多模态维度情感预测综述

李 霞 [1, 2] 卢官明 [1] 闫静杰 [1] 张正言 [1, 3]

摘 要 维度情感模型通过几个取值连续的维度 (如唤醒维、效价维、支配维等) 将情感刻画为一个多维信号. 与传统的离散情感模型相比, 具有表示情感的范围广、能描述情感的演变过程等优点, 近年来受到越来越多情感识别研究者的关注. 多模态维度情感预测是一项复杂的工程, 预测性能受所使用的模态、每个模态的特征提取、信息融合技术、标注人员的标注误差等多方面影响. 为了提高多模态维度情感预测的性能, 研究者在各个方面都做出了不懈努力. 本文综述了维度情感的概念、标注, 维度情感预测的性能评价指标以及多模态维度情感预测的研究现状, 对比和分析了各种因素对多模态维度情感预测性能的影响, 并总结出多模态维度情感预测面临的挑战及发展趋势.

关键词 情感识别, 情感预测, 维度情感模型, 离散情感模型, 信息融合, 特征提取

引用格式 李霞, 卢官明, 闫静杰, 张正言. 多模态维度情感预测综述. 自动化学报, 2018, **44**(12): 2142−2159

DOI 10.16383/j.aas.2018.c170644

A Survey of Dimensional Emotion Prediction by Multimodal Cues

LI Xia[1, 2] LU Guan-Ming[1] YAN Jing-Jie[1] ZHANG Zheng-Yan[1, 3]

Abstract The dimensional emotion model characterizes emotion as a signal in a multi-dimensional space spanned by several continuously valued dimensions (such as arousal, valence, and dominance). Compared with the discrete emotion model, it has the advantages that it can distinguish subtle difference of emotion, can represent evolution of emotion, etc. So the dimensional emotion model has been paid more and more attention in recent years. Dimensional emotion prediction from multimodal cues is a complex task, the prediction performance is influenced by such as modalities used, features extracted from each modality, information fusion technique, annotation errors. In order to improve multimodal dimensional emotion prediction performance, researchers have made persistent efforts in all aspects. In the paper, concept and annotation of dimensional emotion, performance evaluation criteria of dimensional emotion prediction, and research status of multimodal dimensional emotion prediction are reviewed; influences of various factors on emotion prediction performance are analyzed; challenge and development trend of multimodal dimensional emotion prediction are summarized.

Key words Emotion recognition, emotion prediction, dimensional emotion model, discrete emotion model, information fusion, feature extraction

Citation Li Xia, Lu Guan-Ming, Yan Jing-Jie, Zhang Zheng-Yan. A survey of dimensional emotion prediction by multimodal cues. *Acta Automatica Sinica*, 2018, **44**(12): 2142−2159

情感是人们日常生活中常见的一种心理现象. 对情感的准确识别是利用情感进行交流的前提, 在日常人际交往中有着重要的作用. 对于智能机器, 只有能够对人的情感状态进行快速准确的判断, 才有可能进一步理解和响应人类情感, 从而实现与用户进行自然、友好、和谐地交互[1]. 例如在智能汽车系统中, 对驾驶员的情感状态进行实时监测, 并根据监测结果给予必要的响应便能够有效避免事故的发生; 在智能电话服务系统中, 对来电者的情感状态进行自动判断, 根据判断结果给予合适的响应或将控制线转接给人工处理, 便能有效地提高服务效率和质量.

人的情感是通过面部表情、身体姿态、声音以及生理信号等多种模态表现出来的. 情感判断可以基于这些模态中的一个或多个来进行, 但是单个模态的信息存在信息不全面、容易受噪声干扰等缺陷, 目前越来越多的研究者开始综合运用多个模态的信息进行情感判断. 多个模态的信息能够互相印证、互

收稿日期 2017-11-15 录用日期 2018-03-07
Manuscript received November 15, 2017; accepted March 7, 2018
国家自然科学基金 (61501249, 61071167), 江苏省重点研发计划项目 (BE2016775), 江苏省自然科学基金 (BK20150855), 江苏省研究生创新项目 (KYLX15_0827, KYLX16_0660) 资助
Supported by National Natural Science Foundation of China (61501249, 61071167), Key Research and Development Program of Jiangsu Province (BE2016775), Natural Science Foundation of Jiangsu Province (BK20150855), and Jiangsu Innovation Program for Graduate Education (KYLX15_0827, KYLX16_0660)
本文责任编委 黄庆明
Recommended by Associate Editor HUANG Qing-Ming
1. 南京邮电大学通信与信息工程学院 南京 210003 2. 安徽工业大学数理学院 马鞍山 243000 3. 江苏科技大学电子信息学院 镇江 212003
1. College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003 2. School of Mathematics and Physics, Anhui University of Technology, Maanshan 243000 3. School of Electronics and Information, Jiangsu University of Science and Technology, Zhenjiang 212003

相补充, 从而可以为情感判断提供更加全面准确的信息, 提高情感判断的性能. D′Mello 等[2] 对 2009～2013 年出现的多模态情感识别系统进行元数据分析发现, 相比于单模态情感识别系统, 平均性能提高了 9.83 %, 这充分肯定了多模态信息融合对提高情感识别性能的有效性.

要对人的情感状态进行判断, 首要的任务是建立情感状态的表示模型. 在情感识别领域, 常用的情感表示模型主要有离散情感模型和维度情感模型. 离散情感模型使用快乐、悲伤、愤怒等形容词标签表示情感, 虽然此种表示方式非常简单、直观, 但无法区分情感的细微差别, 也无法描述情感的演变过程. 维度情感模型用几个取值连续的维度将情感刻画为一个多维信号[3]. 由于每个维度的取值可以连续变化, 因此可以对情感的细微差别进行区分, 同时可以通过对情感状态的实时标注跟踪情感状态的演变过程. 同时, 心理学的研究表明, 一些情感维度的取值情况与人的记忆、注意等认知行为具有密切联系[1], 这使得机器更容易根据维度情感预测结果来理解和响应用户的情感行为.

随着手机、平板等各种便携录像、录音设备, 以及 iwatch 等智能可穿戴设备的出现, 人们随时随地获取视频、音频及生理信号成为可能, 这为多模态维度情感预测提供了数据基础, 拓展了多模态维度情感预测的应用领域. 在多模态交互式对话系统中, 系统中的虚拟人可以根据用户的语音、面部表情和姿态预测用户的维度情感, 并根据预测结果选择合适的词语与用户进行对话, 将用户的情感状态向某个特定的情感状态进行引导.

多模态维度情感预测是综合运用情感的多个表现模态对各个情感维度的取值进行预测, 是一个复杂工程, 包括建立多模态维度情感数据库、从各个模态中提取特征、选择与设计预测模型、信息融合等环节, 每个环节的处理对最后的预测性能都具有重大影响. 本文综述了多模态维度情感预测各个环节的研究现状, 对比和分析了不同方法对预测性能的影响, 并总结出多模态维度情感预测面临的挑战及发展趋势.

1 维度情感模型

离散情感模型和维度情感模型是情感识别领域广泛使用的两种情感表示模型. 离散情感模型使用形容词标签将情感表示为几种相对独立的情感类别 (例如 Ekman 提出的快乐、悲伤、愤怒、恐惧、厌恶和惊讶六种基本情感[3]). 离散情感模型因其简单直观的优点, 在情感识别领域得到了极其广泛的应用. 但是存在许多缺点: 1) 情感的类别总是运用某个词语表示, 导致运用此模型能够表示的情感范围有限, 同时导致情感的编码与文化和语言具有密切的联系[4], 从而限制了情感编码的普适性; 2) 很多情感类别之间存在高度的相关性[5], 但在此模型下很难对这种相关性进行度量和处理; 3) 情感的产生、发展和消失是一个过程, 而此模型无法描述情感的发展进程.

为了克服离散情感模型的缺点, 研究者建立了维度情感模型. 维度情感模型认为情感是一种高度相关的连续体, 运用几个取值连续的基本维度将情感状态描述为多维空间中的某一个坐标, 每个维度是对情感的某一方面的度量[5]. 对于情感具有哪些维度, 心理学家并没有统一的认识, 其中认同度最高的一种模型为 “愉悦 (Pleasure) – 唤醒 (Arousal) – 支配 (Dominance)” 模型或 PAD 模型, 此模型认为情感具有愉悦维、唤醒维和支配维三个维度. 愉悦维也称为效价 (Valence) 维, 是对人的愉悦程度的度量, 从一个极端 (苦恼) 到另一个极端 (狂喜); 唤醒维也称为激活 (Activation) 维, 是对生理活动和心理警觉水平的度量, 如睡眠、厌倦等为低唤醒, 清醒、紧张等为高唤醒; 支配维也称为注意 (Attention) 维或能量 (Power) 维, 是指影响周围环境及他人或反过来受其影响的一种感受, 高的支配度是一种有力、主宰感, 而低的支配度是一种退缩、软弱感[5−6]. Russell 在对 PAD 模型进行深入研究时发现, 支配维更多地与认知活动有关, 愉悦和唤醒两个维度就可以表示绝大部分不同的情感, 他采用环状结构模型表示复杂的情感[5]. 在环状结构模型中, 每个维度的取值极限构成一个圆, 圆的中心表示中性的情感[7], 愉悦和唤醒是两个相互正交的维度, 情感均匀地分布在圆环的内部[5], 此模型称为愉悦 – 唤醒模型 (也称为效价 – 唤醒模型或 VA 模型), 运用此模型可以表示多数基本情感, 如图 1 所示[8]. 由于愉悦 – 唤醒模型的简单和实用性, 很多维度情感预测的研究都是在这两个维度上进行的. 理论上讲 PAD 模型能够表示无穷多种情感, 但它仍然不能表示人类所能体验的所有情感, 例如 “惊讶” 就处在了此情感空间的外部[2]. 为了更完整地描述情感, 一些研究者将期望 (Expectation/anticipation) 维作为第四个维度, 强度 (Intensity) 维作为第五个维度[9]. 期望维是对个体情感出现的突然性的度量, 即个体缺乏预料和准备程度的度量; 强度指的是个体偏离冷静的程度. Fontaine 等[10] 的研究表明, 第四个维度的加入能够将 “惊讶” 与其他的情感类型区分开来, 基本能够区分日常生活中的所有情感. 因此, 在维度情感预测中, 也有不少是基于前四个维度进行的.

近年来, 维度情感预测受到了越来越多的关注. 其主要优势在于: 1) 维度情感模型相比于离散情感模型具有更强的表示能力, 尤其是在处理自然的数

据时优势更加明显, 此时情感状态的范围非常广泛, 很难用有限的几种情感类型描述[4]; 2) 运用维度情感模型可以对情感的发展变化过程进行跟踪[4]; 3) 运用维度情感模型可以对情感的相似性和差异性进行度量[9]; 4) 心理学研究表明, 人类的决策、推理、记忆、注意等认知都与 PAD 模型中的三个维度存在密切关系, 例如, Lang 等研究表明愉悦维度决定了欲求动机系统和防御动机系统哪个被情感刺激激活, 而唤醒维度决定了每个动机系统被激活的程度[11]. 由此可见, 在人机互动中, 运用维度情感模型比运用离散情感模型更利于机器充分理解人的情感并做出合适的反应.

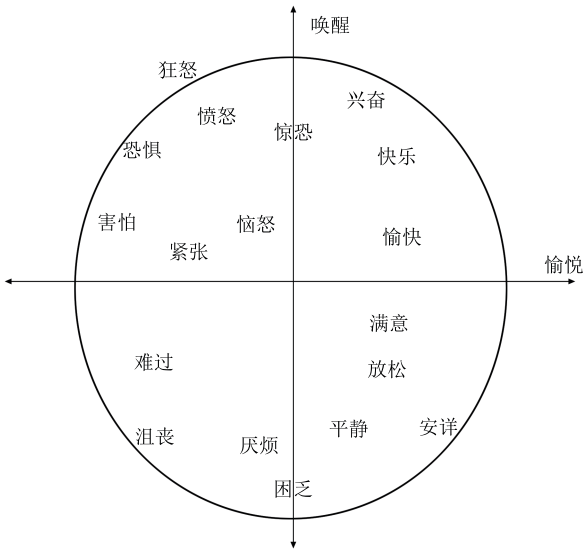


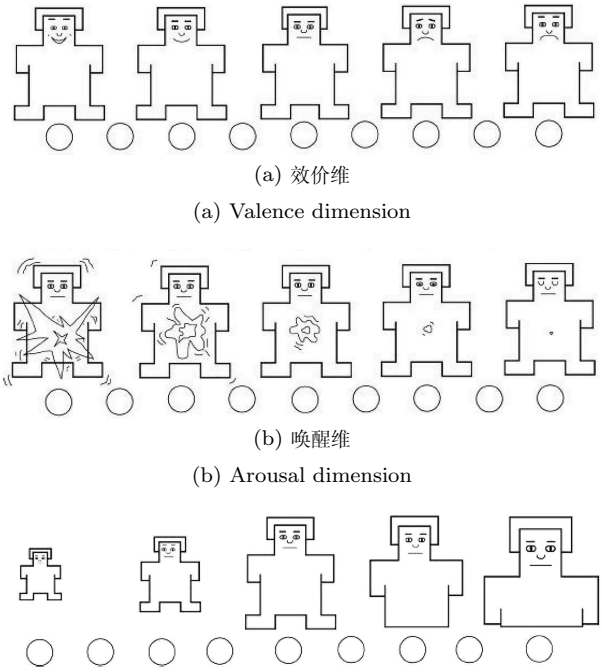
图 1 愉悦–唤醒模型

Fig. 1 Pleasure-arousal model

2 维度情感标注

维度情感模型虽然具有很多优点, 但是维度情感预测直到最近几年才得到人们的更多关注, 主要原因是这种表示方式比较抽象, 标注比较困难.

维度情感标注工作是基于情感量化理论完成的, 目前没有一个统一的方法. SAM (Self-assessment manikin) 系统是一种被多数研究者认可的维度情感量化方法, 它基于 PAD 模型建立[12], 使用卡通小人的形象表示 PAD 模型中三个维度的取值. 图 2 是效价维、唤醒维和支配维的取值分布[12], 以卡通小人眉毛和嘴巴的变化表示效价维的取值; 以心脏位置出现的震动程度以及眼睛的有神程度表示唤醒维的取值; 以图片的大小表示受控制的程度. 在某个维度标注的过程中, 只需从对应的卡通小人中选出一个最符合当前情感状态的即可. 使用的小人数目由对此维度进行量化的数目决定, 一般为 5 个或 9 个. 每个小人对应的具体数值没有一个严格规定, 使用 9 个小人时, 对应的 9 个数字可以是 1 ∼ 9 的整数, 可以是 −4 ∼ 4 的整数, 也可以是 $[-1, 1]$ 的 9 个等间隔的值[13]. 相比于其他情感量化方法, SAM 系统具有简单、快速、直观的优点, 并且避免了不同人对同一词语的不同理解造成的差异, 从而获得的标注结果方差较小、不同标注者间的一致性较高[14], 因此 SAM 系统经常被用于维度情感的标注任务中. 在每个卡通小人的下方标注数字并与小人一起呈现于屏幕上, 允许标注者点击两个数字之间的任意位置, 即可以实现对目标维度的连续赋值[13].



(a) 效价维

(a) Valence dimension

(b) 唤醒维

(b) Arousal dimension

(c) 支配维

(c) Dominance dimension

图 2 SAM 系统

Fig. 2 SAM system

情感是一个不断变化的过程, 为了对每个情感维度的取值进行实时跟踪, 研究者开发了很多标记工具, FEELtrace[7] 和 ANNEMO[15] 是两个常用的标记工具. FEELtrace 是基于效价–唤醒环状模型建立的, 如图 3 所示[7], 将以效价维和唤醒维为主轴的圆呈现于电脑屏幕上, 标注者只需根据自己感知的情感用鼠标拖动圆形光标到合适的位置即可同时对效价维和唤醒维赋值[7]. ANNEMO 是一种基于网页的维度情感标记工具, 如图 4 所示[15], 它将视频和标记光标同时显示于一个窗口, 用户在观看视频的同时, 对视频中对象的某个情感维度进行时间连续的标记[15]. 与 FEELtrace 相比, ANNEMO 使用更加方便, 而且一次只对一个维度进行标记, 得到的结果更加精确.

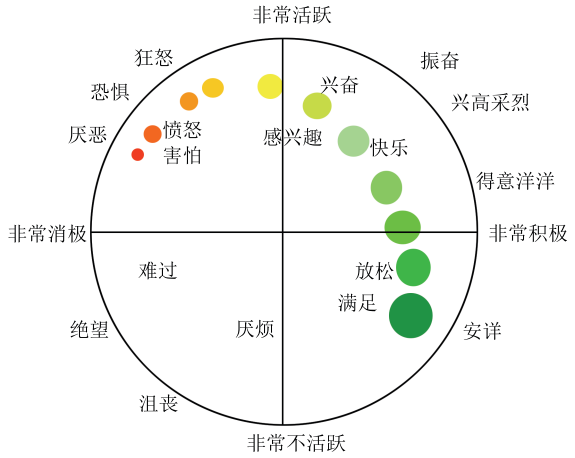


图 3　FEELtrace 标注示例

Fig. 3　Example of FEELtrace annotation

图 4　ANNEMO 标注示例

Fig. 4　Example of ANNEMO annotation

3　维度情感预测的性能评估指标

维度情感预测问题主要可以分为两种类型, 一是根据一个或多个维度的取值将维度情感预测问题退化为一个分类问题[9], 此分类问题既可以是按照某个维度的取值分成正与负 (或积极与消极) 两种类型的两分类问题[16], 又可以是按照某个维度的取值分为低、中、高三种类型的三分类问题[17], 还可以是在效价–唤醒空间中用四个象限代表四个类别的四分类问题[18] 等; 二是对每个维度的连续取值进行预测, 此时维度情感预测问题是一个回归问题[19].

当维度情感预测问题退化为分类问题时, 称为维度情感分类, 此时预测性能的评价指标与离散情感识别使用的评价指标相同, 主要有整体分类准确率 (Accuracy)、召回率 (Recall)、精确率 (Precision)、F1-score 等. 设共有 A, B 两种类别, $n_{\rm TP}$ 是 A 类样本正确分类的样本数, $n_{\rm FN}$ 是 A 类样本错误分类的样本数, $n_{\rm FP}$ 是 B 类样本错误分类的样本数, $n_{\rm TN}$ 是 B 类样本正确分类的样本数. 则整体分类准确率定义为

$$P_{\rm acc}=\frac{n_{\rm TN}+n_{\rm TP}}{n_{\rm TN}+n_{\rm FN}+n_{\rm TP}+n_{\rm FP}} \tag{1}$$

A 类样本的分类准确率或召回率定义为[20]

$$P_{\rm re}=\frac{n_{\rm TP}}{n_{\rm TP}+n_{\rm FN}} \tag{2}$$

A 类样本的分类精确率定义为[20]

$$P_{\rm pre}=\frac{n_{\rm TP}}{n_{\rm TP}+n_{\rm FP}} \tag{3}$$

A 类样本的分类 F1-socre 定义为[20]

$$P_{\rm F1}=\frac{2P_{\rm pre}P_{\rm re}}{P_{\rm pre}+P_{\rm re}} \tag{4}$$

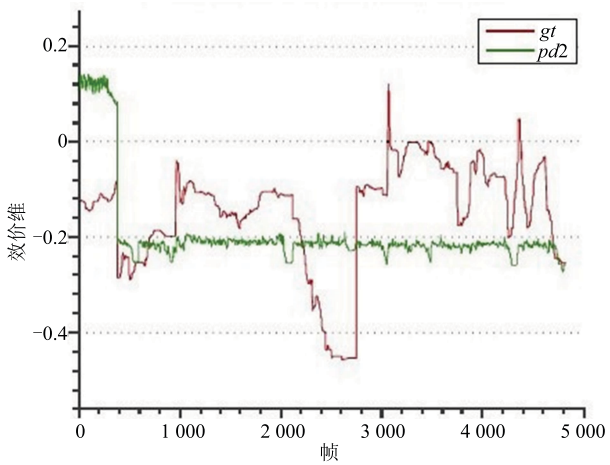
当维度情感预测为回归问题时, 称为连续维度情感预测, 此时预测性能的评价指标是一个不断探索的问题, 早期的文献一般采用均方误差 (Mean squared error, MSE) 度量估计的性能. 设 $\hat{\theta}$ 是估计标签, θ 是真实标签, n 为样本数目, $\sigma_{\hat{\theta}}^2$, σ_{θ}^2 分别是 $\hat{\theta}$ 和 θ 的方差, $\mu_{\hat{\theta}}$, μ_{θ} 分别是 $\hat{\theta}$ 和 θ 的期望, 则 MSE 定义为[21]

$$\mathrm{MSE}=\frac{1}{n}\sum_{f=1}^{n}\Big(\hat{\theta}(f)-\theta(f)\Big)^2 \tag{5}$$

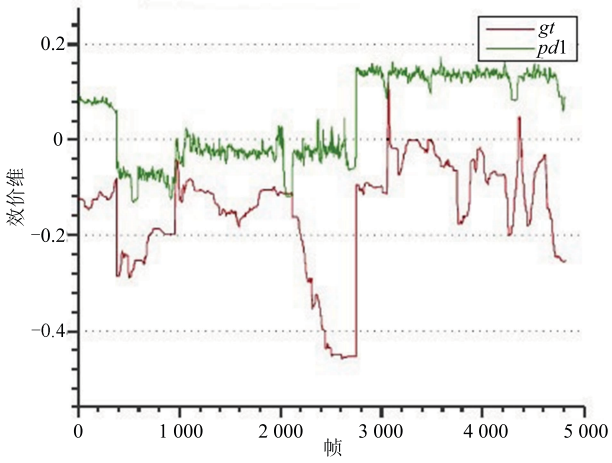
MSE 描述了预测与真值的偏差, 但 MSE 对于异常值敏感, 以及对 θ 与 $\hat{\theta}$ 的相对变化趋势无法进行描述, 因此并不能很好地描述预测与真值的吻合度. 鉴于 MSE 的缺点, Pearson 相关系数 (Pearson correlation coefficient, CC) 被用来作为连续维度情感预测的评价指标, 其定义为[21]

$$\rho=\frac{\dfrac{1}{n}\displaystyle\sum_{f=1}^{n}\Big[(\hat{\theta}(f)-\mu_{\hat{\theta}})(\theta(f)-\mu_{\theta})\Big]}{\sigma_{\hat{\theta}}\sigma_{\theta}}=\frac{\mathrm{E}\Big[(\hat{\theta}-\mu_{\hat{\theta}})(\theta-\mu_{\theta})\Big]}{\sigma_{\hat{\theta}}\sigma_{\theta}} \tag{6}$$

CC 的取值范围为 $[-1,1]$, 反映了预测与真值具有线性关系的紧密程度. 图 5 给出了两组效价维的预测与真值的对比图[21], 从图 5 可以看出, CC 能够很好地反映预测与真值的协同变化关系. 但是, 由于 CC 对预测的幅值不敏感, 无法对 θ 与 $\hat{\theta}$ 的偏差进行度量, 因此仍不能很好地描述预测与真值的吻合程度. 为了更好地描述预测与真值的吻合程度, AV + EC 2015[22] 竞赛中开始使用一致性相关系数



(a) MSE = 0.021, CC = 0.075



(b) MSE = 0.214, CC = 0.473

图 5 具有不同 MSE 和 CC 的效价维的预测与真值的对比图

Fig. 5 Comparison of the prediction and truth values of valence dimension with different MSEs and CCs

(Concordance correlation coefficient, CCC) 作为预测性能的评价指标, 其定义为

$$\rho_c = \frac{2\rho\sigma_\theta\sigma_{\hat{\theta}}}{\sigma_\theta^2 + \sigma_{\hat{\theta}}^2 + (\mu_{\hat{\theta}} - \mu_\theta)^2} \tag{7}$$

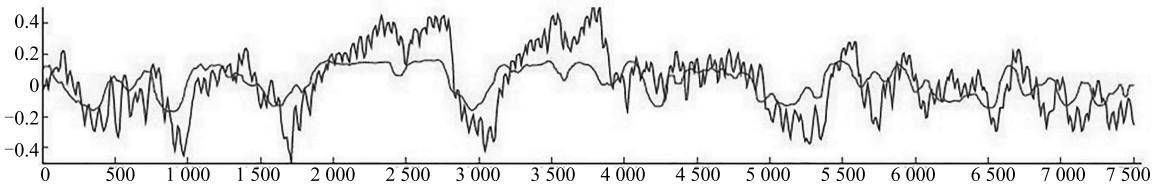
CCC 结合了 CC 与 MSE 的优点, 既反映了预测与真值的协同变化关系, 又反映了预测与真值的偏差, 因此能够更好地反映预测与真值的吻合程度, 是目前广泛使用的连续维度情感预测性能评价指标. 图 6 给出了 CC 相同, 而 CCC 不同的预测与估计的吻合程度对比[23], 显然 CCC 高的吻合程度更高.

4 多模态维度情感预测研究现状

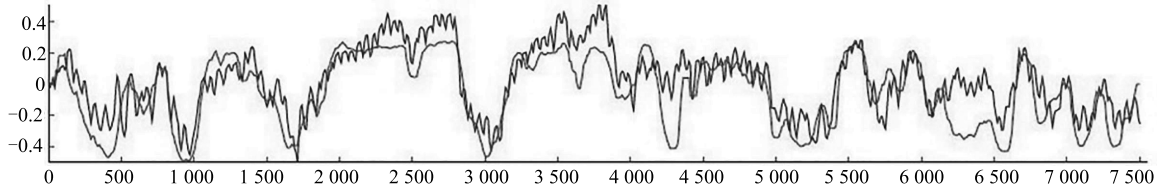
人类的情感可以通过面部表情、身体姿态、语音、生理信号等多个模态表现出来. 面部表情和身体姿态都是可视的, 有时也将它们统一看作视觉模态; 语音信息可以从听觉途径获得, 也称为听觉模态. 从这两个 (或多个) 模态中进行情感判断与我们的日常生活经验相符, 而且它们可以通过非侵入性的传感器获取, 相对来说简单方便成本低, 因此一直以来基于这几个模态中的一个或多个进行情感判断都是一个重要的课题. 近些年随着可穿戴传感器的出现, 使得生理信号的实时获取成为可能, 这促进了生理信号在情感识别研究中的运用.

面部表情是人们日常交流中理解对方情感的主要线索之一[24−25]. 面部表情的最大优点是它对六种基本情感的表现具有普遍性, 并与文化背景无关[26]. 因此早期的情感识别主要集中于运用面部表情进行六种基本情感的识别. 在维度情感模型下进行情感预测, 面部表情自然也是经常使用的重要线索之一.

Ekman 和 Friesen 的研究[27] 表明身体姿态比



(a) CC = 0.481



(b) CCC = 0.763

图 6 CC 相同的条件下唤醒维的预测与真值的对比图

Fig. 6 Comparison of the prediction and truth values of arousal dimension with the same CC

面部表情能够为维度情感预测提供更多的信息. 因此很多维度情感预测的工作是基于身体姿态进行的, Gunes 等[28] 运用头部运动的幅度和方向, 点头和摇头的频率对五个情感维度的取值进行了预测.

目前, 运动捕获系统也经常用于获取面部和身体行为信息, 它通过在面部和身体的固定位置放置一些标记, 记录人的运动行为. IEMOCAP[14] 和 CreativeIT[29] 数据库都提供了由运动捕获系统获取的面部和身体行为数据.

听觉模态是可用于情感识别的一个重要模态, 声音信号中既有明确的语言信息又有非语言的声学信息, 这些信息都可用于情感状态的推断[9, 30]. 很多实验[31−32] 都表明使用音频信息比视频信息能够获得更好的维度情感预测效果. 因此, 不管是进行单模态还是多模态维度情感预测, 听觉模态都非常重要.

心理学的研究表明, 人的情感与人的中枢神经系统和自主神经系统等都具有密切的联系. 人的不同情感活动会引起大脑的不同部位发生不同的反应[33]; 能够激起人的交感神经系统活跃, 从而引起肾上腺素和去甲肾上腺素分泌增多, 心血管系统发生一系列变化[33]; 能够引起内外腺体变化, 从而影响激素分泌量的变化[33] 等. 因此, 脑电图 (Electroencephalography, EEG)、皮肤电活动 (Electrodermal activity, EDA)、肌电图 (Electromyography, EMG)、心电图 (Electrocardiogram, EKG 或 ECG)、眼电图 (Electrooculogram, EOG)、心率、呼吸率等[4, 34] 生理信号也常用来进行维度情感预测.

上述这些模态可以单独用于情感预测, 但是单个模态存在信息不全面、容易受噪声干扰等固有缺陷, 造成依赖单个模态的情感预测系统在鲁棒性、精确性等方面都不能满足使用要求, 这在很大程度上限制了它的应用[2]. 而且, 心理学领域的研究和情感识别领域的实验结果都表明同时考虑多个模态的信息确实能够提高情感识别的效果, 因此, 多模态情感预测受到了人们极大的重视. 构建一个多模态维度情感预测系统一般包括多模态维度情感数据的收集、各个模态中的特征提取、预测模型的设计和选择、信息融合和其他影响因素的处理.

4.1 多模态维度情感数据库

在日常生活中, 各种情感状态的出现具有不平衡性, 为了获取丰富而全面的情感数据, 情感数据的收集一般是在实验室进行的. 由于表演的情感与自然的情感在很多个方面都存在差异, 目前一般不直接要求对象表演某种情感, 而是设计某种场景来诱导对象的情感, 这样获得的情感数据也被认为是自然的数据. 近年来研究者们在多个场景下收集了多模态情感数据, 并在不同的维度上进行了标注, 常用的多模态维度情感数据库有 SEMAINE, RECOLA, IEMOCAP, CreativeIT, DEAP, VAM 等.

SEMAINE (Sustained emotionally colored machine-human interaction using nonverbal expression) 数据库[35] 是为了实现计算机能够与人类进行流畅的、富有情感的对话而建立的. 目前公开的数据是在被称作 Solid SAL (Sensitive artificial listener) 的场景下获取的, 此场景模拟了人机对话的过程, 由人扮演了机器角色与用户进行对话. 机器角色根据用户的情感状态选择词语与用户进行对话, 使得对话不中断, 并将用户的情感状态向某个特定的情感状态引导. 共有 24 个用户分别与四个不同性格的机器角色进行对话, 每次对话都记录了用户和机器角色的正面视频和音频, 以及用户的侧面视频. 标注人员按照视频帧率逐帧给出了用户在对话过程中的情感状态在唤醒维、效价维、支配维、期望维和强度维五个维度上的取值.

RECOLA (Remote collaborative and affective interactions) 数据库[15] 共记录了 46 个参与者的情感数据, 这些参与者两人一组被分成 23 组, 每组通过远程视频会议讨论某个灾难场景下逃生的方案, 并达成一致意见. 数据库中包含所有参与者在讨论过程中的面部视频和音频数据, 以及其中 35 个参与者的 ECG、EDA 数据. 标注人员按照视频帧率逐帧给出了参与者前 5 分钟讨论过程中的情感状态在效价维和唤醒维的值.

IEMOCAP 数据库[14] 共记录了 10 个演员 (5 男, 5 女) 的情感数据, 这些演员一男一女组合被分成 5 组, 每组按照脚本或即兴进行对话表演. 同一对话内容由相同的演员表演两次, 每次使用运动捕获设备记录对话一方的面部表情、头部姿势和手部运动数据, 同时记录对话双方的视频和音频数据. 数据库中共有 174 段对话, 每一段对话都被分割成了语句, 每个语句呈现的情感状态在效价维、唤醒维和支配维三个维度上的值用 1～5 的整数进行了标记.

CreativeIT 数据库[29] 共记录了 16 个演员的情感数据, 这些演员两人一组被分成了 8 组进行即兴表演, 共进行了 50 次表演. 每次表演过程中, 都记录了表演双方的视频和音频数据, 以及使用 Vicon 动作捕获系统获取的演员全身动作数据. 标注人员按照视频帧率逐帧给出了每个演员表演过程中的情感状态在效价维、唤醒维和支配维三个维度的取值.

DEAP 数据库[13] 记录的是 32 个参与者在观看音乐视频时的 EEG 信号、外围生理信号, 以及其中 22 个人的正面视频. 每个参与者都观看了 40 段音乐视频, 并将自己在观看音乐视频过程中感受到的情感在唤醒维、效价维和支配维上给出了 1～9 之

间的连续自我评估.

VAM 数据库[36] 中的素材来自德国的电视脱口秀节目 Vera am Mittag. 其数据分为三部分: VAM-video 集、VAM-audio 集和 VAM-faces 集. VAM-video 集中的数据是从节目中分割出的 1 421 条语句对应的嘉宾视频. VAM-audio 集中的数据是从上述语句中选出的 1 081 条比较好的语句对应的声音信号, 并由标注人员对每条语句展现的情感状态在唤醒维、效价维和支配维三个维度上用 [−1, 1] 的 5 个等间隔值进行标注. 从 VAM-video 集中选取了大部分时间都是说话者正面图像的视频, 并从中提取出说话者的面部图像, 构成了 VAM-faces 集, 共包含 1 867 张图片. 标注人员对 VAM-faces 集的图片中对象的情感状态在唤醒维、效价维和支配维三个维度上用 [−1, 1] 的 5 个等间隔值进行标注.

表 1 总结了常用维度情感数据库的数据获取场景、参与者数目、记录的模态、标注的情感维度、标注者人数、使用的标注工具或标注方法、标签的取值范围及取值类型.

现有的数据库多数是在特定场景下诱导得到的, 在一个场景下训练的系统在另一个场景下或在真正自然的场景下的泛化能力如何, 是一个值得研究的问题, 这依赖于多个场景以及真正自然的场景下多模态维度情感数据库的建立. 构建多模态维度情感数据库与构建多模态离散情感数据库相比, 除了要面临情感状态的出现不平衡、完整的多模态信息不容易捕捉等共同要面临的困难外, 维度情感标签的标注也是一大困难. 众所周知, 情感是一个变化的过程, 对于多模态情感数据给出时间连续的维度情感标签比按段给出维度情感标签要更有使用价值. 但时间连续的维度情感标注不仅是一个耗时、耗力的乏味工作, 而且由于时间连续的维度情感标注是一个比较精细的过程, 因此标注结果与标注者自身的偏好、经验等都有着密切的关系. 为了降低标注者自身的因素对标注结果的影响, 常采取的方法[15] 有: 1) 选择多个标注者共同完成标注任务; 2) 选择与标记对象具有相同母语的标注者; 3) 在标注工作开始之前对标注者进行训练使其能够尽量客观地给出维度情感的标注, 并且能够熟练地使用维度情感标注工具; 4) 对多个标注者的标注结果进行插值、标准化等一系列后期处理, 进一步减少标注偏差.

4.2 特征提取

无论是多模态还是单模态维度情感预测, 也无论是维度情感预测还是离散情感识别, 各个模态的特征提取都是非常关键的. 特征提取后得到的特征维数往往较高, 并且可能包含过多的冗余信息, 从而影响最后的预测性能, 因此常在特征提取之后进行特征选择和降维. 表 2 总结了维度情感预测文献中使用的模态以及各个模态的特征提取、特征选择和降维方法, 同时总结了预测模型和信息融合方法.

所有可以用于情感识别的特征都可以用于多模态维度情感预测中. 如, 视觉模态的几何特征、纹理特征 (Gabor[37], LBP[38], HoG[39], Haar[40] 等)、时空几何特征和时空纹理特征 (LBP-TOP[41], LPQ-TOP[42], LGBP-TOP[43], 时空 Haar[44] 等); 音频信号中的声学特征 (梅尔倒谱系数、对数频率能量系数、线性预测系数、线性预测倒谱系数、谱质心、频谱流量、感知线性预测系数、共振峰频率及其带宽、频率微扰和振幅微扰、声门参数等[4, 8]) 及其函数; 音频信号中的语言特征 (BoW (Bag of words)[4], BoC (Bag of concepts)[4], BoNG (Bag-of-N-grams)[45], BoCNG (Bag-of-character-N-grams)[45] 等); 生理信号的时域特征 (过零率、均值等)、频域特征 (高频能量、低频能量等)、时间−频域特征 (希尔伯特−黄谱、离散小波变换等) 等[46−47], 都可用于维度情感预测中.

特征提取后得到的特征维数往往比较高, 并且可能包含的冗余信息过多, 从而影响最后的识别性能. 因此常在特征提取之后进行特征选择和降维, 常用的特征选择和降维方法 CFS (Correlation-based feature subset selection)[18], PCA (Principal component analysis)[48], SPCA (Supervised PCA)[48],

表 1 常用维度情感数据库总结

Table 1 Summary of the frequently used dimensional emotion database

数据库	场景	参与者数	模态	情感维度	标注者数	工具/方法	标签范围与类型
SEMAINE	Solid SAL	24	Vi + Au	A, V, E, D, I	2～8 人	FEELtrace	[−1, 1] 的连续值
RECOLA	远程视频会议	46	Vi + Au + Ph	A, V	6 人	ANNEMO	[−1, 1] 的连续值
IEMOCAP	双人对话表演	10	Vi + Au	A, V, D	至少 2 人	SAM 系统	1～5 的整数值
CreativeIT	双人对话表演	16	Vi + Au	A, V, D	3～4 人	FEELtrace	[−1, 1] 的连续值
DEAP	观看音乐视频	32	Vi + Ph	A, V, D	1 人	SAM 系统	[1, 9] 的连续值
VAM	电视脱口秀	47	Vi + Au	A, V, D	6～34 人	SAM 系统	[−1, 1] 的 5 点等间隔值

注: Vi — 视觉模态, Au — 听觉模态, Ph — 生理信号, A — 唤醒维, V — 效价维, E — 期望维, D — 支配维, I — 强度维

表 2 维度情感预测文献总结

Table 2 Literature review of the dimensional emotion prediction

文献 (出版日期)	模态		特征	特征选择和降维	维度情感预测模型		信息融合方法
					回归模型	分类模型	
[49] (2008)		Au	声学特征	CFS	LSTM-RNN	CRF	–
[16] (2009)		Au	声学特征	–	–	HMM	–
[28] (2010)	Vi	头部运动	几何特征	–	SVR	–	–
[50] (2010)	Vi	步态	几何特征	PCA, KPCA, LDA, GDA	–	NN	–
[18] (2010)	Au	声音	声学特征	CFS	–	LSTM-RNN	FE
		语言	语言特征				
[51] (2010)	Au	声音	声学特征	–	LSTM-RNN	–	FE
		语言	语言特征				
[52] (2010)		Vi	几何特征	PCA, CFS	–	BLSTM	FE
		Au	声学特征				
[48] (2011)		Vi	LBP 特征	PCA, SPCA	SVR	–	FE + DE
		Au	声学特征				
[53] (2011)		Vi	几何特征	–	GMM	–	MO
		Au	声学特征				
[32] (2011)		Vi	头部姿势、面部运动单元	CFS	SVR	–	FE – 基于串的方法
		Au	笑声、叹息声、关键词				
[45] (2011)	Au	声音	声学特征	CFS	SVR	–	FE
		语言	BoCNG 特征				
[21] (2011)	Vi	面部	几何特征	–	BLSTM	–	OA
		肩膀					
		Au	声学特征				
[54] (2012)		Vi	几何特征	–	–	EWSC-HMM	MO
		Au	声学特征				
[55] (2012)	Vi	面部	面部表情	–	多模态推断系统	–	MO
		身体	几何特征				
		Au	语言与关键词信息				
[56] (2012)		Vi	多尺度动态视频特征	新的基于相关的特征选择	核回归	–	OA – 局部线性回归
		Au	声学特征				
[57] (2012)	Vi	面部	几何特征	–	OA-RVM	–	OA-RVM
		肩膀					
		Au	声学特征				
[31] (2013)		Vi	基于光流的低级特征	CFS	BLSTM	–	FE
	Au	声音	声学特征				
		语言	BoW 特征				
[58] (2013)		Vi	局部时空特征	–	SVR	–	DE – 加权和
		Au	声学特征				
[59] (2013)		Vi	几何特征	CSR	CSR	–	CSR
		Au	声学特征				
[60] (2015)		Vi	EOH, LBP, LBQ	–	PLS	–	DE – 线性回归
		Au	声学特征				

表 2 维度情感预测文献总结 (续)

Table 2 Literature review of the dimensional emotion prediction (continued)

文献 (出版日期)	模态	特征	特征选择和降维	维度情感预测模型		信息融合方法
				回归模型	分类模型	
[23] (2015)	Vi	LBP-TOP, LGBP-TOP, PHOG-TOP, HOG, 时空几何特征	–	随机森林	–	DE – 平均
	Au	声学特征				
	Ph	生理特征				
[61] (2015)	Vi	LGBP-TOP, 时空几何特征	–	SVM, RVM	–	OA
	Au	声学特征				
	Ph	时间和频域特征				
[62] (2015)	Vi	LGBP-TOP, LPQ-TOP, 时空几何特征	–	DBLSTM	–	DE-DBLSTM
	Au	声学特征				
	Ph	时间和频域特征				
[63] (2015)	Vi	LGBP-TOP, 时空几何特征	PCA	LSTM	–	FE, DE – 线性回归
	Au	声学特征				
	Ph	时间和频域特征				
[64] (2016)	Au	声学特征	–	DBLSTM	–	DE-ELM
[65] (2016)	Au	加强后的声学特征	–	SVR	–	–
[66] (2016)	Vi	LBP 特征	CFS	DNN-SKF	–	FE
	Au	声学特征				
	词汇	词汇特征				
[67] (2016)	Vi	CNN 特征	–	LSTM	–	DE – Kalman 滤波
	Au	声学特征				
	Ph	时间和频域特征				
[68] (2016)	Au	CNN 特征	–	LSTM	–	–
[69] (2016)	Vi	LGBP-TOP, 几何特征, CNN 特征	PCA	LSTM	–	DE-LSTM
	Au	声学特征				
	Ph	时间和频域特征				
[20] (2017)	Ph	通过 SAE 进行抽象的传统特征	–	Bayesian 模型	–	FE – 分层的特征融合网络

注: 若文中使用多种方法进行对比分析, 这里只列出性能最好的一种方法. Vi — 视觉模态, Au — 听觉模态, Ph — 生理信号, FE — 特征层融合, DE — 决定层融合 (决定层融合使用的具体方法), MO — 模型层融合, OA — 输出相关融合

KPCA (Kernel principal component analysis)[50], LDA (Linear discriminant analysis)[50], GDA (General discriminant analysis)[50] 等都可以用于维度情感预测中. 这些经典的特征提取、特征选择和降维方法使用广泛, 在很多综述文章 (如文献 [4, 8−9, 70] 等) 都有论述.

近年来, 深度学习技术得到了突飞猛进的发展, 在很多领域都得到了比较成功的应用. 运用深度学习技术进行特征提取和选择, 不仅可以减少人工的干预, 减少手工提取和选择特征的复杂性和盲目性, 而且提取的特征对于识别问题来说能够突出目标本质的差异性而忽略无关的差异性, 从而能够提高目标识别的准确性[71]. 因此, 研究者们也将深度学习技术应用到情感识别领域进行各个模态的特征提取和选择.

最常用于特征提取的深度网络是卷积神经网络 (Convolutional neural network, CNN), 它由多个单层卷积神经网络进行多次堆叠而成. 单层卷积神经网络一般包括卷积、非线性变换和下采样三个阶段, 如图 7 所示[72]. 每层的输入和输出为由一组向

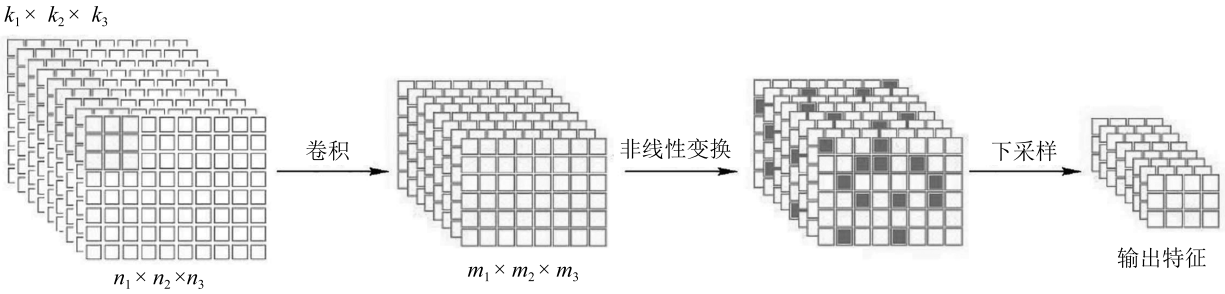


图 7 单层卷积神经网络的三个阶段

Fig. 7 The tree phases of the single layer convolutional neural network

量构成的特征图. 卷积阶段的卷积核决定了对输入特征图的观测模式, 不同的卷积核得到不同的特征; 非线性变换对卷积阶段得到的特征进行筛选; 下采样也称作池化, 其在减少数据量的同时能保留有用的信息. 在 CNN 最后一层的输出特征图后接一个全连接层和分类器, 即可实现分类或识别. CNN 每一层的输出都可看作是输入信号的特征, 可以单独用于不同的任务. 卷积神经网络的特点决定了其非常适合对图像进行特征提取, 因此在多模态维度情感预测中, CNN 经常被用于提取视觉模态的特征 (如文献 [67−69]). 对于其他模态的特征也可使用 CNN 进行特征提取, 例如 Zheng 等[73] 将语音信号转换为一系列二维矩阵, 作为 CNN 的输入来提取语音特征; Poria 等[74] 将文本中的每个词语根据 word2vec 词典以及词性表示成一个 306 维的向量, 一个句子中的所有词语对应的向量连接成一个向量作为 CNN 的输入进行特征提取. 遗憾的是运用 CNN 提取非视觉模态的特征只是用于离散情感识别中, 在维度情感预测中未见文献报告. 使用 CNN 进行特征提取遇到的问题主要是数据量不足, 从而导致过拟合现象, 为了解决此问题一般采取的方法是, 先使用其他库训练 CNN, 然后在目标库上进行特征提取, 例如 Chao 等[69] 使用在 CFW 和 FaceSrub 数据库上训练的 CNN 获取面部的表示.

由于情感的产生、发展和消退是一个过程, 为了获取更多的情感信息, 研究者们试图使用各种时空特征 (时空几何特征[23, 61]、时空纹理特征[23, 61] 等) 来提高维度情感预测的性能. 由于 LSTM (Long short-term memory) 具有对时间序列进行建模的能力, 因此也经常用来提取特征或提高特征的区分能力. Zhang 等[65] 为了消除自然环境下的加性噪声和卷积噪声对维度情感预测的影响, 基于 LSTM 的结构构建了循环去噪自编码 (Recurrent denoising autoencoder, RDA) 系统, 对传统声学特征进行特征增强, 获得了很好的效果. Wöllmer 等[18] 将 LSTM 与动态 Bayesian 网络 (Dynamic Bayesian networks, DBN) 相结合得到 LSTM-DBN 关键词检查器来获取二值的语言特征.

堆叠自编码 (Stacked autoencoder, SAE) 可以通过无监督的预训练和有监督的微调来确定系统的参数、提高特征的可区分性, 因此也常用来进行特征提取或对传统特征进行抽象. SAE 是以自编码器 (AutoEncoder, AE) 为基本单元堆叠而成的一种深度网络. AE 的结构如图 8 所示, 包括编码器和解码器两部分, 输入信号通过编码器得到编码, 再通过一个解码器得到输入信号的重构, 重构与输入信号对比得到重构误差. 编码器的输出编码即为抽象化的特征并作为下一层 AE 的输入. 逐层最小化重构误差, 确定编码和解码参数, 即可以实现 SAE 的无监督预训练, 在最顶层添加一个分类器, 运用有标签样本, 通过有监督学习可以实现对系统的参数微调. 但是对于 SAE 的层数以及每层神经元的个数一般需要使用者根据自己的经验确定. Yin 等[20] 提出了一种生理数据驱动的方法确定 SAE 的结构, 并使用 SAE 获取了各种传统生理信号特征的抽象表示, 进而实现维度情感分类.

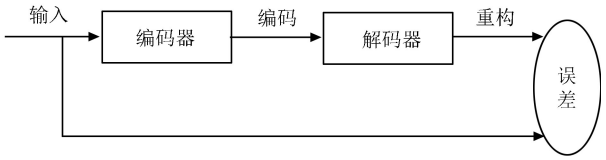


图 8 自编码器的结构

Fig. 8 Structure of autoencoder

4.3 预测模型

维度情感预测可以是一个分类问题也可以是一个回归问题, 当其是一个分类问题时, 常用的分类器如支持向量机 (Support vector machine, SVM)、K-最近邻分类器、隐马尔科夫模型 (Hidden Markov model, HMM) 等[9] 都可用于完成维度情感分类的任务; 当维度情感预测是一个回归问题时, 常用的回归模型如支持向量回归 (Support vector regression, SVR)、关联向量机 (Relevance vector ma-

chine, RVM) 等[9] 都可用于连续维度情感预测.

情感的产生、发展和消退是一个过程, 能够对各个模态的时间动态信息进行建模, 对提高维度情感预测的性能是有益的, 而 RNN (Recurrent neural networks) 正具有这样的优点, 因此 RNN 及其变形经常被用于维度情感预测中. RNN 的网络结构如图 9 所示, 图 9 右边是左边网络按时间展开的结果[71]. t 时刻的输出不仅与 t 时刻的输入有关, 而且还与历史状态有关, 因此它能够对时间序列进行建模. 但是当 t 时刻依赖的信息越来越久远时, RNN 学习到这些信息会越来越困难, 此时 RNN 的变形 LSTM 显示了优越性, 它对长期信息进行有选择的记忆是一种默认行为, 不需要付出很大的代价, 因此 LSTM 更加适合进行维度情感预测, 很多文献都使用了此模型 (如 [67−69] 等). LSTM 模型只能使用历史信息, 但未来信息对维度情感预测也是有用的, 为了将未来信息也用于维度情感预测中, 一些文献 (如 [21, 31]) 使用了 BLSTM (Bidirectional LSTM) 模型, 为了充分发掘特征与标签之间复杂的关系, 也有很多文献 (如 [62, 64]) 使用了由 BLSTM 堆叠构成的深度 BLSTM (Deep BLSTM, DBLSTM) 模型.

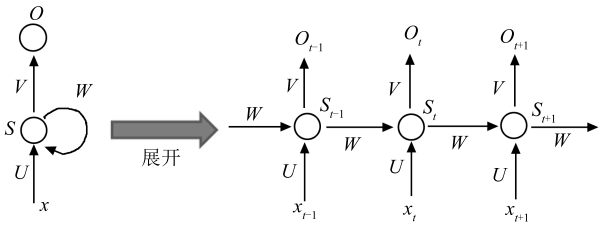


图 9 RNN 的网络结构

Fig. 9 Network structure of RNN

传统 RNN 以平方误差为代价函数, 而维度情感预测的目标是最大化预测标签与实际标签的相关性, 同时最小化它们的平均偏差, 为了更好地实现这个目的, Weninger 等[75] 将 RNN 的代价函数由平方误差更改为 CCC, 大大提高了连续维度情感预测的性能. Banda 等[76] 为了发挥 RNN 能够对较长的上下文依赖性进行建模的优点, 并加快收敛速度提高泛化能力, 使用了 NARX-RNN (Nonlinear AutoRegressive with eXogenous inputs recurrent neural network) 模型进行情感预测, 也获得了不错的效果. Pei 等[66] 将深度神经网络 (Deep neural network, DNN) 与切换卡尔曼滤波器 (Switching Kalman filter, SKF) 相结合提出了 DNN-SKF 框架, 先对输入特征和情感维度之间复杂的非线性关系用 DNN 进行建模, 然后用分段线性的 SKF 对情感的时间动态进行建模, 进而实现连续维度情感预测.

4.4 信息融合

理论上讲, 综合考虑多个模态以及其他信息能够提高情感识别系统的性能, 但是一个不恰当的融合方法不仅不能提高识别的性能, 可能还会降低识别的性能, 文献 [77] 仅用音频或视频模态进行情感识别, 所得平均识别率分别为 0.506 和 0.500, 但是运用音视频双模态融合进行情感识别的平均识别率仅为 0.47. 近些年研究者对信息融合进行了非常广泛的研究, 提出了很多融合方法, 其中用于维度情感预测的融合方法除了常见的特征层融合、决定层融合和模型层融合方法外, 针对维度情感预测的特殊性, 很多研究者将各个维度之间的关系用于维度情感预测过程中, 这类融合方法称为标签层融合.

特征层融合也称早期融合, 概念简单、容易理解和操作, 被广泛应用于维度情感预测中[51, 78]. Eyben 等[32] 为了将多个模态的行为事件 (例如微笑、摇头、叹息等) 用于各个情感维度的预测中, 使用特征层融合的思想提出了基于串的融合方法, 这也可以看作特征层融合的一个变形. 为了充分发掘不同模态之间复杂的非线性关系, 研究者提出了很多深层的特征融合方法, 并将其应用于维度情感预测中, Yin 等[20] 提出的基于多融合层的 SAE 集成分类器 (Multiple-fusion-layer based ensemble classifier of SAE, MESAE) 框架中, 多个模态的生理信号特征先经过 SAE 进行抽象, 再通过一个基于连通图的分层融合网络进行融合得到最后的抽象融合特征. 特征层融合中, 最难处理是不同模态数据的异步性, 为了处理这个难题, Chen 等[63] 在 LSTM 框架中将具有不同持续时间的特征输入到网络的不同层, 短时音频特征输入到第一隐层, 长时视频特征输入到第二隐层, 最长时间的 ECG 特征输入到第三隐层.

决定层融合也称后期融合, 也是一种操作简单的融合方法, 有着广泛应用. 在多模态维度情感预测任务中, 常用的决定层融合方法有求加权和[60]、求平均[79]、求中值[23] 和线性回归[22] 等. 为了对不同模态的预测结果之间复杂的关系进行建模, 近年来一些先进的机器学习技术也被用来进行决定层融合, 如 Kalman 滤波器[67]、极端学习机 (Extreme learning machine, ELM)[64]、DLSTM[62] 等. 但是, 决定层融合中默认的各个模态相互独立的假定与实际情形不符, 这也限制了最后的预测性能.

模型层的融合是设计一个模型将多个模态的信息以及其他方面的信息相结合来获取最终的情感预测结果. 设计同时实现多模态信息融合和维度情感预测的模型技巧性较强、困难较大, 文献中的工作也不是太多. Soladié 等[55] 设计了一个模糊推断系统, 将视频、音频和上下文相关特征进行融

合, 并对情感的效价维、唤醒维等四个维度的取值进行预测; Metallinou 等[53] 提出了一个高斯混合模型 (Gaussian mixture model, GMM) 融合多个音视频特征, 并对情感的唤醒维和支配维进行跟踪; Lin 等[54] 使用了误差加权半耦合隐马尔科夫模型 (Error weighted semi-coupled hidden Markov model, EWSC-HMM) 将音视频特征在模型层面进行融合, 并实现维度情感分类; Wu 等[80] 提出了双层半耦合隐马尔科夫模型 (Two-level hierarchical alignment-based SC-HMM, 2H-SC-HMM), 能够对视频和音频两个模态的时间阶段内部以及时间阶段之间的关系进行对齐矫正, 在此基础上对音视频信息进行融合并实现维度情感分类.

上面三类融合方法是经典的信息融合方法, 在多模态离散情感识别和多模态维度情感预测中都有应用, 但是对多模态维度情感预测来说, 所能使用的信息除了多模态信息外, 还有各个维度之间的关系, 将这些信息融入到多模态维度情感预测的过程中对于提高维度情感预测的性能是有益的, 这种融合方法称为标签层融合. Nicolaou 等[21] 基于心理学的研究结果 (情感的各个维度之间是有密切联系的) 首次将情感的各个维度之间的关系应用于多模态维度情感预测中, 提出了一个输出相关 (Output-associative, OA) 融合框架来利用各个情感维度间的相关性. 在此框架中, 对每个模态都使用 LSTM 分别对唤醒维和效价维进行预测, 将每个维度在每个模态上的预测结果作为输入再一次使用 LSTM 得到每个维度的最终估计, 如图 10 所示[21]. 此种 OA 融合框架与决定层融合类似, 最大的特点是使用了不同维度的预测结果来进一步得到某一维度最后的预测; 此融合框架中共进行了前后两次回归运算, 这两次回归运算使用的回归模型并不限于 LSTM, 可以使用其他的回归模型代替. 实际上很多文献也做了这样的工作, 例如 Nicolle 等[56] 使用了局部线性回归来融合基于不同模态的各个维度的预测. Nicolaou 等[57] 使用 RVM 代替 LSTM, 提出了 OA-RVM 回归框架, 并将输入特征与初步预测一起输入到一个 RVM 中, 得到最后的预测. Huang 等[61] 在使用 OA 和 OA-RVM 时将某一个时刻及其之前某一段时间的预测和输入特征连接, 输入到下一个回归模型中实现对这一时刻的维度情感预测, 以此来对上下文信息进行建模. Nicolaou 等[59] 为了利用每个情感维度之间以及每个维度与各个模态的特征之间的关系, 借助 CCA 的思想提出了 CSR (Correlated-spaces regression) 模型, 此模型先将所有模态的特征和标签运用 CCA 映射到变换空间, 然后在变换空间中学习特征到标签的映射, 在测试集中只需将在变换空间中的估计映回原始标签空间即可. CSR 模型使用了各个维度的相关性并且同时实现了特征的有监督降维和多模态融合, 也获得了较好的效果.

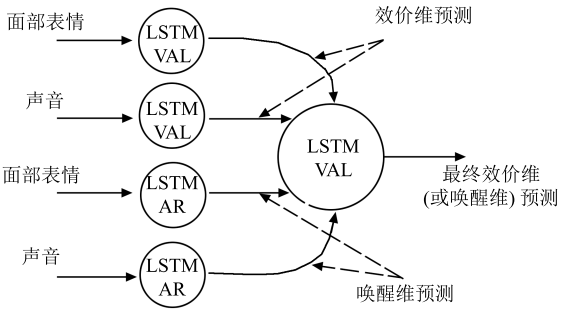


图 10　OA 融合框架

Fig. 10　OA fusion framework

4.5　其他信息的影响和应用

多模态维度情感预测的性能不仅受多个模态的特征提取、预测模型选取以及信息融合的影响, 而且受许多其他因素的影响, 要获取好的预测性能需要全面考察所有的影响因素.

在对每个情感维度进行实时标注时, 人的观察、评估以及反应都需要时间, 这造成了标注结果与情感表现之间有一个延时, 此延时与标注者、标注的维度、观察的行为都有关系[81]. 用合适的方法处理这种延时有助于提高维度情感预测的性能. Huang 等[61] 将标签的前 N 帧和特征的后 N 帧去掉实现标签和特征在时间上的对齐, 对最后的预测标签采用光滑滤波实现预测标签的延时以与基准标签在时间上对齐. 文中根据最后的预测性能寻找最佳延时, 获得了很好的预测效果. Nicolle 等[56] 认为特征与实际维度情感之间具有更强的相关性, 于是利用特征与延时标签的相关系数构建了延时概率分布, 基于此概率分布进行特征选择, 大大增强了预测结果的健壮性. Mariooryad 等[81−82] 通过最大化情感表现与延时标签的互信息获取最佳延时, 并对标签进行平移弥补延时造成的影响, 在基于面部和声音特征的维度情感分类中, 这种弥补相对于基准获得了超过 7 % 的增益.

5　对比与分析

维度情感预测一般是在自然的数据库上进行的, 这是一个比较困难的任务. 为了提高情感预测的性能, 研究者在特征提取、信息融合、预测模型的设计以及发掘维度情感预测性能的影响因素等方面都做了不懈努力. 但是, 由于文献使用的数据库、实验方法、分析的时间粒度、性能评价指标、使用的维度以及对每个维度的处理方法等都不尽相同, 因此很难进行详尽的对比分析. 这里仅对一些具有可比性的结果进行对比分析. 表 3 和表 4 是在常用数据库上

表 3 连续维度情感预测对比总结

Table 3 Comparison and summary of continuous dimensional emotion prediction

文献	数据库	模态	情感维度	特征	回归模型	融合方法	延时弥补 (Y/N)	应用维度间相关性 (Y/N)	最好平均预测性能 CC	最好平均预测性能 CCC
[78] (基准)	AVEC 2012	Vi	A, V, E, D	Vi LBP	SVR	–	N	N	0.09	–
		Vi + Au		Au 声学特征		FE	N	N	0.11	–
[58]	AVEC 2012	Vi	A, V, E, D	Vi 局部时空特征	SVR	–	N	N	0.41	–
		Vi + Au		Au 声学特征		DE – 加权和	N	N	0.42	–
[79] (基准)	AVEC 2014	Vi	A, V, D	Vi LGBP-TOP	SVR	–	N	N	0.20	–
		Vi + Au		Au 声学特征		DE – 加权和	N	N	0.36	–
[22] (基准)	AVEC 2015	Vi	A, V	Vi LGBP-TOP + 时空几何特征	SVR	–	N	N	0.29	0.20
		Vi + Au + Ph		Au 声学特征 Ph 时间和频域特性		DE – 线性回归	N	N	0.42	0.41
[47] (基准)	AVEC 2016	Vi	A, V	Vi LGBP-TOP + 时空几何特征	SVR	–	N	N	–	0.40
		Vi + Au + Ph		Au 声学特征 Ph 时间和频域特性		DE – 线性回归	Y	N	–	0.66
[55]	AVEC 2012	Vi + Au	A, V, E, D	Vi 面部表情 + 身体语言 Au 语句和关键词信息	多模态模糊推断系统	MO	N	N	0.43	–
[56]	AVEC 2012	Vi + Au	A, V, E, D	Vi 多尺度动态视频特征 Au 声学特征	核回归	OA – 局部线性回归	Y	Y	0.46	–
[61]	AVEC 2015	Vi + Au + Ph	A, V	Vi LGBP-TOP + 时空几何特征 Au 声学特征 Ph 时间和频域特性	SVM, RVM	OA-Regression	Y	Y	–	0.66
[62]	AVEC 2015	Vi + Au + Ph	A, V	Vi LGBP-TOP + LPQ-TOP + 时空几何特征 Au 声学特征 Ph 时间和频域特性	DBLSTM	DE-DBLSTM	Y	N	0.68	0.68

注: Vi — 视觉模态, Au — 听觉模态, Ph — 生理信号, A — 唤醒维, V — 效价维, E — 期望维, D — 支配维, FE — 特征层融合, DE — 决定层融合 (决定层融合使用的具体方法), MO — 模型层融合, OA — 输出相关融合

表 4 维度情感分类对比总结

Table 4 Comparison and summary of dimensional emotion classification

文献	数据库	模态	情感维度	特征	识别模型	信息融合方法	最好平均性能 (%) WA	最好平均性能 (%) UA
[83] (基准)	AVEC 2011	音频	A, V, E, D	声学特征	SVM	–	45.05	51.95
[31]	AVEC 2011	音频	A, V, E, D	声学特征	LSTM	–	65.2	58.5
[54]	SEMAINE	音频 + 视频	A,V	视频 几何特征 音频 声学特征	EWSC-HMM	模型层融合	–	78.1
[80]	SEMAINE	音频 + 视频	A,V	视频 几何特征 音频 声学特征	2H-SC-HMM	模型层融合	–	87.5

注: A — 唤醒维, V — 效价维, E — 期望维, D — 支配维, UA — 未加权准确性, WA — 加权准确性

进行连续维度情感预测和维度情感分类的对比总结, 给出的预测性能是相应文献中各个维度预测性能的平均值, 其中文献 [22, 47] 中基于视频特征的预测结果是基于纹理特征和几何特征所得预测结果的平均值, 文献使用多种方法的, 这里只列出获得最好预测性能使用的方法.

情感的产生、发展和消退是一个动态过程, 在特征提取时考虑时间变化, 在模型设计时考虑上下文的依赖关系, 都被证明对提高维度情感预测的性能是有效的. 文献 [58, 78] 基于视频的预测中, 在相同条件下使用局部时空特征的预测结果明显比使用静态 LBP 特征的预测结果好. 从 2014 年开始, AVEC 比赛都是以时空特征 (包括时空纹理特征和几何特征) 为基准视频特征, 虽然与 AVEC 2012 使用的数据库不同, 也大概可以看出, 与 AVEC 2012 基于视频特征的基准预测结果相比有了大幅的提高. 在选择分类/回归模型时, 使用能够对上下文的动态依赖关系建模的模型比使用静态模型的效果要好. 文献 [31] 采用 LSTM 模型对上下文信息进行建模, 使用 AVEC 2011 大赛组提供的音频特征进行维度情感分类, 平均准确率比 AVEC 2011 的基准平均准确率有了大幅度的提高.

各个模态的信息具有互为补充、互为印证的关系, 合理地利用它们来提高各个情感维度的预测性能也是非常有效的. 从表 3 可以看出, 多模态维度情感预测系统的性能普遍优于单模态维度情感预测系统. 多模态信息融合算法对预测性能的影响是巨大的, 文献 [55] 使用的多模态模糊推断系统的预测结果与 AVEC 2012 基准双模态预测结果相比具有很大的提升. 文献 [80] 使用的 2H-SC-HMM 模型, 具有对音视频两个模态的时间阶段内部以及时间阶段之间的关系进行对齐矫正的能力, 在 SEMAINE 库上进行维度情感分类的平均准确率达到了 87.5 %, 相比文献 [54] 使用的 EWSC-HMM 模型完成相同任务的平均准确率 78.13 % 有了大幅的提高.

多模态维度情感预测是一项复杂的工程, 其性能受到多种因素的影响, 好的预测系统往往综合考虑了各个方面的影响因素. 文献 [56] 使用多尺度动态视频特征, 考虑了反应延时问题, 使用局部线性回归融合从每个模态获得的各个维度的预测结果, 获得了目前 AVEC 2012 数据库上最好的预测性能 (平均 CC = 0.46). 文献 [61] 处理了标注延时的问题, 考虑了情感的各个维度的相关性问题, 使用基于输出相关融合框架的多模态系统在 AVEC 2015 数据库上获得了优异的预测性能 (平均 CCC = 0.66). 文献 [62] 利用 DBLSTM 具有对上下文的依赖性进行建模的优点, 将其应用于单模态预测和对每个模态的预测结果进行融合的过程中, 而且在进行单模态预测时进行了特征选择, 同时处理了标注延时的问题, 获得了 AVEC 2015 数据库上目前最好的预测性能 (平均 CCC = 0.68).

6 总结与展望

多模态维度情感预测涉及了心理学、生理学、社会科学等多个学科, 它的发展依赖于多个领域的成果和发现. 随着人工智能的发展和人机互动的迫切需要, 多模态维度情感预测受到越来越多研究者的关注, 近年来取得了很大进展. 本文通过对多模态维度情感预测研究现状的认识, 思考总结出其面临的挑战及发展趋势如下:

1) 各个情感维度的标记是一个十分耗费时间和精力并且需要一定技巧的工作, 这限制了维度情感数据集的建立. 因此, 充分应用有限的现有数据, 采用弱监督或半监督学习提升预测的泛化能力是一个亟待解决的问题.

2) 多个模态的情感数据一般是通过多种传感器获取的, 在获取过程中很难做到记录的同步性, 并且不同的模态对情感状态的表现也不是同步的, 在进行多模态维度情感预测中如何更好地处理这些异步性是一个挑战性的问题.

3) 各个模态蕴含的情感信息互为补充、互为印证, 而且受数据的获取条件以及个体的刻意控制等很多因素的影响, 会出现一个或多个模态信息的缺失, 因此如何更好地建立模型实现多模态信息融合是一个需要研究的问题.

4) 情感的维度信息与其他信息 (如情感的类别信息、社会行为信息等) 都具有密切的关系, 在维度情感预测过程中如何充分利用这些信息提高维度情感预测的性能是一个有趣的问题.

5) 在现有的多模态维度情感预测中, 对于生理信号和语言信息 (语音识别出的语言或文本中的语言) 的使用十分有限, 但是显然这两种信号能够为维度情感预测提供有用的信息. 因此如何从这两种信号中挖掘出对维度情感预测有用的信息是值得研究的.

6) 随着深度学习技术的发展以及在各个领域的成功应用, 多模态维度情感预测领域也不可避免地受到影响, 并且目前也有了一些应用. 但是如何更好的将深度学习技术应用于维度情感预测的各个环节, 深度学习技术在各个环节的应用能否优于传统的机器学习技术, 以及运用深度学习技术提升的预测性能相对计算成本的增加是否相匹配等, 都是需要充分研究的问题.

7) 由于人机互动的实时性需要, 提高多模态维度情感预测性能的同时降低计算量, 使多模态维度情感预测能够实时地进行具有很大的实际应用意义.

References

1 Liu Ye, Fu Qiu-Fang, Fu Xiao-Lan. The interaction between cognition and emotion. *Chinese Science Bulletin*, 2009, **54**(22): 4102−4116
(刘烨, 付秋芳, 傅小兰. 认知与情绪的交互作用. 科学通报, 2009, **54**(18): 2783−2796)

2 D′Mello S K, Kory J. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, 2015, **47**(3): Article No. 43

3 Zeng Z H, Pantic M, Roisman G I, Huang T S. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009, **31**(1): 39−58

4 Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. *Information Fusion*, 2017, **37**: 98−125

5 Yue Guo-An, Dong Ying-Hong. On the categorical and dimensional approaches of the theories of the basic structure of emotions. *Nankai Journal (Literature and Social Science Edition)*, 2013, (1): 140−150
(乐国安, 董颖红. 情绪的基本结构: 争论、应用及其前瞻. 南开学报(哲学社会科学版), 2013, (1): 140−150)

6 Arifin S, Cheung P Y K. Affective level video segmentation by utilizing the pleasure-arousal-dominance information. *IEEE Transactions on Multimedia*, 2008, **10**(7): 1325−1341

7 Cowie R, Douglas-Cowie E, Savvidou S, McMahon E, Sawey M, Schröder M. "FEELTRACE": an instrument for recording perceived emotion in real time. In: Proceedings of the 2000 ISCA Tutorial and Research Workshop on Speech and Emotion. Northern Ireland: ISCA, 2000. 19−24

8 Han Wen-Jing, Li Hai-Feng, Ruan Hua-Bin, Ma Lin. Review on speech emotion recognition. *Journal of Software*, 2014, **25**(1): 37−50
(韩文静, 李海峰, 阮华斌, 马琳. 语音情感识别研究进展综述. 软件学报, 2014, **25**(1): 37−50)

9 Gunes H, Schuller B. Categorical and dimensional affect analysis in continuous input: current trends and future directions. *Image and Vision Computing*, 2013, **31**(2): 120−136

10 Fontaine J R J, Scherer K R, Roesch E B, Eiisworth P C. The world of emotions is not two-dimensional. *Psychological Science*, 2007, **18**(12): 1050−1057

11 Zou Ji-Lin, Zhang Xiao-Cong, Zhang Huan, Yu Liang, Zhou Ren-Lai. Beyond dichotomy of valence and arousal: review of the motivational dimensional model of affect. *Advances in Psychological Science*, 2011, **19**(9): 1339−1346
(邹吉林, 张小聪, 张环, 于靓, 周仁来. 超越效价和唤醒 — 情绪的动机维度模型述评. 心理科学进展, 2011, **19**(9): 1339−1346)

12 Morris J D. Observations: SAM: the self-assessment manikin — an efficient cross-cultural measurement of emotional response. *Journal of Advertising Research*, 1995, **35**: 63−68

13 Koelstra S, Muhl C, Soleymani M, Lee J S, Yazdani A, Ebrahimi T, et al. DEAP: a database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 2012, **3**(1): 18−31

14 Busso C, Bulut M, Lee C C, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 2008, **42**(4): 335−359

15 Ringeval F, Sonderegger A, Sauer J, Lalanne D. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In: Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. Shanghai, China: IEEE, 2013. 1−8

16 Schuller B, Vlasenko B, Eyben F, Rigoll G, Wendemuth A. Acoustic emotion recognition: a benchmark comparison of performances. In: Proceedings of the 2009 IEEE Workshop on Automatic Speech Recognition and Understanding. Merano, Italy: IEEE, 2009. 552−557

17 Tarasov A, Delany S J. Benchmarking classification models for emotion recognition in natural speech: a multi-corporal study. In: Proceedings of the 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops. Santa Barbara, CA, USA: IEEE, 2011. 841−846

18 Wöllmer M, Schuller B, Eyben F, Rigoll G. Combining long short-term memory and dynamic Bayesian networks for incremental emotion-sensitive artificial listening. *IEEE Journal of Selected Topics in Signal Processing*, 2010, **4**(5): 867−881

19 Espinosa H P, García C A R, Pineda L V. Features selection for primitives estimation on emotional speech. In: Proceedings of the 2010 IEEE International Conference on Acoustics Speech and Signal Processing. Dallas, TX, USA: IEEE, 2010. 5138−5141

20 Yin Z, Zhao M Y, Wang Y X, Yang J D, Zhang J H. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer Methods and Programs in Biomedicine*, 2017, **140**: 93−110

21 Nicolaou M A, Gunes H, Pantic M. Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space. *IEEE Transactions on Affective Computing*, 2011, **2**(2): 92−105

22 Ringeval F, Schuller B, Valstar M, Jaiswal S, Marchi E, Lalanne D, et al. AV + EC 2015: the first affect recognition challenge bridging across audio, video, and physiological data. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. Brisbane, Australia: ACM, 2015. 3−8

23 Kächele M, Schels M, Thiam P, Schwenker F. Fusion mappings for multimodal affect recognition. In: Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence. Cape Town, South Africa: IEEE, 2015. 307−313

24 Sun Xiao, Pan Ting, Ren Fu-Ji. Facial expression recognition using ROI-KNN deep convolutional neural networks. *Acta Automatica Sinica*, 2016, **42**(6): 883−891
(孙晓, 潘汀, 任福继. 基于 ROI-KNN 卷积神经网络的面部表情识别. 自动化学报, 2016, **42**(6): 883−891)

25 Xu Feng, Zhang Jun-Ping. Facial microexpression recognition: a survey. *Acta Automatica Sinica*, 2017, **43**(3): 333−348
(徐峰, 张军平. 人脸微表情识别综述. 自动化学报, 2017, **43**(3): 333−348)

26 Ekman P. Universal facial expressions of emotion. *California Mental Health Research Digest*, 1970, **8**(4): 151−158

27 Kleinsmith A, Bianchi-Berthouze N. Affective body expression perception and recognition: a survey. *IEEE Transactions on Affective Computing*, 2013, **4**(1): 15−33

28 Gunes H, Pantic M. Dimensional emotion prediction from spontaneous head gestures for interaction with sensitive artificial listeners. In: Proceeding of the 10th International Conference on Intelligent Virtual Agents. Berlin, Heidelberg, Germany: Springer-Verlag, 2010. 371−377

29 Metallinou A, Yang Z J, Lee C C, Busso C, Carnicke S, Narayanan S. The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations. *Language Resources and Evaluation*, 2016, **50**(3): 497−521

30 Wang Ke, Xia Rui. A survey on automatical construction methods of sentiment lexicons. *Acta Automatica Sinica*, 2016, **42**(4): 495−511
(王科, 夏睿. 情感词典自动构建方法综述. 自动化学报, 2016, **42**(4): 495−511)

31 Wöllmer M, Kaiser M, Eyben F, Schuller B, Rigoll G. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 2013, **31**(2): 153−163

32 Eyben F, Wöllmer M, Valstar M F, Gunes H, Schuller B, Pantic M. String-based audiovisual fusion of behavioural events for the assessment of dimensional affect. In: Proceedings of the 2011 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops. Santa Barbara, CA, USA: IEEE, 2011. 322−329

33 Peng Ran-Ling. *General Psychology*. Beijing: Beijing Normal University Press, 2001.
(彭聃龄. 普通心理学. 北京: 北京师范大学出版社, 2001.)

34 Calvo R A, D′Mello S. Affect detection: an interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 2010, **1**(1): 18−37

35 Mckeown G, Valstar M, Cowie R, Pantic M, Schroder M. The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 2012, **3**(1): 5−17

36 Grimm M, Kroschel K, Narayanan S. The Vera am Mittag German audio-visual emotional speech database. In: Proceedings of the 2008 IEEE International Conference on Multimedia and Expo. Hannover, German: IEEE, 2008. 865−868

37 Lades M, Vorbruggen J C, Buhmann J, Lang J, von der Malsburg C, Wurtz R P, et al. Distortion invariant object recognition in the dynamic link architecture. *IEEE Transactions on Computers*, 1993, **42**(3): 300−311

38 Ahonen T, Hadid A, Pietikainen M. Face description with local binary patterns: application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, **28**(12): 2037−2041

39 Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 886−893

40 Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, USA: IEEE, 2001. I-511−I-518

41 Zhao G Y, Pietikäinen M. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, **29**(6): 915−28

42 Jiang B H, Valstar M, Martinez B, Pantic M. A dynamic appearance descriptor approach to facial actions temporal modeling. *IEEE Transactions on Cybernetics*, 2014, **44**(2): 161−174

43 Almaev T R, Valstar M F. Local Gabor binary patterns from three orthogonal planes for automatic facial expression recognition. In: Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. Geneva, Switzerland: IEEE, 2013. 356−361

44 Yang P, Liu Q, Metaxas D N. Boosting coded dynamic features for facial action units and facial expression recognition. In: Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE, 2007. 1−6

45 Schuller B. Recognizing affect from linguistic information in 3D continuous space. *IEEE Transactions on Affective Computing*, 2011, **2**(4): 192−205

46 Jenke R, Peer A, Buss M. Feature extraction and selection for emotion recognition from EEG. *IEEE Transactions on Affective Computing*, 2014, **5**(3): 327−339

47 Valstar M, Gratch J, Schuller B, Ringeval F, Lalanne D, Torres M T, et al. AVEC 2016: depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. Amsterdam, The Netherlands: ACM, 2016. 3−10

48 Sayedelahl A, Araujo R, Kamel M S. Audio-visual feature-decision level fusion for spontaneous emotion estimation in speech conversations. In: Proceedings of the 2013 IEEE International Conference on Multimedia and Expo Workshops. San Jose, CA, USA: IEEE, 2013. 1−6

49 Wöllmer M, Eyben F, Reiter S, Schuller B, Cox C, Douglas-Cowie E, et al. Abandoning emotion classes — towards continuous emotion recognition with modelling of long-range dependencies. In: Proceedings of the 2008 Interspeech. Brisbane, Australia: DBLP, 2008. 597−600

50 Karg M, Kuhnlenz K, Buss M. Recognition of affect based on gait patterns. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2010, **40**(4): 1050−1061

51 Eyben F, Wöllmer M, Graves A, Schuller B, Douglas-Cowie E, Cowie R. On-line emotion recognition in a 3-D activation-valence-time continuum using acoustic and linguistic cues. *Journal on Multimodal User Interfaces*, 2010, **3**(1−2): 7−19

52 Wöllmer M, Metallinou A, Eyben F, Schuller B, Narayanan S. Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling. In: Proceedings of the 11th Annual Conference of the International Speech Communication Association. Makuhari, Chiba, Japan: DBLP, 2010. 2362−2365

53 Metallinou A, Katsamanis A, Wang Y, Narayanan S. Tracking changes in continuous emotion states using body language and prosodic cues. In: Proceedings of the 2011 IEEE International Conference on Acoustics, Speech, and Signal Processing. Prague, Czech: IEEE, 2011. 2288−2291

54 Lin J C, Wu C H, Wei W L. Error weighted semi-coupled hidden Markov model for audio-visual emotion recognition. *IEEE Transactions on Multimedia*, 2012, **14**(1): 142−156

55 Soladié C, Salam H, Pelachaud C, Stoiber N, Séguier R. A multimodal fuzzy inference system using a continuous facial expression representation for emotion detection. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction. Santa Monica, California, USA: ACM, 2012. 493−500

56 Nicolle J, Rapp V, Bailly K, Prevost L, Chetouani M. Robust continuous prediction of human emotions using multiscale dynamic cues. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction. Santa Monica, California, USA: ACM, 2012: 501−508

57 Nicolaou M A, Gunes H, Pantic M. Output-associative RVM regression for dimensional and continuous emotion prediction. In: Proceedings of the 2012 IEEE International Conference on Automatic Face and Gesture Recognition and Workshops. Santa Barbara, CA, USA: IEEE, 2012. 16−23

58 Song Y, Morency L P, Davis R. Learning a sparse codebook of facial and body microexpressions for emotion recognition. In: Proceedings of the 15th ACM on International Conference on Multimodal Interaction. Sydney, Australia: ACM, 2013. 237−244

59 Nicolaou M A, Zafeiriou S, Pantic M. Correlated-spaces regression for learning continuous emotion dimensions. In: Proceedings of the 21st ACM International Conference on Multimedia. Barcelona, Spain: ACM, 2013. 773−776

60 Gaus Y F A, Meng H Y, Jan A, Zhang F, Turabzadeh S. Automatic affective dimension recognition from naturalistic facial expressions based on wavelet filtering and PLS regression. In: Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. Ljubljana, Yugoslavia: IEEE, 2015. 1−6

61 Huang Z, Dang T, Cummins N, Stasak B, Le P, Sethu V, et al. An investigation of annotation delay compensation and output-associative fusion for multimodal continuous emotion prediction. In: Proceedings of the 2015 International Workshop on Audio/Visual Emotion Challenge. New York, USA: ACM, 2015. 41−48

62 He L, Jiang D M, Yang L, Pei E C, Wu P, Sahli H. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. Brisbane, Australia: ACM, 2015. 73−80

63 Chen S Z, Jin Q. Multi-modal dimensional emotion recognition using recurrent neural network. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. Brisbane, Australia: ACM, 2015. 49−56

64 Li X X, Xianyu H, Tian J S, Chen W X, Meng F H, Xu M X, et al. A deep bidirectional long short-term memory based multi-scale approach for music dynamic emotion prediction. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing. Shanghai, China: IEEE, 2016. 544−548

65 Zhang Z X, Ringeval F, Han J, Deng J, Marchi E, Schuller B. Facing realism in spontaneous emotion recognition from speech: feature enhancement by autoencoder with LSTM neural networks. In: Proceedings of the 2016 Conference of the International Speech Communication Association. San Francisco, USA: ISCA, 2016. 3593−3597

66 Pei E C, Xia X H, Yang L, Jiang D M, Sahli H. Deep neural network and switching Kalman filter based continuous affect recognition. In: Proceedings of the 2016 IEEE International Conference on Multimedia and Expo Workshops. Seattle, WA, USA: IEEE, 2016. 1−6

67 Brady K, Gwon Y, Khorrami P, Godoy E, Campbell W, Dagli C, et al. Multi-modal audio, video and physiological sensor learning for continuous emotion prediction. In: Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge. Amsterdam, The Netherlands: ACM, 2016. 97−104

68 Trigeorgis G, Ringeval F, Brueckner R, Marchi E, Nicolaou M A, Schuller B, et al. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In: Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016. 5200−5204

69 Chao L L, Tao J H, Yang M H, Li Y, Wen Z Q. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge. Brisbane, Australia: ACM, 2015. 65−72

70 Sariyanidi E, Gunes H, Cavallaro A. Automatic analysis of facial affect: a survey of registration, representation, and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, **37**(6): 1113−1133

71 Lecun Y, Bengio Y, Hinton G. Deep learning. *Nature*, 2015, **521**(7553): 436−444

72 Yin Bao-Cai, Wang Wen-Tong, Wang Li-Chun. Review of deep learning. *Journal of Beijing University of Technology*, 2015, **41**(1): 48−59
(尹宝才, 王文通, 王立春. 深度学习研究综述. 北京工业大学学报, 2015 **41**(1): 48−59)

73 Zheng W Q, Yu J S, Zou Y X. An experimental study of speech emotion recognition based on deep convolutional neural networks. In: Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction. Xi′an, China: IEEE, 2015. 827−831

74 Poria S, Chaturvedi I, Cambria E, Hussain A. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In: Proceedings of the 16th IEEE International Conference on Data Mining. Barcelona, Spain: IEEE, 2016. 439−448

75 Weninger F, Ringeval F, Marchi E, Schuller B. Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence. New York, USA: AAAI Press, 2016. 2196−2202

76 Banda N, Engelbrecht A, Robinson P. Continuous emotion recognition using a particle swarm optimized NARX neural network. In: Proceedings of the 2015 International Conference on Affective Computing and Intelligent Interaction. Xi′an, China: IEEE, 2015. 380−386

77 Glodek M, Tschechne S, Layher G, Schels M, Brosch T, Scherer S, et al. Multiple classifier systems for the classification of audio-visual emotional states. In: Proceedings of the 2011 International Conference on Affective Computing and Intelligent Interaction. Berlin, Heidelberg, German: Springer-Verlag, 2011. 359−368

78 Schuller B, Valstar M, Cowie R, Pantic M. AVEC 2012: the continuous audio/visual emotion challenge — an introduction. In: Proceedings of the 14th ACM International Conference on Multimodal Interaction. Santa Monica, California, USA: ACM, 2012. 361−362

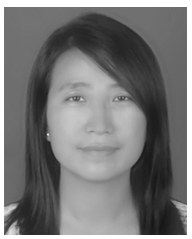
79 Valstar M, Schuller B, Smith K, Almaev T, Eyben F, Krajewski J, et al. AVEC 2014: 3D dimensional affect and depression recognition challenge. In: Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge. Orlando, Florida, USA: ACM, 2014. 3−10

80 Wu C H, Lin J C, Wei W L. Two-level hierarchical alignment for semi-coupled HMM-based audiovisual emotion recognition with temporal course. *IEEE Transactions on Multimedia*, 2013, **15**(8): 1880−1895

81 Mariooryad S, Busso C. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Transactions on Affective Computing*, 2015, **6**(2): 97−108

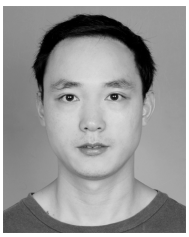
82 Mariooryad S, Busso C. Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations. In: Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. Geneva, Switzerland: IEEE, 2013. 85−90

83 Schuller B, Valstar M, Eyben F, McKeown G, Cowie R, Pantic M. AVEC 2011 — the first international audio/visual emotion challenge. In: Proceedings of the 2011 International Conference on Affective Computing and Intelligent Interaction. Berlin, German: Springer-Verlag, 2011. 415−424



李 霞 南京邮电大学通信与信息工程学院博士研究生. 2002 年获得曲阜师范大学数学与应用数学系学士学位, 2005 年获得南京大学应用数学系硕士学位. 主要研究方向为情感计算, 模式识别, 机器学习和计算机视觉.
E-mail: lx19800102@163.com
(**LI Xia** Ph. D. candidate at the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications. She received her bachelor degree in mathematics and applied mathematics from Qufu Normal University in 2002 and master degree in applied mathematics from Nanjing University in 2005, respectively. Her research interest covers affective computing, pattern recognition, machine learning, and computer vision.)

卢官明 南京邮电大学通信与信息工程学院教授. 1985 年和 1988 年获得南京邮电大学无线电工程学士学位和通信与电子系统硕士学位, 1999 年获得上海交通大学通信与信息系统博士学位. 主要研究方向为图像处理, 情感计算, 机器学习. 本文通信作者.
E-mail: lugm@njupt.edu.cn
(**LU Guan-Ming** Professor at the College of Communication and Information Engineering, Nanjing University of Posts and Telecommunications. He received his bachelor degree in radio engineering and master degree in communication and electronic systems from Nanjing University of Posts and Telecommunications in 1985 and 1988, respectively, and Ph. D. degree in communication and information systems from Shanghai Jiao Tong University in 1999. His research interest covers image processing, affective computing, and machine learning. Corresponding author of this paper.)



闫静杰 南京邮电大学通信与信息工程学院讲师. 2006 年和 2009 年获得中国矿业与技术大学电子科学与技术学士学位和信号与信息处理硕士学位. 2014 年获得东南大学信息与通信工程博士学位. 主要研究方向为模式识别, 情感计算, 计算机视觉和机器学习.
E-mail: yanjingjie1212@163.com
(**YAN Jing-Jie** Lecturer at the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications. He received his bachelor degree in electronic science and technology in 2006 and master degree in signal and information processing in 2009 from China University of Mining and Technology, and Ph. D. degree in signal and information processing from Southeast University in 2014. His research interest covers pattern recognition, affective computing, computer vision, and machine learning.)

张正言 南京邮电大学通信与信息工程学院博士研究生. 2004 年和 2007 年获得江苏科技大学电子信息工程学士学位和信号与信息处理硕士学位. 主要研究方向为模式识别, 机器学习和计算机视觉.
E-mail: zhangzhengyan@just.edu.cn
(**ZHANG Zheng-Yan** Ph. D. candidate at the College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications. He received his bachelor degree in electronic information engineering and master degree in signal and information processing from Jiangsu University of Science and Technology in 2004 and 2007, respectively. His research interest covers pattern recognition, machine learning, and computer vision.)