基于多模态信息的情感计算综述

彭小江[1,2]

(1. 衡阳师范学院 计算机科学与技术学院，湖南 衡阳 421002；
2. 中国科学院 深圳先进技术研究院，广东 深圳 518055)

摘 要：情感计算是计算机视觉和人工智能领域重要方向，在服务机器人、法政、娱乐等方面有较大应用价值。人类情感的外在表露和内在生物信息是情感计算重要依据。该文分别综述了基于外在表露的视觉、音频、姿态和言词四个模态情感计算和基于脑部生物信息的情感计算方法进行全面综述。揭示了目前多模态情感计算的显著问题和发展方向。

关键词：情感计算；计算机视觉；多模态；人脸表情识别

中图分类号：TN34 文献标志码：A 文章编号：1673-0313(2018)03-0031-06

DOI:10.13914/j.cnki.cn43-1453/z.2018.03.008

情感是生物神经系统对外界价值关系产生的主观反映，也是生物智能的重要组成部分[1]。在所有的情感生物中，人类的情感最具表达力、最复杂、社会 性也最强。在人际交往与人们的日常生活过程中，情感的表达是其中不可或缺的重要部分，其传递的信息非常丰富，我们可以通过人们的语音的变化、语言内容、脸部表情及肢体动作姿态等来判别人们当前的情感状态。

情感计算通常指利用机器设备对人类情感进行分类识别、解释、模仿，这些任务可以在各种表现形式如人脸图像视频、音频、生物信号上进行。情感计算作为一门融合视觉信号处理、心理学生理学、模式识别和人工智能等领域的交叉学科，目前在法政(如微表情测谎)、娱乐、安全、服务机器人等领域有较大应用价值。最具代表的案例是 Apple 公司最近收购的一个人脸表情公司 Emotinet，其旨在健康、照片管理、观众情感反馈等领域发挥作用。2006 年，Minsky 在其著作《情感机器—The Emotion Machine》[2]一书中指出人工智能＝认知智能＋情感智能。现阶段人工智能在国内外的研究可以说是异常火热，但是绝大部分人工智能工作都停留在认知智能层面。另外，作为人们在现代社会进行信息交流的重要途径，互联网俨然已经成为涉及广泛主题的意见和情绪资源库。在发帖评论、浏览行为以及分享的媒体对象中处处可见操作、发布者的情感信息。对这些文字信息的分析称为意见挖掘、情感计算或者情感分析，它在智能对话、舆情发现、社交等诸多领域中都起着不可或缺的作用。

由于情感计算的重要性，有关情感计算的研究长期以来受到研究人员的广泛关注，并在表情识别、语音情感分类与合成、表情生成等领域开展了大量的研究工作，取得了重要的进展。最早的在情感方面进行的有关研究可以追溯上世纪六七十年代。1971 年，Ekman 和 Friesen 提出，人类的情感如果强制地分为某些类别的话，可以分为：高兴，悲伤，愤怒，厌恶，吃惊和恐惧六种[3]，并进一步从解剖学角度提出了人脸表情动作编码系统 FACS[4]。目前大多数研究都是基于这 6 类表情外加“自然”类进行分析，如图 1(b)。

收稿日期：2018-03-08

基金项目：国家自然科学基金项目(61502152)

作者简介：彭小江(1987-)，男，江西莲花人，副研究员，博士，主要研究方向为人工智能和计算机视觉。

在多模态情感分析方面，目前大多数都是基于视觉、语音、文本之类的其中一两种模态进行研究。由于人类情感的表达方式是多种多样的，这种基于单双模态信息的分析是不完善且不丰富的，对于情感的准确判别是远远不够的，比如观察者在给图1(a)表情标签时很难判断是悲伤还是愤怒。其实，人类表现各个模态的情感信息之间是相辅相成，缺一不可的[5]。

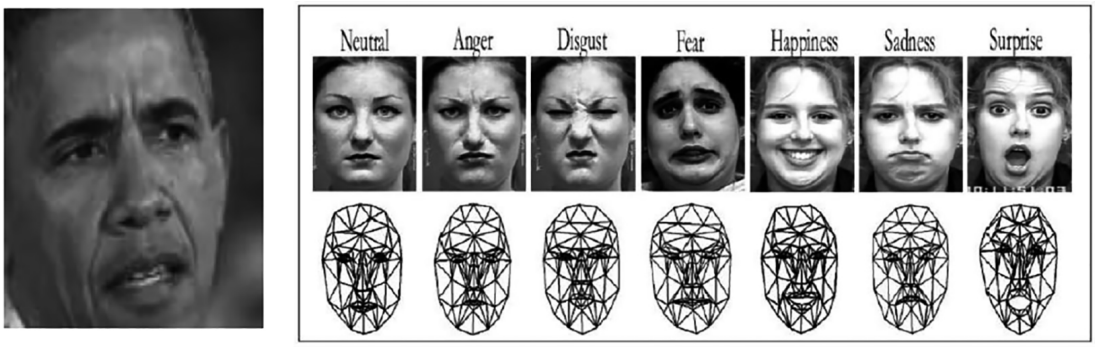

图1　(a)视觉表情样例。(b)人类的基本情感及人脸表情动作编码系统(FACS)[4]

本文将从情感计算所利用的多模态信息角度对基于视频、音频、语言内容(即文本)、姿态以及脑部信号(如：EEG，fMRI—功能性核磁共振图)的情感计算方法进行全面综述，揭示目前多模态情感分析的问题和方向。

1　基于视觉的情感分析

视觉是人类感知外界的主要途径，在基于视觉情感分析方面，面部表情无疑是情感最重要的表现模态。面部表情识别本质上是归属机器学习及模式识别等相关领域，机器学习及模式识别的算法可以应用于面部表情识别中。

从现有研究文献算法来看，研究人员往往将面部表情识别的算法实现划分为人脸检测及预处理、表情特征提取、表情特征分类三大步骤[6]。

在人脸检测方面，传统比较典型的方法有基于Adaboost和Haar特征的VJ检测器[7]、基于LBP[8]特征的人脸检测器[9]、基于HOG[10]特征和可形变模型(DPM)的检测器[11]等。这些人脸检测方法在可控实验环境下性能表现优异，但是在实际室外场景中性能下降非常严重，主要原因在于手工设计的浅层特征很难对实际场景的各种光照、遮挡、姿态等因素建模。近年来，由于深度学习在视觉任务的优异表现，基于深度学习的人脸检测成为主流方法。从深度网络结构来看，基于深度学习的人脸检测方法可以分为三类：级联CNN方法[12-14]，基于proposal的两步法[15-16]，单步法[17-20]。级联CNN人脸检测方法主要是吸取adaboost的思想从粗到精对人脸进行检测 。基于proposal的两步法主要是沿袭物体检测中R-CNN思想，先生成proposal然后对proposal进行人脸分类，目前使用较多的是基于faster R-CNN[15]框架。单步法主要是沿袭物体检测中SSD[18]、YOLO[20]以及faster R-CNN里的RPN方法，该方法在速度方面比较有优势。人脸检测后，通常的预处理为人脸对齐操作以消除图像尺寸和人脸部位偏移带来的影响。人脸对齐方法有主动外观模型(AAM)法[21]、级联回归方法[22-23]和基于深度学习的回归方法[14,24-25]等。目前，人脸检测和对齐方面使用最多的是申请人所在研究所提出的MTCNN[14]方法，该方法能够同时完成检测和对齐，操作方便。

在视觉表情特征提取方面，早期都是使用传统的手工设计特征，如文献[26]使用Gabor滤波器提取表情特征，文献[27]使用LBP特征，文献[28]使用HOG特征等。这些手工设计的特征通常在现实室外场景中缺乏对光照、姿态、清晰度等因素足够的泛化建模能力，实际场景应用效果较差。深度神经网络能够从大量数据中学习出比较适合任务的特征.近年来，由于深度神经网络在视觉特征学习方面的优异表现，表情特征的提取大部分采用的是深度神经网络方法[29-32]。比如文章作者在2017年ICMIEmotiW竞赛的群体情感分类中，使用人脸识别深度网络进行微调出多通道深度神经网络获得冠军[33]。

表情特征提取后需要对特征进行分类识别，常用的模式识别分类方法比如支持向量机(SVM)、朴素贝叶斯(Naïve Bayes)、随机森林等在早期手工设计特征时代使用广泛。目前，在深度学习全面普及

时代，一般直接使用 Softmax 来训练分类器，通常该分类器的训练和特征学习同时进行。

2 基于语音的情感分析

基于语音的情感分析是从说话人语音中提取情感特征，然后对相应特征进行识别分类等过程。语音情感数据库是语音情感识别的基础，语音情感库质量的好坏对语音情感识别的效果起到重要作用。因此建立一个高质量的语音数据库是进行语音情感识别的第一关健步。目前，许多国家建立了不同语种的情感语料库，如：德语德国柏林 EMODB 情感语料库[34]、Semaine 数据库[35]、VAM 数据库[36]等不同语言的语音情感库。

语音情感识别中特征提取是后续处理的基础，即是指从语音信号中提取可以表征语音情感的特征。国内外研究者们从语音学和心理学方面对情感特征进行了大量的研究。一般提取的情感特征主要分为韵律特征、音质特征和谱特征。

韵律特征被认为是主要的语音情感参数，反映的是" 唤醒度" 信息。人类语言的时常、语调、轻重各不相同，这些韵律特征的变化构成了美妙的语言。常见韵律特征有基频、时长、能量等，韵律特征的统计特征分析着眼于整体语音，反映出一段时间之上韵律参数的变化规律。在 Basque 情感数据上，Luengo 等人[37]研究发现能量的平均值、方差、能量对数和基频对数的动态变化范围、基频均值和对数斜交共 6 个特征是最具有情感区分能力的特征。

此外，音质将征与情感的关联性也很大。音质特征主要指语谱和音色方面的特性，取决于说话的音波形式，由说话人喉部及以上器官决定。情感不同，同一人的音质也会有差异。共振峰参数是用在语音情感识别中最主要的音质特征，Tato 等人的探索说明音质特征对于辨别像生气和高兴这样的情感有不错的成效[38]。

此外，谱特征参数也是反映语音情感状态的主要参数。谱特征反映信号的频域持性。Nwe 研究出不同频谱区间的频谱能量分布和情感状态有很大相关性，指出高频段高兴情感能量较高，悲伤情感能量很低[39]。线性预测倒谱系数（LPCC）、梅尔频率倒谱系数（MFCC）等参数也因此被广泛用于语音情感识别中，且这些参数一般被认为鲁棒性较好。

语音特征提取好后，情感识别算法可以使用常规的 SVM 算法、隐马尔科夫法、高斯混合模型等方法。

近年来，深度学习在语音情感分析中也有成功的应用。比如文献[40，41]直接在语音频谱图上直接使用简单的深度卷积神经网络，也取得了不错的效果。

3 基于文本的情感分析

文本情感分析是自然语言处理（NLP）领域一个重要方向指利用自然语言处理和文本挖掘技术，对带有情感色彩的主观性文本进行分析、处理和抽取的过程[42]。

近年来，情感分析得到了越来越多研究机构和学者的关注，在 SIGIR、ACL、WWW、CIKM、WSDM 等著名国际会议上，针对这一问题的研究成果层出不穷，国内外研究机构组织了众多相关评测来推动情感分析技术的发展。

由国际文本检索会议 TREC 针对英文文本观点检索任务的博客检索任务（Blog Track），篇章情感分类任务，以及其他一些有趣的情感分析任务；由日本国立信 息学研究所主办的搜索引擎评价国际会议 NTCIR（NII Test Collection for IR Systems）针对日、韩、英、中文文本 的情感分类以及观点持有者抽取任务。由中文信息学 会信息检索委员会主办的每年一次的中文倾向性分析 评测 COAE（Chinese Opinion Analysis Evaluation）已举办了 5 届，在关注情感词语和观点句子的抽取以及倾 向性识别的基础上重点对于否定句、比较句以及微博观点句进行评测[43]。

众多研究机构的评测推动了情感分析研究的发展，出现了很多有代表性的情感分析语料库资源，文献[44]对语料库构建进行了详细阐述，如康奈尔影 评数据集（Cornell Movie-Review Datasets），多视角问答（Multiple-Perspective Question Answering，MPQA）语料库，TREC 测试集，NTCIR 多语言语料库（NTCIRmultilingual corpus），中文 COAE 语料库等。

早期的文本情感分析借助包含积极和消极词的字典。每个词在情感上都有分值，通常 +1 代表积极情绪，-1 代表消极。接着，我们简单累加句子中所有词的情感分值来计算最终的总分。显而易见，这样的做法存在许多缺陷，最重要的就是忽略了语境（context）和邻近的词。另一个常见的做法是以文本进行“词袋（bag of words）”建模。经过 BOW 后的特征向量可以作为诸如逻辑回归、支持向量机的机器学习算法的输入，以此来进行分类。此方法和早期方法相比有了明显的进步，但依然忽

略了语境，而且数据的大小会随着词汇的大小增加。目前，由 Google 提出来的基于深度学习的 Word2Vec 和 Doc2Vec 方法是非常有效的方法。

4 基于脑部信息和多模态信息的情感分析

除了上述 3 节中的面部视觉、语音、文本（即语言内容）模态，姿态行为、脑电（EEG）及核磁共振图（fMRI）通常也和情感密切相关。目前，含有多模态的情感分析数据库主要有：（1）DEAP 数据集[45]—包含无声视频和 EEG 信号，（2）MEC[46] 和 AFEW 数据库—包含语音的电影视频片段，（3）GEMEP（the Geneva Multimodal Emotion Portrayal）数据集[47]—包含语音、面部表情和姿态 3 个模态。

由于信号差异比较大，目前多模态融合大多数采用两类融合方法：分类器前情感特征表达融合和分类器分数融合。文献[48]提出一种 ModDrop 的多模态融合方法，实验证明比多模态特征直接拼接效果更好。文献[49]在 AFEW 7.0 音视频情感分类任务中评估了 2 种分数融合方法和包括 ModDrop 在内的 3 种多模态特征融合方法。其他常用的双模态特征融合方法如典型相关分析（Canonical Correlation Analysis，CCA）[50] 衍生出了多重典型相关分析（Multiple-Set Canonical Correlation Analysis，MCCA）[51] 等。

5 总 结

情感分析已经受到研究人员的广泛关注，并开展了大量的工作，在基于面部表情、语音、文本和生物信号的情感识别方面均取得了一系列的重要进展。但受制于情感表达模态的复杂多变等特性，目前的情感分析往往主要集中在单/双模态、离散的基本情感类别等方面，缺乏更多模态及连续动态情感分析的系统研究。这种现状导致了目前服务机器人在人机交互环境下，情感表达单一死板，缺乏多通道多模态协同的自然表达模式，影响了用户对情感的感知和理解。针对情感生物信息如脑部 fMRI 和 EEG 信号，大部分研究都是手工设计一些特征并利用传统的机器学习方法，效果并不理想。未来情感计算可能的发展方向有：1）建立大规模视觉、音频 、姿态和言词四模态现实场景数据，建立视频、fMRI、EEG 组成的多模态生物信息情感数据，2）基于深度学习的脑部情感信息分析，3）基于强人工智能的多模态外在情感生成以提供更真实的人机交互。

参考文献：

[1] 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.

[2] MINSKY M. The Emotion Machine [M]. New York: Pantheon, 2006.

[3] EKMAN P, Friesen W V. Constants across cultures in the face and emotion[J]. Journal of Personality & Social Psychology, 1971, 17(2): 124-129.

[4] EKMAN P, FRIESEN W. Facial action coding system: a technique for the measurement of Facial Movement [M]. Palo Alto: Consulting Psychologists Press, 1978.

[5] SOLEYMANI M, Pantic M, Pun T. Multimodal emotion recognition in response to videos[J]. IEEE Transactions on Affective Computing (TAC), 2012, 3(2): 211-223.

[6] PANTIC M, ROTHKRANTZ L J M. Automatic analysis of facial expressions: the state of the art [J]. IEEE Transactions on Pattern Analysis Machine Intelligence (TPAMI), 2000, 22(12): 1424-1445.

[7] VIOLA P, JONES M J. Robust Real-Time Object Detection [J]. International Journal of Computer Vision (IJCV), 2004, 57(2): 137-154.

[8] OJALA T, PIETIKäINEN M, HARWOOD D A Comparative Study of Texture Measures with Classification Based on Feature Distributions[J]. Pattern Recognition, 1996, 29(1): 51-59.

[9] ZHANG L, CHU R, XIANG S, et al. Face detection based on multi-block lbp representation [C]. International Conference on Biometrics, 2007: 11-18.

[10] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]. Computer Vision and Pattern Recognition (CVPR), 2005: 886-893.

[11] YAN J, LEI Z, WEN L, et al. The fastest deformable part model for object detection [C]. CVPR, 2014: 2497-2504.

[12] LI H, LIN Z, SHEN X, et al. A convolutional neural network cascade for face detection [C]. CVPR, 2015: 5325-5334.

[13] YANG S, LUO P, LOY C.-C., et al. Tang. From facial parts responses to face detection: A deep learning approach [C]. IEEE International Conference on Computer Vision (ICCV), 2015: 3676-3684.

[14] ZHANG K, ZHANG Z, LI Z, ea al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. Signal Processing Letters, 2016, 23(10): 1499-1503.

[15] REN S, HE K, GIRSHICK R, et al. Faster r-cnn: To-

wards real-time object detection with region proposal networks[J].TPAMI,2017,39(6): 1137-1149.

[16] JIANG H,LEARNED-MILLER E.Face Detection with the Faster R-CNN[C].IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2017:650-657.

[17] NAJIBI M,SAMANGOUEI P,CHELLAPPA R,et al. Ssh: single stage headless face detector[C]. ICCV, 2017:4875-4884.

[18] LIU W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector [C]. European Conference on Computer Vision (ECCV),2016:21-37.

[19] S F ZHANG,X Y ZHU,Z LEI,et al.S3FD: single shot Scale-invariant face detector[C].ICCV,2017:192-201.

[20] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]. CVPR,2016:779-788.

[21] COOTES T.F,EDWARDS G.J.,TAYLOR C.J.Active appearance models[J].TPAMI,2001,23(6): 681-685.

[22] KAZEMI V,SULLIVAN J.One millisecond face alignment with an ensemble of regression trees[C].CVPR, 2014:1867-1874.

[23] REN S,CAO X,WEI Y,et al.Face alignment at 3000 fps via regressing local binary features[C]. CVPR, 2014:1685-1692.

[24] SUN Y, WANG X, TANG X. Deep convolutional network cascade for facial point detection[C].CVPR, 2013:3476-3483.

[25] ZHANG Z,LUO P,LOY C C,et al.Facial landmark detection by deep multi-task learning[C].ECCV,2014: 94-108.

[26] LIU C,WECHSLER H.Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition[J].IEEE Transaction on Image Processing (TIP),2002,11(4): 467-476.

[27] SHAN C,GONG S,MCOWAN P W.Facial expression recognition based on local binary patterns: A comprehensive study[J].Image and Vision Computing,2009, 27(6):803-816.

[28] MAVADATI S M,MAHOOR M H,BARTLETT K, et al.Disfa: A spontaneous facial action intensity database[J].TAC,2013,4(2): 151-160.

[29] MOLLAHOSSEINI A,HASANI B,SALVADOR M J, et al.Facial expression recognition from world wild web [C].CVPR Workshops,2016:1509-1516.

[30] MOLLAHOSSEINI A, CHAN D, MAHOOR M H. Going deeper in facial expression recognition using deep neural networks[C].IEEE Winter Conference on Applications of Computer Vision (WACV), 2016: 1-10.

[31] HE L, JIANG D, YANG L, et al. Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks[C]. ACM International Workshop on Audio/Visual Emotion Challenge,2015:73-80.

[32] FAN Y,LU X,LI D,et al.Video-based emotion recognition using cnn-rnn and c3d hybrid networks[C]. ACM International Conference on Multimodal Interaction (ICMI),2016:445-450.

[33] TAN L, ZHANG K, WANG K, et al. Group emotion recognition with individual facial emotion CNNs and global image based CNNs[C].ICMI,2017:549-552.

[34] BURKHARDT F,PAESCHKE A,ROLFES M,et al.A database of German emotional speech[C]. INTERSPEECH,2005.

[35] MCKEOWN G, VALSTAR M F, Cowie R, et al. The SEMAINE corpus of emotionally coloured character interactions[C]. International Conference on Multimedia and Expo (ICME),2010:1079-1084.

[36] GRIMM M, KROSCHEL K, NARAYANAN S. The vera am Mittag German audio-visual emotional speech database[C].ICME,2008:865-868.

[37] LUENGO I ,NAVAS E,HERNáEZ I,et al.Automatic emotion recognition using prosodic parameters[C].INTERSPEECH,2015:493-496.

[38] SCHULLER B, BATLINER A, STEIDL S, et al. Emotion recognition from speech: Putting ASR in the loop[C].IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2009: 4585-4588.

[39] NWE T L,FOO S W,SILVA L C D.Speech emotion recognition using hidden Markov models[J]. Speech Communication,2003,41(4):603-623.

[40] BADSHAH A M,AHMAD J,RAHIM N,et al.Speech emotion recognition from spectrograms with deep convolutional neural network[C].International Conference on Platform Technology and Service,2017:1-5.

[41] ZHANG S, ZHANG S, HUANG T, et al. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching [J].IEEE Transactions on Multimedia (TMM),2018, 20(6):1576-1590.

[42] PANG B, LEE L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval,2008,2(2): 1-135.

[43] **杨立公,朱俭,汤世平.文本情感分析综述[J].计算机应**

用,2013,33(6):1574-1607.
[44] LIU B,ZHANG L.A survey of opinion mining and sentiment analysis[J].Springer US:Mining Text Data,2012:415-463.
[45] KOELSTRA S,MUHL C,SOLEYMANI M,et al.Deap:A database for emotion analysis;using physiological signals[J].IEEE Transactions on Affective Computing,2012,3(1):18-31.
[46] LI Y,JAO J,SCHVUER B,et al.MEC 2016:the multimodal emotion recognition challenge of CCPR[C].Pattern Recognition.Springer sigapore,2016:667-678.
[47] BäNZIGER T,SCHERER K R.Introducing the geneva multimodal emotion portrayal (GEMEP) corpus[M].Oxford:Oxford University Press,2010.
[48] Neverova N,Wolf C,Taylor G,et al.Moddrop:adaptive multi-modal gesture recognition[J].TPAMI,2016,38(8):1692-1706.
[49] VIELZEUF V,PATEUX S,JURIE F.Temporal multimodal fusion for video emotion classification in the wild[C].ICMI,2017:569-576.
[50] WEENINK D.Canonical correlation analysis inference for functional data with applications[M].Springer New York,2012.
[51] TAKANE Y,HWANG H,ABDI H.Regularized multiple-set canonical correlation analysis[J].Psychometrika,2008,73(4):753-761.

(编校 陈志阳)

Multi-modal Affective Computing:A Comprehensive Survey

PENG Xiao-jiang[1,2]

(1. College of Computer Science and Technology,Hengyang Normal University,Hengyang Hunan 421002,China;
2. Shenzhen Institutes of Advanced Technology,Chinese Academy of Sciences,Shenzhen Guangdong 518055,China)

Abstract:Affective computing is an important research area in computer vision andartificial intelligence due to its wider range of potential applications such asservice robot,forensic community,entertainments,etc.The extrinsic behavior andintrinsic biological information are two important foundations for emotioncomputing.This paper presents a comprehensive survey on affective computing methods based on extrinsic behavior information such as vision,audio,pose,and context information,and based on brrain biological information.In addition,it analyzes the problems of current multi-modal affective computing and give an insight on the future direction.

Key words:affective computing;computer vision;multi-modal;facial expression recognition