基于姿态的情感计算综述

付心仪[1,2,3], 薛程[4], 李希[1], 张玥泽[5], 蔡天阳[6]

[1] (清华大学美术学院 北京 100084)
[2] (清华大学未来实验室 北京 100084)
[3] (清华大学-阿里巴巴自然交互体验联合实验室 北京 100084)
[4] (中国传媒大学动画与数字艺术学院 北京 100024)
[5] (Department of Informatics, Technical University of Munich Munich 80333)
[6] (School of Information Engineering, University of Technology of Compiegne Compiegne 60200)
(fuxy@tsinghua.edu.cn)

摘 要: 情感计算的理论与算法研究是近年来人机交互领域的热点话题. 当前, 常见的情感计算集中在基于面部表情、语音、文本、人体姿态等方向, 既有单一模态的算法, 又有多模态的综合算法. 基于面部表情和语音模态的算法占据多数, 国内外基于人体姿态的算法相对较少. 文中针对基于姿态的情感计算所面临的几个关键科学问题展开了综述, 包括情感的心理学模型、人体姿态估计算法、姿态的情感特征提取算法、情感分类与标注算法、姿态情感数据集、基于姿态的情感识别算法等. 具体来说, 首先介绍了几种常用的情感计算心理学模型, 评述了各类模型的适用场景; 随后从人体检测和姿态估计 2 个角度对人体姿态估计的常用算法进行了总结, 并讨论了 2D 和 3D 姿态估计的应用前景. 针对特征提取算法, 分析了基于全身和上半身身体动作的姿态特征提取算法. 在情感标注方面, 介绍了表演数据和非表演数据的情感标注算法, 并指出了半自动或自动的标注非表演数据将是未来的重要发展趋势之一. 针对姿态情感数据集, 列举了近年来常见的 14 个数据集, 并主要从是否是表演数据、数据维度、静态或动态姿势、全身或非全身数据等几个方面进行了总结. 在基于姿态的情感识别算法方面, 主要介绍了基于人工神经网络的情感识别算法, 指出了不同算法的优劣之处和适用的数据集类型. 文中的综述研究, 总结提炼了国内外该领域经典且前沿的工作, 希望为相关的研究者提供研究帮助.

关键词: 情感计算; 姿态; 情感特征; 情感标注; 情感识别
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2020.18350.z43

A Review of Body Gesture Based Affective Computing

Fu Xinyi[1,2,3], Xue Cheng[4], Li Xi[1], Zhang Yueze[5], and Cai Tianyang[6]

[1] (*Academic of Art & Design, Tsinghua University, Beijing* 100084)
[2] (*The Future Laboratory, Tsinghua University, Beijing* 100084)
[3] (*Tsinghua University-Alibaba Joint Research Laboratory for Natural Interaction Experience, Beijing* 100084)
[4] (*School of Animation and Digital Art, Communication University of China, Beijing* 100024)
[5] (*Department of Informatics, Technical University of Munich, Munich* 80333)
[6] (*School of Information Engineering, University of Technology of Compiegne, Compiegne* 60200)

Abstract: Research on the theory and method of affective computing has been a hot topic in the field of human-computer interaction in recent years. At present, the common research on affective computing in related fields focuses on facial expression, speech, text, human gesture, and other directions. There are both sin-

收稿日期: 2020-03-03; 修回日期: 2020-04-19. 基金项目: 国家重点研发计划(2019YFF0302902); 清华大学自主科研计划(20197010003). 付心仪(1989—), 女, 博士, 助理研究员, CCF 会员, 主要研究方向为人机交互、情感计算、用户体验、文化遗产数字化; 薛程(1994—), 男, 硕士, 主要研究方向情感计算、大视场虚拟现实设备、游戏制作; 李希(1996—), 女, 硕士研究生, 主要研究方向情感计算、差分渲染、非真实感绘制; 张玥泽(1995—) 男, 硕士研究生, 主要研究方向为计算机视觉, 混合现实; 蔡天阳(1996—), 男, 硕士, 主要研究方向为计算机视觉, 情感计算.

gle-modality research and multi-modality comprehensive research. Among them, the researches based on facial expressions and speech modalities are the majority, and the research based on human gesture is relatively few. In this paper, we conduct a research survey on several key problems faced by gesture-based affective computing, including the emotional psychological model, human pose estimation, body emotional feature extraction method, emotion classification and labeling method, gesture-emotion dataset, and gesture-based emotion recognition algorithm. Specifically, we first introduce several commonly used emotional computing psychology models, review the applications of various models. Then we summarize the common methods of human pose estimation from two perspectives of human detection and pose estimation, and discuss the application prospects of 2D and 3D pose estimation. For feature extraction methods, we analyze feature extraction methods based on body movements of the whole body and the upper body. In the aspect of emotion annotation, we introduce the emotion annotation methods of performance data and non-performance data. We also point out that semi-automatic or automatic labeling of non-performance data would be one of the important development trends in the future. For the posture and emotion datasets, we list 14 most commonly used datasets in recent years classified by performance or non-performance data, data dimensions, static or dynamic poses, full-body or non-full-body data. In terms of gesture-based emotion recognition algorithms, we mainly introduce emotion recognition algorithms based on artificial neural networks, pointing out the advantages and disadvantages of different methods and their applicable datasets. This article reviews and summarizes the classic and cutting-edge research work in related fields, hoping to provide a good research basis for researchers in similar directions.

Key words: affective computing; body gesture; emotional features; emotion annotation; emotion recogonition

情感是人类表达内心感受的重要途径，是每个人所必需的精神活动. 情感的分类是多元而复杂的，人们通常根据情感表达时的主要特征对其进行离散的分类. 20 世纪 70 年代, Ekman 提出开心(happiness)、伤心(sadness)、害怕(fear)、嫌弃(disgust)、生气(anger)和惊讶(surprise)6 种基本情感(emotion)分类[1]. 此外，也可以对情绪(sentiment)进行 2 极的分类：积极(positive)和消极(negative). 更进一步，情感也可以使用激励(arousal)-效价(valence)等模型[2]进行量化表示.

近年来，情感计算的理论与方法是人机交互领域的热点话题，也是国际上前沿的研究方向. 目前，常见的情感计算集中在基于面部表情、语音、文本、人体姿态等方向，既有单一模态的，又有多模态的综合研究. 多年来，情感计算的研究主要集中在面部表情、语音分析和文本分析上，尽管已有一些身体姿态的情感计算相关文献[3-7]，但总体上相对偏少.

根据 Ekman[8]提出的观点，人们在试图理解他人情感时会更注重面部表情而不是身体姿势. 然而，相关的经典实验心理学研究表明，肢体语言也是情感信息的重要来源. Bull[9]发现，一部分情感与不同的身体姿势和动作有关，如兴趣或无聊，同意或不同意. Pollick 等[10]发现，借助于特定的手臂运动，人们能够以显著高于基准水平的准确度辨别基本情感. Coulson[11]强调静态身体姿势在识别任务中的作用. 相关研究表明，对于人类表达的影响，实际语言只占 35%，非语言的信号占 65%[12]. 所以，身体姿态是人类表达情感的重要环节.

基于人体姿态的情感计算有广阔的应用空间[13]，在很多大尺度场景(如商场、车站、广场等公共场所)中，用户的表情、声音等属于微观情感，需要近距离地交互才可以采集到，而用户的动作姿态也是表达感情的重要载体，目前尚未得到充分的利用. 此外，对于失聪失语人群、面部表情障碍人群等，语音和表情的情感表达较难实现，动作姿态是他们表达感情的主要通道.

用户动作姿态具有空间尺度大、数据容易获得、不同情感数据变化明显等特征. 因此，以用户动作姿态为基础的多模态情感计算，在智能家居、智能宠物、智能传媒、智能健康、安防保全等领域都有转化为产品的落地空间.

基于姿态的情感计算涉及一系列必要环节，包括情感建模、数据获取、数据集构建、数据标注、特征提取、识别算法等. 本文针对这些环节所面临的几个关键问题和采用的方法，展开综述，包括情感计算所使用的心理学模型、人体姿态估计方法、姿态的情感特征提取及其方法、情感的分类与标注方法、姿态情感数据集总结、基于姿态的情感识别算法等，总结提炼了国内外该领域经典且前沿的研究工作，针对基于姿态的情感计算所面临的问题进行诠释，同时以独特的视角评述了相关研究的优劣，希望为相似研究者提供帮助.

1 情感计算的心理学模型

情感计算的基点是人类的情感，计算机要量化人类的情感，就少不了基于心理学的情感建模. 情感理论的研究可以追溯到 19 世纪，人们已经提出多种理论与分类方法. 当前，最流行的情感模型可以分为离散模型、维度模型和成分模型 3 类[14].

在情感计算中，最先被 Ekman 等应用的是离散的分类情感模型，将情感分为 6 种或更多种的基础情感.

为了更精确地描述情感，在 21 世纪以来数据驱动的发展背景下，出现了连续的维度情感模型. 1980 年，Russell[15]提出双极圆周模型(bipolar circumplex model, BCM)，将情感分为价值维度和唤醒维度，如图 1 所示. Plutchik[16]提出 3D 圆周模型(3D circumplex model)，将离散与连续的情感模型结合起来，把人的情感分为 8 种基础情感并与色环上的 8 种颜色一一对应，高度则表示情感的密度. 3D 圆周模型在使用方面也可以更加灵活，如同颜色一般，混合的情感可以用来表达二级情感. 1996 年，Mehrabian[17]提出愉悦度、激活度和优势度(pleasure-arousal-dominance, PAD)模型，在价值维度和唤醒维度的基础上增加了支配维度，用来衡量人对于他人或者环境的支配能力.

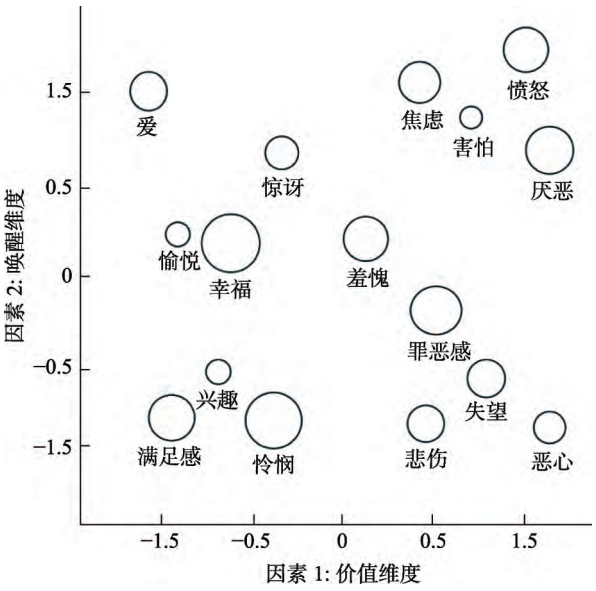

图 1 基于唤醒与价值的 2D 情感模型[15]

2001 年，Plutchik[18]提出介于离散模型和连续模型之间的成分模型，它是一个层级的结构，每层的情感由单独的情感和上一层的结果合成而来.

情感模型的选择由研究目的和应用场景决定，各个模型之间并没有明显的优劣差别. 值得一提的是，尽管 Fontaine 等[19]2007 年提出情感不应仅仅被分为 2D，但在情感计算的各种实践中，Russell 的情感模型依旧被广泛认可与使用.

2 人体的姿态估计

人体的姿态估计分为 2 个步骤：(1) 对人体进行检测，将背景中的无关信息裁剪掉；(2) 检测和跟踪人体姿势，以减少由姿势引起的数据的不相关变化.

2.1 人体检测

人体的身体、姿势、服饰等都不是刚体，导致外观会有较大的差别；同时，环境中光照和遮挡的变化也会增加检测难度. 通常，图像中的人体检测利用一个矩形框进行确认，首先提取出潜在的候选区域，然后将这些区域分为人类区域和非人类区域，得出最终的判定. 如果可以获取到区域的深度信息，就可以对人类和非人类空间的搜索区域进行限制，并对背景的裁剪进行简化. Viola 等[20]首先使用层级结构进行人体检测，并使用自适应增强算法(AdaBoost)进行特征选择. 为了提升性能，Dalal 等[21]使用基于梯度的特征，并对方向梯度直方图(histogram of oriented gradient, HOG)进行了推广. 早期对于人体的检测是先对人体部位分别进行检测，然后进行基于几何形状去拼装，而不是基于动力学去链接. Felzenszwalb 等[22]使用可变形部件模型(deformable part models, DPM)建立身体部位的关联，并将未确定部件的位置信息作为潜变量，提出潜在支持向量机(latent support vector machine, SVM). 传统方法虽然时间效率较高，但无法充分利用图像信息，导致识别准确率并不高. 随着数据量和计算性能的提升，神经网络变得更

加实用. 早期的神经网络速度较慢，为了提升速度，Angelova 等[23]使用多层级的神经网络，通过小型神经网络筛选出可能含有人体的图像区域，再使用大型神经网络去识别精准的位置. 通过这种方式，速度和性能可以得到一个平衡，基于特征的卷积神经网络(convolutional neural network, CNN)也因而被广泛使用. Girshick[24]提出基于区域的快速 CNN (fast R-CNN)，通过将分类和区域选择联合起来，提升了速度. Zhang 等[25]将图像的卷积特征与检测网络共享，引入区域生成网络(region proposal network, RPN)进一步提升了速度. He 等[26]使用残差神经网络(ResNet)，极大地提升了网络层数，达到实时检测人体的目标.

2.2　姿态估计

搜索空间的高维度、背景的杂乱、人体信息的复杂、光照等环境的干扰、身体遮挡等因素，提升了身体姿态估计的难度[27]. 身体姿态估计主要分为 2D 姿态估计和 3D 姿态估计.

2D 姿态估计包含人体关键骨骼点的坐标信息以及它们的连接状态. 2D 姿态的检测，又可分为单人姿态估计与多人姿态估计. 单人姿态估计中，人们最先采用坐标回归的方法去计算关节点位置. DeepPose 是第 1 个使用深度学习进行姿态估计的模型，它利用多阶段回归方法，得到 2D 坐标[28]. 由于姿态数据较为复杂，通过回归得到的坐标信息不够准确，因此出现了基于概率的热图检测. 热图检测首先计算像素与关节点的距离，并转换为概率值，然后进行训练. Tompson 等[29]将 CNN 与结构图结合，对关节点的局部信息和全局信息进行综合，确保了定位的准确. 多人姿态估计主要分为 2 种方法: (1) 自上到下的估计. 先估计人体后，再估计他们的姿态. 自上到下的人体姿态估计中，第 1 步是进行人体检测，常见的模型包括 R-CNN[30] 和 Faster R-CNN[31]等. He 等[32]先使用 Faster R-CNN 模型，再使用全卷积神经网络(fully convolutional network, FCN)进行进一步的检测，提出基于掩模区域的卷积神经网络(mask R-CNN)模型. (2) 自下到上的估计. 先估计身体的各个部位后，再将它们分别拼接到各个身体部位上. 自下到上估计的难点在于对于身体部位的聚类. 常见的聚类算法有部件亲和场[33]、语义部分分割[34]和 DeeperCut[35]等.

3D 姿态估计分为 2 种: (1) 基于人体骨骼点的稀疏模型，根据 2D 人体骨骼点坐标以及 2D 姿态特征进行直接预测. Wang 等[36]使用自监督机制，使得 2D 数据与 3D 数据可以互相矫正; Zhou 等[37]利用人体结构的比例信息，将 2D 坐标进行转换得到关节点的 3D 坐标. (2) 密集型的 3D 姿态模型建模，不仅需要得到骨骼关节的信息，也要将人体皮肤像素映射到 3D 模型表面. DensePose[38]使用 UV 坐标将多人皮肤线性(skinned multi-person linear, SMPL)模型表面的顶点参数化，采用人工构建数据集，并进行标记; BodyNet[39]则基于 2D 坐标，进行 3D 坐标的体素重建，使用堆叠沙漏网络结构(stacked hourglass networks, SHN)，相较于 SMPL 模型，体素网格的结构简单，可以提升效率.

当前，2D 姿态估计已经较为成熟，而 3D 姿态估计的识别准确率和识别时间效率均有很大的研究空间.

3　姿态的情感特征提取

不同于普通的姿态检测特征，基于姿态的情感计算特征依赖于姿态情感特征的提取，相关研究包括对整个身体运动的分析和对上身运动的分析.

3.1　基于全身姿态的情感特征提取

已有的研究中，人们发现，肢体情感语言具有初等动作原语和组合的语法特性，这为基于肢体运动的情感计算奠定了基础. 与传统的对面部表情的情感分析相比，对于全身现有的特征分析的难点在于肢体在 3D 空间范围内的运动具有 6 个自由度，并且存在各种动作组合的可能性. 此外，实现情感识别分类的难点还在于个体运动的偏差问题，即不同的人对于相同的情感倾向于使用不同的方式表达，这种个人风格增加了分类的难度.

现有的常见情感姿态特征分析基于身体运动学的粗粒度特征，如前倾或者后仰的分析. 采集系统一般都基于多摄像头同步录制系统，以重现和追踪肢体运动的 3D 轨迹.

Kapur 等[40]展示了基于简单的运动动力学的统计量度(如速度和加速度)，已经足以完成对自动分类器的构造(SVM 和决策树). 他们采用的采集系统 VICON 由 6 个摄像头和轻质可穿戴标记组成. 原始数据包括对每种情感表现进行 25 次 10 s 的 120 Hz 采样，根据插值计算肢体动作的 3D 空间轨迹，之后利用 Matlab 程序对每个 3D 坐标点求得一阶和二阶导数，即速度和加速度. 由于实验更加关注在长时间范围内的动态运动，Kapur 等[40]还求得了速度和加速度的中值和标准差. 实验结果表明，尽管这种数据分析看似简单，却已经足以构造相当精确的分类器(识别率高达 93%).

Bianchi-Berthouze 等[41]基于身体关节之间的角度和距离给出对于姿势的一般描述，并由此创建了一个情感姿势识别系统，该系统使用关联神经网络将姿势特征集映射到情感类别.

进一步，Bernhardt 等[42]基于对运动学的统计量度分析，在非样式化的身体动作中分析情感表达，用 15 个关节表示采样的身体骨骼结构，并根据身体体型统一化身体局部坐标系来获得保持不变的旋转和测量比例. 通过对肢体能量的阈值化分析，可以对每个动作进行分段并判断是否在关键帧中存在肢体动作.

拉邦动作分析(Laban movement analysis, LMA)理论是一种描述、解释、记录人类动作的理论体系，被广泛用于表演艺术和舞蹈艺术中. 近年来，受 LMA 理论的启发，人体姿态的情感特征有了新的进展. Aristidou 等[43]基于 LMA 理论设计了一组 86 维的 3D 肢体特征，用于分类戏剧表演中肢体动作的情感；他们定义了关节点之间最大(最小/平均)距离、速度、加速度等，并使用极度随机树(extremely randomized trees, ET)与 SVM 实现情感分类任务. 与其他方法相比，该方法具有更大的普适性，适用于更多场景的姿态情感特征计算. Luo 等[44]同样利用 LMA 理论来提取姿态的情感特征，构建了肢体语言数据集(body language dataset, BoLD)，并在多种学习模型上进行测试. Ajili 等[45]根据 LMA 理论，提出一种带有底层语法结构的人体动作描述语言，用于量化不同类型的人类情感.

3.2 基于上部肢体运动的情感特征提取

情感特征提取的另一个研究趋势是关注上身运动. 正如 Freedman[46]所指出的，除了步态之外，手部解释是用于表达能力研究的最常见的可量化的公开行为. Siple[47]认为，以手语作为辅助交流工作时，即使交流者看着对方的脸，他依旧可以非常自然地且高灵敏度地捕捉到对方在面部和上半身区域内的微小的动作细节.

除去手单独的动作捕捉外，手与头、肩的相对位置关系与运动同样包含着情感信息. McNeill[48]证实，上部肢体，尤其是头部和手部的运动学信息最能表现被测试对象的情感表达，人们自发地会试图根据他人的上肢运动来推断交谈过程中他人的情感.

2008 年，Gunes 等[49]提出一种基于面部和手部的双模态模型，来识别 12 种情感状态，使用光流法跟踪了测试对象的手、头和肩部区域，并分析了质心、旋转角度、面积的长度宽度以及面积大小，最后使用 SVM 进行分类.

针对面向未来移动设备上情感计算的需求，如何在相对较低运算能力和资源相对有限的传感器系统中设计实现情感检测，Glowinski 等[50]在 2011 年提出使用简化的视觉信息输入，即头部和手部的位置以及速度来推断情感信息，并提出了常用的计算情感的上肢特征，如能量、空间属性、平滑度、对称性、头部倾向等. 与其他研究不太相同，Glowinski 等[50]分析了非语言的手势特征，以及采用低分辨率和低视觉信息量的数据作为输入.

Patwardhan 等[51]使用 3D 的头部和上半身的骨骼点运动特征，有助于提高基于机器学习的多模态情感识别系统的准确率. 与 2D 特征相比，3D 特征可以记录更多的位置信息，对于姿态情感计算，具有更大的优势.

从总体上说，全身和半身的姿态选择可以对应于不同的应用场景，应该选择更符合自己研究场景的姿态部位.

4 情感标注方法

情感标注是情感计算领域中最棘手的问题之一，原因是真相很难获得，标注也仅限于 1 个或 2 个评分者，主观情感和标注情感之间的一致性很差，特别是在自然姿态的数据中. 这会促使研究人员关注具有表演性质的姿态，然而即使在已知情感表达的一组动作数据中，解释其中某个姿势或者是手势的表达时，也有很大的主观空间，选择表达情感内容来影响所选数据质量的情况并不少见. Xia 等[52]指出，根据姿态行为和风格来自动地标注动作非常困难，尤其是将动作标注为一个特定的类别.

对于表演性数据集，研究者往往根据预先设定的情感类别直接进行标注. Aristidou 等[53]首先从 BCM[15]中挑选情感作为情感标签，然后通过动作捕捉方法记录舞蹈家表达情感的动作，最后给定情感标签.

由于没有情感的预设，非表演数据集情感标注的主观性和准确性就变得十分具有挑战性. 2011 年，在对于非表演动作的姿态识别中，Kleinsmith 等[54]提出一种多人协同的半自动标注方法. 他们认为，人们通常不会给自己的姿势贴上标签，或是清晰地记住自己在有某种情感时的动作，所以在测试结束后让被试者自己提交情感报告是不可靠

的，且在测试过程中阻止被试者且询问他们当前的情感状态也并不可行. 如图 2 所示，Kleinsmith 等[54]将整个观察者池随机分成 3 个子集，每个观察者子集都被要求给姿势标注真实情况(ground truth, GT)标签. 将观察者子集 1 的 GT 标签与观察者子集 2 的 GT 标签进行比较，确定人类识别一致性的基准率，2 个观察者子集的一致性水平和可靠性是通过计算两者在情感类别的整个姿势集上匹配的 GT 标签的数量而得到的. 第 3 个子集确定的 GT 被保留下来，供以后建立自动识别模型. 然后根据观察者子集 1 的 GT 对模型进行测试. 整个过程将重复 10 次.

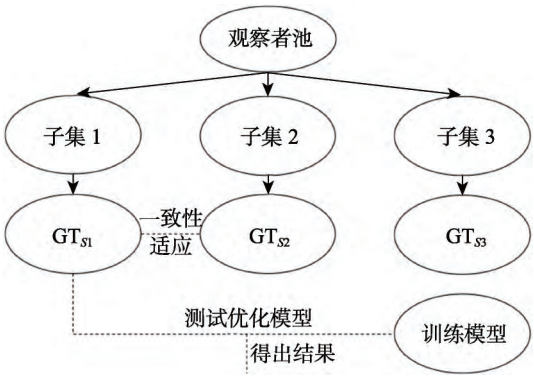

图 2　基于外部观察者的 3 个观察者群轮换的情感标注方法[54]

Malatesta 等[55]旨在解决对同一段动作视频序列中不同人的标注主观性太大而导致千差万别的问题，设计并进行两阶段的用户感知研究，目的是收集人类评分者对可用视频的注释数据，并将这些感知评分与所用计算机视觉组件的结果进行比较. 他们将标注过程分为实验标注阶段和全面标注阶段. 根据这一策略，使用模糊逻辑建立一个分析框架，将处理过的视觉特征映射到来自大量人群(在第 2 阶段雇佣 100 多名评分员)的可靠情感注释. 与之前的工作一样，该系统不是基于对被试者离散的主观表演，而是基于大量外部观众的感知来标记情感和表现力.

随着越来越多的工作开始使用非表演数据集，非表演数据的标注就显得更为重要，而当前对于姿态情感计算还大多围绕着识别算法这一层面. 尤其是当前姿态数据的获取成本在不断地降低，数据集在不断地扩大，非表演数据的人工标注需要过多的精力和成本，而准确的标注是情感识别算法等后续工作的基础，因此半自动或自动的非表演数据标注方法应该受到更多的关注.

5　姿态情感数据集

目前已有一些单一模态和多模态的姿态情感数据集，根据姿态数据集的数据类型、特征提取和标注结果，可以分为表演(act)数据集/非表演(non-act)数据集、3D 数据集/2D 数据集、动作(movement)数据集/姿势(posture)数据集和全身数据集/非全身数据集等类型. 现有的常用姿态情感数据集统计如表 1 所示.

表 1　现有常用姿态情感数据集统计

模态	文献	年份	数据集		特征	姿态
			类型	维数		
单一	[11]	2004	表演	3D	姿势	全身
	[56]	2016	表演	3D	动作	全身
	[57]	2014	表演	3D	动作	全身
	[58]	2014	表演	3D	动作	非全身
	[59]	2012	表演	2D	动作	全身
	[60]	2010	表演	2D	动作	全身
	[61]	2009	表演	3D	动作	全身
	[62]	2006	表演	3D	动作	全身
	[63]	2006	表演	2D	动作	半身
	[64]	2006	表演	3D	姿势	全身
多	[65]	2015	表演	2D	均有	均有
	[66]	2011	非表演	2D	动作	非全身
	[67]	2008	表演	3D	动作	非全身
	[68]	2007	均有	2D	动作	均有

目前的数据集尚存在一些问题，例如，有些数据来源于网络视频，质量参差不齐、数量较少等；有些数据产生于实验室环境，对于真实场景的情感计算不太适用；另外，数据集的内容大多要求被试根据剧本进行表演，不是自然的情感流露，与应用级别的真实情感识别尚有一定的距离. 在姿态情感数据集方面，还需要高质量、大数据、非表演、真实场景的优秀数据集来推动相关研究的进展. 此外，2D 数据集具有成本较低、计算更为简单快速的特点，可以应用在需要快速计算情感的场合；而 3D 数据集包含的信息更多，很多情感的表达也仅可以通过 3D 的微观变化来表现，加之目前 3D 人体姿态估计的相关成果飞速发展，所以未来应有更低成本、更快速、更准确的 3D 姿态获取方法，今后 3D 数据集将具有更广阔的应用空间.

6 基于姿态的情感识别算法

在适当的姿态情感数据集的基础上，常见的情感自动识别算法通常使用合适的分类器，对情感进行分类识别. 情感识别算法分为 2 种: (1) 传统的基于统计特征的一些分类器，如 SVM[69-70], *K* 最近邻算法(*K*-nearest neighbors, *K*NN)[71]和决策树(decision tree, DT)[72-73]. Aristidou 等[43]基于 LMA 系统，对指定的动作特征进行统计分析，使用 SVM 和 ET 算法进行情感分类计算. Gunes 等[73]使用 AdaBoost, C4.5 等 DT 算法，结合面部表情与静态的人体姿势数据进行情感分类. 但这些分类算法都使用姿势的数据集，使用的特征也是基于视频中单帧数据提取而来的. (2) 利用人工神经网络(artificial neural network, ANN)的算法[74-81]. 针对复杂的非表演动作数据集，更好的解决方案是使用较为复杂的计算模型，使用 ANN 可以得到更好的分类效果，如将时序模型应用于情感计算分类，比常规的分类算法性能更强. Savva 等[77]利用 Wii 游戏机进行用户动作数据的采集，并且使用循环神经网络(recurrent neural network, RNN)对时序特征进行处理和情感分类. 然而，在遇到长输入数据时 RNN 表现的性能并不理想，受限于梯度消失问题，有着较差的长期记忆能力. Sapiński 等[78]则使用长短期记忆神经网络(long-short term memory, LSTM)，采用门电路的方式有选择地记忆与遗忘输入信息，使用 RNN-LSTM 分类器有效地解决了问题，提升了识别准确率. Wang 等[79]利用循环门控单元(gated recurrent unit, GRU)，在 LSTM 的基础上减少门电路的个数和参数数量，简化了模型. Lefebvre 等[80]在双向长短期记忆神经网络(bidirectional long-short term memory, BLSTM)和双向循环门控单元(bidirectional gated recurrent unit, BGRU)中加入双向传播的思想，综合考虑了前后因素总结出分类结果，因此更具鲁棒性. 针对数据集特征维度高、特征空间复杂等特点，夏添等[81]利用宏观深度神经网络(macroscopic deep network, MAC-NN)和微观深度神经网络(microcosmic deep network, MIC-NN)，并将二者联合训练为融合神经网络(fusion network, FUS-NN)模型，全方面地捕捉肢体运动序列的信息. 其中，在 MAC-NN 部分，为了在保证模型性能的基础上减少参数个数、简化模型，使用 BGRU 代替 BLSTM; 在 MIC-NN 部分添加 Dropout 层，进一步加强了模型的抗过拟合能力.

对于较为简单的数据集，如静态的姿势数据集，用简单的分类器(如 SVM, *K*NN, DT 等)，即可得到较好的分类结果; 当数据集较复杂时，如基于时序的动作数据集，由于使用的特征维度高，特征空间复杂，就需要使用结构较为复杂的 ANN，如 RNN, LSTM, GRU 等神经网络. 如果数据集采用非表演数据集，动作数据会将更加杂乱，需要对传统神经网络进行修改，可以使用 BLSTM, BGRU, FUS-NN 等神经网络.

7 结 语

基于姿态的情感计算是情感计算领域的重要课题. 本文针对基于姿态的情感计算面临的几个关键问题，包括情感的心理学模型、人体姿态估计方法、姿态的情感特征提取方法、情感分类与标注方法、姿态情感数据集、基于姿态的情感识别算法进行综述，以独特的视角评述了相关工作的优劣，供相关领域科研人员参考.

参考文献(References):

[1] Ekman P. An argument for basic emotions[J]. Cognition & Emotion, 1992, 6(3/4): 169-200

[2] Barrett L F. Discrete emotions or dimensions? The role of valence focus and arousal focus[J]. Cognition & Emotion, 1998, 12(4): 579-599

[3] Noroozi F, Kaminska D, Corneanu C, *et al*. Survey on emotional body gesture recognition[J]. IEEE Transactions on Affective Computing, 2018: 1-1

[4] Glowinski D, Coll S Y, Baron N, *et al*. Body, space, and emotion: a perceptual study[J]. Human Technology: An Interdisciplinary Journal on Humans in ICT Environments, 2017: 13(1): 32-57

[5] Poria S, Cambria E, Bajpai R, *et al*. A review of affective computing: from unimodal analysis to multimodal fusion[J]. Information Fusion, 2017, 37: 98-125

[6] Kleinsmith A, Bianchi-Berthouze N. Affective body expression perception and recognition: asurvey[J]. IEEE Transactions on Affective Computing, 2012, 4(1): 15-33

[7] D'melloS K, Kory J. A review and meta-analysis of multimodal affect detection systems[J]. ACM Computing Surveys, 2015, 47(3): 1-36

[8] Ekman P. Differential communication of affect by head and body cues[J]. Journal of Personality and Social Psychology, 1965, 2(5): Article No.726

[9] Bull P E. Posture & gesture[M]. Amsterdam: Elsevier, 2016

[10] Pollick F E, Paterson H M, Bruderlin A, *et al*. Perceiving affect from arm movement[J]. Cognition, 2001, 82(2): B51-B61

[11] Coulson M. Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence[J].

Journal of Nonverbal Behavior, 2004, 28(2): 117-139
[12] Elman J L. Encyclopedia of language and Linguistics[M]. 2nd ed. Oxford: Elsevier, 2005
[13] Dou Jinhua, Qin Jingyan. Affective computing service platform of product appearance image based on deep learning [J]. Packaging Engineering, 2020, 41(6): 20-25(in Chinese)
(窦金花，覃京燕. 基于深度学习的产品外观意象情感计算服务平台研究[J]. 包装工程, 2020, 41(6): 20-25)
[14] Kołakowska A, Landowska A, Szwoch M, *et al.* Modeling emotions for affect-aware applications[M] //Information Systems Development and Applications. Gdańsk: Faculty of Management University of Gdańsk, 2015: 55-69
[15] Russell J A. A circumplex model of affect[J]. Journal of Personality and Social Psychology, 1980, 39(6): 1161
[16] Plutchik R. A general psychoevolutionary theory of emotion[M]. Cambridge: Academic Press, 1980: 3-33
[17] Mehrabian A. Analysis of the big five personality factors in terms of the PAD temperament model[J]. Australian Journal of Psychology, 1996, 48(2): 86-92
[18] Plutchik R. The nature of emotions: human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice[J]. American Scientist, 2001, 89(4): 344-350
[19] Fontaine J R, Scherer K R, Roesch E B. *et al*. The world of emotions is not two-dimensional[J]. Psychological Science, 2007, 18(12): 1050-1057
[20] Viola P, Jones M J, Snow D. Detecting pedestrians using patterns of motion and appearance[J]. International Journal of Computer Vision, 2005, 63(2): 153-161
[21] Dalal N, Triggs B, Schmid C. Human detection using oriented histograms of flow and appearance[C] //Proceedings of the 9th European Conference on Computer Vision. Heidelberg: Springer, 2006: 428-441
[22] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2008: 1-8
[23] Angelova A, Krizhevsky A, Vanhoucke V, *et al*. Real-time pedestrian detection with deep network cascades[C] //Proceedings of the British Machine Vision Conference. Durham: BMVA Press, 2015: 31.2-32.12
[24] Girshick R. Fast R-CNN[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2015: 1440-1448
[25] Zhang L, Lin L, Liang X, *et al*. Is faster R-CNN doing well for pedestrian detection?[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2016: 443-457
[26] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778
[27] Anbarjafari G, Izadpanahi S, Demirel H. Video resolution enhancement by using discrete and stationary wavelet transforms with illumination compensation[J]. Signal, Image and Video Processing, 2015, 9(1): 87-92
[28] Toshev A, Szegedy C. DeepPose: human pose estimation via deep neural networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 1653-1660
[29] Tompson J J, Jain A, LeCun Y, *et al*. Joint training of a convolutional network and a graphical model for human pose estimation[C] //Proceedings of Advances in Neural Information Processing Systems. Montreal: Curran Associates, 2014: 1799-1807
[30] Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2014: 580-587
[31] Ren S Q, He K M, Girshick R, *et al*. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149
[32] He K, Gkioxari G, Dollár P, *et al.* Mask R-CNN[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 2961-2969
[33] Cao Z, Simon T, Wei S E, *et al*. Realtime multi-person 2D pose estimation using part affinity fields[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 7291-7299
[34] Xia F, Wang P, Chen X, *et al*. Joint multi-person pose estimation and semantic part segmentation[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 6769-6778
[35] Insafutdinov E, Pishchulin L, Andres B, *et al*. DeeperCut: a deeper, stronger, and faster multi-person pose estimation model[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2016: 34-50
[36] Wang K, Lin L, Jiang C, *et al*. 3D human pose machines with self-supervised learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(5): 1069-1082
[37] Zhou X, Huang Q, Sun X, *et al*. Towards 3D human pose estimation in the wild: a weakly-supervised approach[C] //Proceedings of the IEEE International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2017: 398-407
[38] Alp Güler R, Neverova N, Kokkinos I. DensePose: dense human pose estimation in the wild[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 7297-7306
[39] Varol G, Ceylan D, Russell B, *et al*. BodyNet: volumetric inference of 3d human body shapes[C] //Proceedings of the European Conference on Computer Vision. Heidelberg: Springer, 2018: 20-36
[40] Kapur A, Kapur A, Virji-Babul N, *et al.* Gesture-based affective computing on motion capture data[C] //Proceedings of the 1st International Conference on Affective Computing and Intelligent. Heidelberg: Springer, 2005: 1-7
[41] Bianchi-Berthouze N, Kleinsmith A. A categorical approach to affective gesture recognition[J]. Connection Science, 2003, 15(4): 259-269
[42] Bernhardt D, Robinson P. Detecting affect from non-stylised body motions[C] //Proceedings of International Conference on Affective Computing and Intelligent Interaction. Heidelberg:

Springer, 2007: 59-70

[43] Aristidou A, Charalambous P, Chrysanthou Y. Emotion analysis and classification: understanding the performers' emotions using the LMA entities[J]. Computer Graphics Forum, 2015, 34(6): 262-276

[44] Luo Y, Ye J, Adams R B, *et al*. ARBEE: towards automated recognition of bodily expression of emotion in the wild[J]. International Journal of Computer Vision, 2019: 1-25

[45] Ajili I, Mallem M, Didier J Y. Human motions and emotions recognition inspired by LMA qualities[J]. The Visual Computer, 2019, 35(10): 1411-1426

[46] Freedman N. Hands, words, and mind: on the structuralization of body movements during discourse and the capacity for verbal representation[M]. Heidelberg: Springer, 1977: 109-132

[47] Siple P. Understanding language through sign language research[M]. Cambridge: Academic Press, 1978

[48] McNeill D. Hand and mind: what gesture reveals about thought[M]. Chicago: University of Chicago Press, 1992

[49] Gunes H, Piccardi M. Automatic temporal segment detection and affect recognition from face and body display[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 2008, 39(1): 64-84

[50] Glowinski D, Dael N, Camurri A, *et al*. Toward a minimal representation of affective gestures[J]. IEEE Transactions on Affective Computing, 2011, 2(2): 106-118

[51] Patwardhan A, Knapp G. Augmenting supervised emotion recognition with rule-based decision model[OL]. [2020-03-03]. https://arxiv.org/abs/1607.02660

[52] Xia S, Wang C, Chai J, *et al*. Realtime style transfer for unlabeled heterogeneous human motion[J]. ACM Transactions on Graphics, 2015, 34(4): 1-10

[53] Aristidou A, Zeng Q, Stavrakis E, *et al*. Emotion control of unstructured dance movements[C] //Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation. New York: ACM press, 2017: 1-10

[54] Kleinsmith A, Bianchi-Berthouze N, Steed A. Automatic recognition of non-acted affective postures[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B, 2011, 41(4): 1027-1038

[55] Malatesta L, Asteriadis S, Caridakis G, *et al*. Associating gesture expressivity with affective representations[J]. Engineering Applications of Artificial Intelligence, 2016, 51: 124-135

[56] Senecal S, Cuel L, Aristidou A, *et al*. Continuous body emotion recognition system during theater performances[J]. Computer Animation and Virtual Worlds, 2016, 27(3/4): 311-320

[57] Fourati N, Pelachaud C. Emilya: emotional body expression in daily actions database[C] //Proceedings of the 9th International Conference on Language Resources and Evaluation. Lisboa: European Language Resources Association, 2014: 3486-3493

[58] Volkova E P, Mohler B J, Dodds T J, *et al*. Emotion categorization of body expressions in narrative scenarios[J]. Frontiers in Psychology, 2014, 5: 623

[59] Dael N, Mortillaro M, Scherer K R. The body action and posture coding system (BAP): development and reliability[J]. Journal of Nonverbal Behavior, 2012, 36(2): 97-121

[60] Scherer K R, Bänziger T, Roesch E B. A blueprint for affective computing: a sourcebook and manual[M]. Oxford: Oxford University Press, 2010: 271-294

[61] Karg M, Jenke R, Seiberl W, *et al*. A comparison of PCA, KPCA and LDA for feature extraction to recognize affect in gait kinematics[C] //Proceeding of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Los Alamitos: IEEE Computer Society Press, 2009: 1-6

[62] Ma Y, Paterson H M, Pollick F E. A motion capture library for the study of identity, gender, and emotion perception from biological motion[J]. Behavior Research Methods, 2006, 38(1): 134-141

[63] Gunes H, Piccardi M. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior[C] //Proceedings of the 18th International Conference on Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2006: 1148-1153

[64] Kleinsmith A, de Silva P R, Bianchi-Berthouze N. Cross-cultural differences in recognizing affect from body posture[J]. Interacting with Computers, 2006, 18(6): 1371-1389

[65] Gavrilescu M. Recognizing emotions from videos by studying facial expressions, body postures and hand gestures[C] //Proceedings of the 23rd Telecommunications Forum Telfor. Los Alamitos: IEEE Computer Society Press, 2015: 720-723

[66] 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, 2011, 3(1): 5-17

[67] Busso C, Bulut M, Lee C C, *et al*. IEMOCAP: interactive emotional dyadic motion capture database[J]. Language Resources and Evaluation, 2008, 42(4): ArticleNo.335

[68] Douglas-Cowie E, Cowie R, Sneddon I, *et al*. The HUMAINE database: addressing the collection and annotation of naturalistic and induced emotional data[C] //Proceedings of International Conference on Affective Computing and Intelligent Interaction. Heidelberg: Springer, 2007: 488-500

[69] George F P, Shaikat I M, Ferdawoos P S, *et al*. Recognition of emotional states using EEG signals based on time-frequency analysis and SVM classifier[J]. International Journal of Electrical & Computer Engineering, 2019, 9(2): 1012

[70] Song N, Yang H, Wu P. A gesture-to-emotional speech conversion by combining gesture recognition and facial expression recognition[C] //Proceedings of the 1st Asian Conference on Affective Computing and Intelligent Interaction. Los Alamitos: IEEE Computer Society Press, 2018: 1-6

[71] Yan J, Lu G, Bai X, *et al*. A novel supervised bimodal emotion recognition approach based on facial expression and body gesture[J]. IEICE Transactions on Fundamentals of Electronics, 2018, E101. A(11): 2003-2006

[72] Noroozi F, Kaminska D, Sapinski T, *et al*. Supervised vocal-based emotion recognition using multiclass support vector machine, random forests, and Adaboost[J]. Journal of the Audio Engineering Society, 2017, 65(7/8): 562-572

[73] Gunes H, Piccardi M. Fusing face and body gesture for machine recognition of emotions[C] //Proceedings of the 14th IEEE International Workshop on Robot and Human Interactive Communication. Los Alamitos: IEEE Computer Society Press, 2005: 306-311

[74] Ly S T, Lee G S, Kim S H, *et al*. Gesture-based emotion recog-

nition by 3D-CNN and LSTM with keyframes selection[J]. International Journal of Contents, 2019, 15(4): 59-64

[75] Tzirakis P, Trigeorgis G, Nicolaou M A, *et al*. End-to-end multimodal emotion recognition using deep neural networks[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1301-1309

[76] Devineau G, Moutarde F, Xi W, *et al*. Deep learning for hand gesture recognition on skeletal data[C] //Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 106-113

[77] Savva N, Scarinzi A, Bianchi-Berthouze N. Continuous recognition of player's affective continuous expression as dynamic quality of aesthetic experience[J]. IEEE Transactions on Computational Intelligence and AI in Games, 2012, 4(3): 199-212

[78] Sapiński T, Kamińska D, Pelikant A, *et al*. Emotion recognition from skeletal movements[J]. Entropy, 2019, 21(7): Article No.646

[79] Wang W, Yang N, Wei F, *et al*. Gated self-matching networks for reading comprehension and question answering[C] //Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics.Vancouver: Association for Computational Linguistics, 2017: 189-198

[80] Lefebvre G, Berlemont S, Mamalet F, *et al*. BLSTM-RNN based 3D gesture classification[C] //Proceedings of International Conference on Artificial Neural Networks. Heidelberg: Springer, 2013: 381-388

[81] Xia Tian, Zhang Yifeng, Liu Yuan. Landmark-based facial expression recognition by joint training of multiple networks[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(4): 552-559(in Chinese)
(夏添，张毅锋，刘袁．基于特征点与多网络联合训练的表情识别[J].计算机辅助设计与图形学学报，2019，31(4): 552-559)