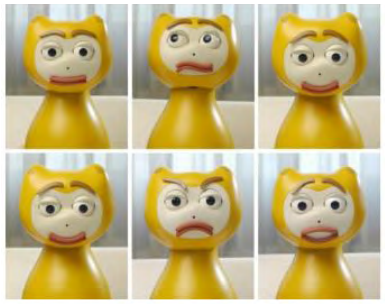
多模态融合的目的是对多个数据源信息进行聚合，从而产生比任何单一模态信息更加一致、准确的信息。一般来说，信息融合会在传感器（sensor）级别、特征（feature）级别和决策（decision）级别进行。传感器级别的融合是硬件设计问题，在此不过多讨论。对于最常使用的特征级别融合，直接进行特征拼接（concatenation），操作简洁，但特征冗余度较高。特征的冗余可通过子空间学习（subspace learning），包括经典的主成分分析和典型相关分析进行优化。另一方面，决策层融合是指每个模态的建模及分类都是独立的，最终通过一些统计学的方法将各模态的分类决策整合成最终结果，常见的方法包括多数票决（majority votes）、总数（sums）、积（products）和加权平均数（weighted average）等。

在对用户进行多模态融合的情感识别的同时，需要对用户的 ID 进行识别与跟踪。这里用到的技术有人脸识别或者行人重识别。每个用户的情感画像将会从初次建立后被不断地补充、更新，从而能够越来越准确地获得每个用户的学习需求以及情感特征。这对提供智能化教育的机构和体系维护用户，进而改善服务，是至关重要的。举一个根据用户情感画像改善教育服务体验的例子：若发现某 ID 用户在一段时间内多次呈现分心（distracted）和紧张（stressful）的状态，系统会辅助教师选择难度更低、更有趣的知识点进行传授，并提醒其需要用更饱满情绪及肢体语言来激发用户兴趣。

助教机器人的情感行为生成

除了上述的情感识别技术，情感行为的模拟生成（synthesis）也是智能化教育需要的

重要技术。当教学场景为教学软件或助教机器人时，就需要根据情感识别及用户画像的结果进行情感行为生成（见图 5）。只有机器也具有情感表达能力，才能有效地安抚、鼓励或者鞭策用户；用户才能有更好的学习体验，提升学习效率。对情感生成建模的同时，需要考虑到不同用户各自的性格、社会关系和文化等。



机器人表情生成

虚拟助手表情生成

图 5：情感生成

机器的情感行为表达主要通过以下三种方式：脸部表情、姿态动作和语音。三种表达方式在理论上应当是高度同步的，因此最后生成出来的脸部表情、姿态动作和语音需在时间维度上尽量保持一致性。1）对于脸部表情来说，一般分为离散（Discrete）表征、维度（Dimensional）表征和估价（Appraisal）表征。离散表征定义生成的表情为原型表情的离散集合，例如著名的 EmotionDisc[14]。维度表征将表情生成看成是在 2D 情感空间中对距离最近的两个情感进行插值。估价表征利用时间及 intra-SEC（Sequential Evaluation Checks）上的相关性来自适应地构造 AU。2）对于姿态动作生成，由于情感一般通过基本的身体移动来进行表达，因此首先建模和生成基本动作，然后再对这些动作进行组合上色。3）对于语音生成，主要有以下几种生成方式：共振峰（Formant）生成、双音拼接（Diphone Concatenation）以及非均一单元选择（Non-uniform Unit Selection）。共振峰生成是基于气动声学模型进行系统建模，因此其生成的语音很机械化不自然。双音拼接和非均一单元选择都需要先构造双音数据集，但由于后者更海量的数据和更优的求解策略，因此其生成的语音最为真实、自然。

三、结语

当前的人工智能技术已经应用在教育考勤、机器人助教、知识梳理等教育场景中，一定程度上实现了智能化教育。但是由于缺乏对用户情感的精准理解以及与用户真实情感的沟通，当前的“AI+ 教育”系统还比较粗糙，无法从个人水平上理解用户需求并做出回应，从而限制了用户的学习体验和学习效率。随着情感计算技术的发展，智慧校园、助教机器人等“AI+ 教育”的落地产品致力于更精准地理解情感、生成情感并与用户进行情感交互。相信情感计算技术可以使未来的智能化教育更加以人为本，为人们提供更加准确、高效的教育服务。

查看内容精选

[14] Zsofia Ruttkay, Han Noot and Hagen Paul, Emotion Disc and Emotion Squares: Tools to Explore the Facial Expression Space, Blackwell Publishers, Inc and the Eurographics Association, 1467-8659, 2003,