基于深度学习的表情动作单元识别综述

邵志文[1,2],周 勇[1,2],谭 鑫[3],马利庄[3,4],刘 兵[1,2],姚 睿[1,2]
(1. 中国矿业大学计算机科学与技术学院,江苏徐州 221116; 2. 矿山数字化教育部工程研究中心,江苏徐州 221116;
3. 上海交通大学计算机科学与工程系,上海 200240; 4. 华东师范大学计算机科学与技术学院,上海 200062)

摘 要: 基于深度学习的表情动作单元识别是计算机视觉与情感计算领域的热点课题. 每个动作单元描述了一种人脸局部表情动作,其组合可定量地表示任意表情. 当前动作单元识别主要面临标签稀缺、特征难捕捉和标签不均衡3个挑战因素. 基于此,本文将已有的研究分为基于迁移学习、基于区域学习和基于关联学习的方法,对各类代表性方法进行评述和总结. 最后,本文对不同方法进行了比较和分析,并在此基础上探讨了未来动作单元识别的研究方向.

关键词: 表情动作单元识别; 标签稀缺性; 特征难捕捉性; 标签不均衡性; 迁移学习; 区域学习; 关联学习
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2022)08-2003-15
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210639

Survey of Expression Action Unit Recognition Based on Deep Learning

SHAO Zhi-wen[1,2], ZHOU Yong[1,2], TAN Xin[3], MA Li-zhuang[3,4], LIU Bing[1,2], YAO Rui[1,2]
(1. *School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, Jiangsu* 221116, *China*;
2. *Engineering Research Center of Mine Digitization, Ministry of Education of the People's Republic of China, Xuzhou, Jiangsu* 221116, *China*;
3. *Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai* 200240, *China*;
4. *School of Computer Science and Technology, East China Normal University, Shanghai* 200062, *China*)

Abstract: Expression action unit(AU) recognition based on deep learning is a hot topic in the fields of computer vision and affective computing. Each AU describes a facial local expression action, and the combinations of AUs can quantitatively represent any expression. Current AU recognition mainly faces three challenging factors, scarcity of labels, difficulty of feature capture, and imbalance of labels. On this basis, this paper categorizes the existing researches into transfer learning based, region learning based, and relation learning based methods, and comments and summarizes each category of representative methods. Finally, this paper compares and analyzes different methods, and further discusses the future research directions of AU recognition.

Key words: expression action unit recognition; scarcity of labels; difficulty of feature capture; imbalance of labels; transfer learning; region learning; relation learning

1 引言

近年来,“以人为本,服务于人”得到人工智能研究越来越广泛的关注,面部表情是人类情感最自然和直接的表现方式,对其的分析和识别[1~3]是计算机视觉与情感计算领域的热门研究方向,在医疗健康[4]、公共安全[5]等领域具有广泛的应用前景. 由于人们在日常生活中较少表现大幅度的面部动作,更多是通过局部细微表情来表达情感,如悲伤时眉毛下垂、惊讶时张开嘴,因此许多工作关注对局部表情动作而不仅仅是整体表情的识别.

人脸动作编码系统(Facial Action Coding System, FACS)[6,7]定义了几十个表情动作单元(Action Unit, AU),是目前描述人脸局部细微表情最全面和客观的系统之一. 如图1所示,快乐、悲伤、惊讶等整体表情被定量地解析为多个AU的组合,每个AU是一个基本面部动作,与一或多个人脸局部肌肉动作有关. 在一个人脸表情中,可能只出现一个AU,也可能同时出现多个

收稿日期:2021-05-18;修回日期:2021-09-16;责任编辑:王天慧
基金项目:国家自然科学基金(No.62106268);江苏省自然科学基金(No.BK20201346);江苏省“六大人才高峰”项目(No.2015-DZXX-010);中央高校基本科研基金(No.2021QN1072)

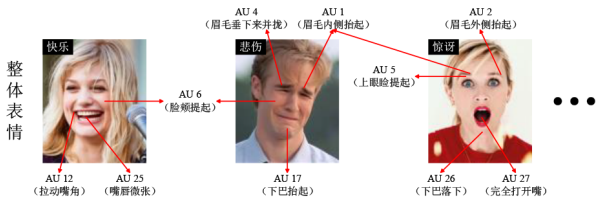


图1 整体表情与AU的关系示例

AU. 虽然FACS只定义了几十个AU,但是每个AU具有从低到高的多个强度级别,因而AU的组合可表示7 000种以上真实存在的表情[8],满足了精细刻画表情的需要.

深度学习在计算机视觉的各个领域都获得了巨大成功,近些年越来越多的人脸表情识别工作采用深度神经网络,基于其强大的特征提取能力,显著提升了表情识别的精度. 然而早期的人脸表情识别综述[9~11]主要介绍传统的非深度学习方法,由于这类方法采用人工设计的特征,限制了表情识别的性能. 近年来,Corneanu等人[12]总结了基于RGB图像、3D、热成像或多模态数据的人脸表情识别工作,Li等人[13]将讨论范围限定在基于深度学习的方法. 然而,上述综述仅关注识别整体表情的工作,忽视了表情AU识别. 另外,贲晛烨等人[14]和徐峰等人[15]对微表情识别进行了综述,但也没有关注AU识别. Martinez等人[16]和Zhi等人[17]虽然详细回顾了AU识别工作,但其中大部分仍是基于传统的非深度学习方法.

鉴于此,本文主要讨论基于深度学习的表情AU识别工作,对这一领域的代表性方法进行分类、评述和总结,弥补现有人脸表情识别综述的不足. 本文接下来首先介绍AU识别的问题定义、挑战和评测数据集,然后从迁移学习、区域学习和关联学习3个角度对已有工作进行概述,之后将一些主流AU识别方法的性能进行了比较,最后探讨了AU识别未来的研究趋势.

2 问题定义、挑战和评测数据集

2.1 AU的定义

人脸表情出现时,一些局部区域会发生肌肉动作. 人脸动作编码系统(FACS)[6,7]基于人脸解剖学所划分的局部肌肉,定义了一个基本面部动作即动作单元(AU)的集合. 每个AU涉及一个或多个局部肌肉,具有0,1,2,3,4,5这6个强度级别,其中0表示不出现而5则表示出现的强度最大,因而可以客观且定量地描述人脸精细表情. 图2展示了常见的27个AU的示例图片及定义,其中9个AU出现在上半脸,18个AU出现在下半脸. 可以发现,每个AU都是一种局部的面部动作,刻画了细微表情. 形式化地,任一人脸表情可以由这些AU出现的强度所构成的向量$\boldsymbol{p}^{(\mathrm{int})}=(p_1^{(\mathrm{int})},p_2^{(\mathrm{int})},\cdots,p_m^{(\mathrm{int})})$来表示,其中未出现的AU的强度即为0.

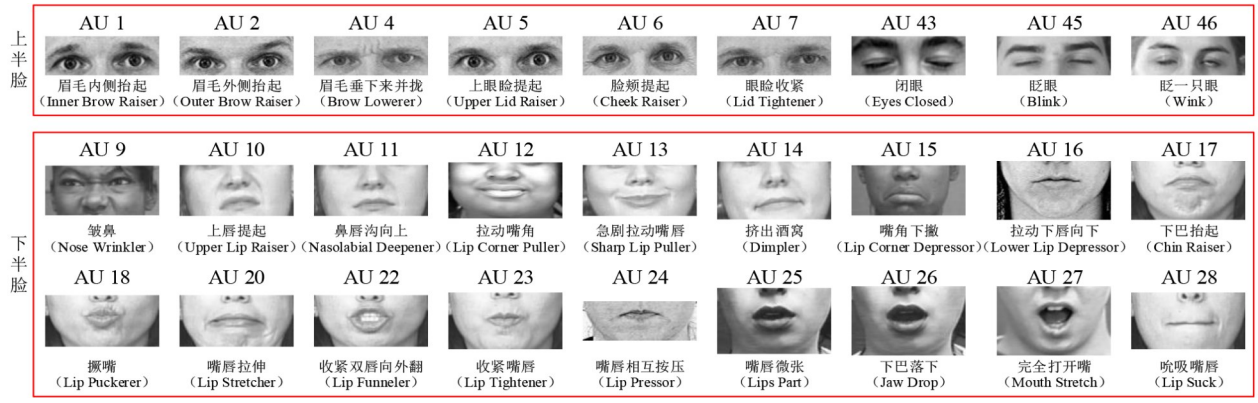


图2 常见的27个AU的示例图片及定义[6,18]

表1列出了每类整体表情中可能出现的AU[16],这些AU同时出现或部分同时出现于整体表情,例如快乐表情可以由AU 6,AU 12和AU 25的组合来表示,悲伤表情可以由AU 1,AU 4,AU 6和AU 17的组合来表示. 值得注意的是,人们在意识到自身表露出一种可能不合适的表情时经常会试图抑制它来隐藏真实的情绪,而只要试图掩盖原来的表情其面部便会自发地出现微表情(Micro-Expression)[19]. 微表情的持续时间很短,一般的界定标准为持续时间不超过500 ms[20],这是其区别于宏表情(Macro-Expression)的主要特征[21]. 微表情也可以用AU的组合进行描述,表2具体定义了每类微表情对应的AU组合[22],其中I,II,III,IV,V和VI类分别与快乐、惊讶、愤怒、厌恶、悲伤和恐

表1 每类整体表情所关联的AU[16]

整体表情	关联的AU
快乐	6, 12, 25
惊讶	1, 2, 5, 26, 27
愤怒	4, 5, 7, 10, 17, 22, 23, 24, 25, 26
厌恶	9, 10, 16, 17, 25, 26
悲伤	1, 4, 6, 11, 15, 17
恐惧	1, 2, 4, 5, 20, 25, 26, 27

惧相关,VII类与蔑视等其他微表情相关. 例如,微表情I类可以由AU 6,AU 7和AU 12的组合或单个AU 6来表示. 因此,研究AU识别对微表情识别同样具有重要意义.

表2　每类微表情对应的AU组合[22]

微表情	AU组合
I	6, 12, 6+12, 6+7+12, 7+12
II	1+2, 5, 25, 1+2+25, 25+26, 5+24
III	23, 4, 4+7, 4+5, 4+5+7, 17+24, 4+6+7, 4+38
IV	10, 9, 4+9, 4+40, 4+5+40, 4+7+9, 4 +9+17, 4+7+10, 4+5+7+9, 7+10
V	1, 15, 1+4, 6+15, 15+17
VI	1+2+4, 20
VII	其他AU组合

注:"+"表示AU的组合

经过观察,AU的组合可以形成7 000多种真实存在的表情[8]. 在某一人脸表情中,可能单独出现一个AU,也可能同时出现多个AU. 当多个AU同时出现时,若它们是可加性的(Additive),则AU的组合出现并不改变各AU的外观;若它们是不可加性的(Non-Additive),即它们的肌肉动作存在交叠区域,会融合成新的肌肉动作,则各AU的外观会被改变. 此外,一些AU组合如AU 1和AU 4在悲伤和恐惧表情中都会出现,比其他组合出现的频率更高. 另外,某些AU之间是相互排斥的,如AU 1和AU 7,两者不会同时出现在任一表情中,若一个AU出现则另一个AU不会出现.

2.2　基于深度学习的AU识别的定义

基于深度学习的AU识别主要包含3个环节,即人脸检测、人脸对齐和AU识别,如图3所示. 人脸检测指在输入图像上检测人脸的位置;人脸对齐指基于人脸配准所定位的面部特征点对人脸进行变换,使得变换后人脸与参照人脸(一般为平均脸)的对应特征点位置相同或相近;AU识别是基于深度神经网络实现,无须额外提取人工设计的特征,其从每张对齐后的人脸图像所提取的特征都对应于相同的面部语义位置,这有利于提升网络的特征学习以及进一步的分类或回归能力.

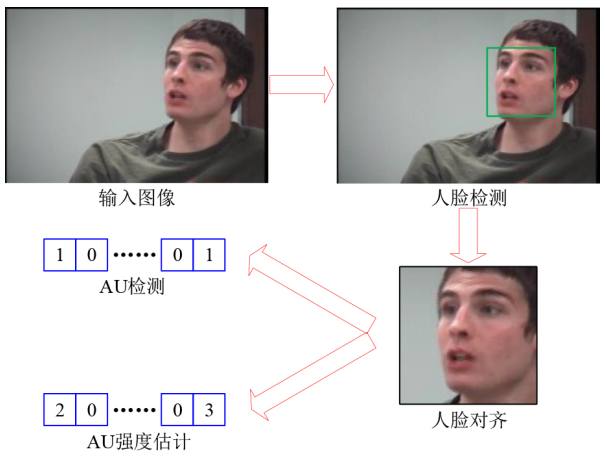


图3　基于深度学习的AU识别的定义

AU识别涵盖AU检测和AU强度估计2个子任务,其中前者指预测输入的人脸图像上每个AU是否出现,输出为每个AU出现的概率$\hat{\boldsymbol{p}}=(\hat{p}_1,\hat{p}_2,\cdots,\hat{p}_m)$,后者则进一步预测每个AU出现的强度$\hat{\boldsymbol{p}}^{(\mathrm{int})}=(\hat{p}_1^{(\mathrm{int})},\hat{p}_2^{(\mathrm{int})},\cdots,\hat{p}_m^{(\mathrm{int})})$,这里$m$为AU的个数. AU识别网络[23,24]的末端一般为全连接层,预测每个AU出现的概率$\hat{p}_i$,对于AU检测,通常采用多标签交叉熵损失(Loss):

$$L^{(\mathrm{det})}=-\frac{1}{m}\sum_{i=1}^{m}[p_i\log\hat{p}_i+(1-p_i)\log(1-\hat{p}_i)] \tag{1}$$

其中,p_i为第i个AU真实出现的概率,值为1表示出现,值为0则表示不出现. 对于AU强度估计,需要将第i个AU被预测的概率转换为强度:

$$\hat{p}_i^{(\mathrm{int})}=\hat{p}_iT \tag{2}$$

其中,T=5为最大的强度级别. 然后采用L2损失:

$$L^{(\mathrm{int})}=\frac{1}{m}\sum_{i=1}^{m}(p_i^{(\mathrm{int})}-\hat{p}_i^{(\mathrm{int})})^2 \tag{3}$$

在测试时,为了获得精确值,对于第i个AU,需要将其被预测出现的概率和强度分别离散化为$[\hat{p}_i]\in\{0,1\}$和$[\hat{p}_i^{(\mathrm{int})}]\in\{0,1,\cdots,T\}$,这里$[\cdot]$表示四舍五入取整.

2.3　基于深度学习的AU识别的挑战

AU作为出现在面部局部区域的细微表情动作,较难被准确捕捉,且人工地对其标注也较困难,因此基于深度学习的AU识别主要面临如下3个挑战因素.

(1)标签稀缺性:AU需要由经过培训的专家来标注,且标注过程较耗时,因而人工标注的成本很高[6],使得目前大多数被标注的数据集规模较小、样本多样性较低. 由于深度学习方法通常需要大量的训练数据,因此标签稀缺性是限制模型精度的重要因素.

(2)特征难捕捉性:AU是非刚性的,其外观随人和表情的变化而变化,且每个AU的形状不规则、不同

AU的大小一般不相同. 而且,人脸表情中时常会同时出现2个以上具有交叠区域的AU,存在不可加性,例如AU 1和AU 4在图1的悲伤表情中同时出现,它们会改变各自原来的外观,融合成新的面部肌肉动作. 这些都导致各AU所关联的局部表情细节难以被准确地捕捉.

(3)标签不均衡性:在人们经常表现的表情中,某些AU出现的频率比其他AU更高,且每一AU出现的频率时常低于不出现的频率,即AU的标签具有不均衡性,而当前AU数据集规模小、多样性低的情况加剧了这种不均衡性. 这些导致了AU识别模型对多个AU同时预测时容易偏向于提升出现频率较高AU的精度,而其他AU的精度则受到抑制,且容易偏向于将AU预测为不出现.

尽管深度学习显著提升了AU识别的性能,上述挑战仍是导致AU识别精度较低、不同AU精度差异较大的主要因素,如何克服这样的挑战是当前AU识别研究的热门方向.

2.4 AU数据集

自FACS[6,7]被提出以来,学术界克服AU数据采集、标注的困难,发布了多个AU数据集,促进了AU识别技术的发展. 早期的数据集,如CK[25]和MMI[18],是在受控环境下采集的具有良好光照和简单背景的正面或近似正面的人脸图片,受试者被要求人为地显露出指定的面部表情. 最近十年,研究人员更多地关注受试者被诱发而自发产生的表情,代表性数据集包括受控环境下采集的BP4D[26]和DISFA[27]等. 近年来,学术界发布了几个非受控场景下采集的数据集,如EmotioNet[28]和Aff-Wild2 (AU Set)[29],其包含的图片来自互联网等野外(Wild)场景,在光照、遮挡、姿态等方面变化多样.

本文接下来对一些流行的AU数据集进行介绍,由于数据集的采集环境(受控、非受控)、表情激发方式(人为、自发)、样本多样性(人脸身份数、图片或视频数)、数据形式(2D,3D)都会影响模型的训练效果,表3对数据集的这些属性进行了总结. 此外,图4展示了这些数据集的示例图片.

表3 AU数据集的属性

数据集	采集环境	表情激发方式	人脸身份数	图片或视频数	数据形式	发布年份
CK[25]	受控	人为	97	486个视频	2D	2000
BP4D[26]	受控	自发	41	328个视频	2D, 3D	2014
DISFA[27]	受控	自发	27	27个视频	2D	2013
EmotioNet[28]	非受控	自发	—	约975 000张图片	2D	2016
Aff-Wild2 (AU Set)[29]	非受控	自发	63	63个视频	2D	2019
CK+[30]	受控	自发	26	107个视频	2D	2010
MMI[18,31]	受控	人为、自发	67	2 390个视频和493张图片	2D	2005, 2010
Bosphorous[32]	受控	人为	105	4 652张图片	3D	2008
ICT-3DRFE[33]	受控	人为	23	345张图片	2D, 3D	2011
D3DFACS[34]	受控	人为	10	519个视频	2D, 3D	2011
BP4D+[35]	受控	自发	140	1 400个视频	2D, 3D	2016
GFT[36]	受控	自发	96	96个视频	2D	2017
CASME II[37]	受控	自发	35	247个视频	2D	2014
SAMM[38]	受控	自发	32	159个视频	2D	2018
MMEW[39]	受控	自发	30	300个视频和900张图片	2D	2021

注:"—"表示数据集的这一属性没有被公布

不难发现,所有受控环境下采集的数据集只包含数十个或100多个不同身份的人脸,虽然每个人脸可能被录制一或多个场景,产生数千至数万视频帧,但整体上样本多样性仍较低. 另外,非受控环境下采集的数据集EmotioNet和Aff-Wild2(AU Set)具有显著更高的样本多样性,然而它们仅被标注AU的出现和不出现2种状态,没有被标注AU的由0到5的强度,限制了其适用范围. 这些都是由AU的人工标注成本很高所导致的. 从数据集的演变趋势也可以看出,研究人员由关注受控环境逐渐转向非受控环境,由于非受控环境采集的图片具有丰富的多样性,所以训练深度学习模型需要更大规模的数据,而对多样性变化的样本进行标注也会面临更高的成本. 因此标签稀缺性是当前AU数据集存在的普遍问题,尤其是在非受控场景.

(*a*) **CK**[25]和**CK+**[30]. 左图红色框中出现AU 6和AU 7，蓝色框中出现AU 10，AU 12，AU 20和AU 25；右图红色框中出现AU 4，AU 6，AU 7和AU 9，蓝色框中出现AU 17

(*b*) **MMI**[18,31]. 左图红色框中出现AU 7，蓝色框中出现AU 12；右图红色框中出现AU 5和AU 6，蓝色框中出现AU 10，AU 12，AU 14，AU 20和AU 25

(*c*) **Bosphorous**[32]. 左图红色框中出现AU 1和AU 5；右图红色框中出现AU 5，蓝色框中出现AU 12，AU 17，AU 20和AU 23

(*d*) **ICT-3DRFE**[33]. 左图中未出现AU；右图红色框中出现AU 6，蓝色框中出现AU 12，AU 14，AU 20和AU 23

(*e*) **D3DFACS**[34]. 左图红色框中出现AU 1和AU 5；右图红色框中出现AU 1和AU 5，蓝色框中出现AU 10

(*f*) **BP4D**[26]和**BP4D+**[35]. 左图红色框中出现AU 7，蓝色框中出现AU 12；右图红色框中出现AU 2，蓝色框中出现AU 10，AU 12和AU 25

(*g*) **DISFA**[27]和**GFT**[36]. 左图红色框中出现AU 6和AU 7，蓝色框中出现AU 12，AU 14，AU 17和AU 20；右图红色框中出现AU 2和AU 6，蓝色框中出现AU 10，AU 12，AU 14和AU 25

(*h*) **CASME II**[37]. 左图红色框中出现AU 4，蓝色框中出现AU 10和AU 15；右图蓝色框中出现AU 17

(*i*) **SAMM**[38]. 左图红色框中出现AU 4，AU 6和AU 7，蓝色框中出现AU 12，AU 14和AU 20；右图红色框中出现AU 1和AU 2，蓝色框中出现AU 10，AU 14，AU 15，AU 20和AU 25

(*j*) **MMEW**[39]. 左图红色框中出现AU 2；右图红色框中出现AU 1和AU 2，蓝色框中出现AU 15

(*k*) **EmotioNet**[28]. 左图蓝色框中出现AU 15和AU 26；右图红色框中出现AU 2和AU 5，蓝色框中出现AU 10，AU 12，AU 14和AU 26

(*l*) **Aff-Wild2(AU Set)**[29]. 左图红色框中出现AU 4；右图红色框中出现AU 1，AU 2和AU 5，蓝色框中出现AU 12和AU 26

图4　AU数据集的示例图片(每张图片所出现的AU被红色或蓝色框标出)

3　基于深度学习的AU识别方法进展

针对标签稀缺性，可以利用迁移学习将有用的知识迁移到当前任务；针对特征难捕捉性，可以从准确捕捉AU的关联区域从而提取AU特征来切入；针对标签不均衡性，可以考虑利用AU间的关联对不均衡的AU进行平衡. 本文接下来分别予以介绍.

3.1　基于迁移学习的方法

迁移学习的目标是弥补有人工标签的训练样本的不足，将相关联的样本、标签、模型或先验知识等迁移过来，提升当前任务的模型性能.

3.1.1　基于已有模型的迁移学习

最常见的迁移学习方法是在当前数据集上微调其他图像数据集上预训练的模型，由于不同类型的图像时常具有相似的颜色分布和背景环境等属性，预训练模型所携带的知识也有利于当前模型的训练. Zhou等人[40]基于一个在ImageNet[41]上预训练的VGG16[42]网络，实现AU强度估计和头部姿态估计. Ji等人[43]在整体表情识别和人脸识别这2个与AU相关联任务的数据集上分别预训练ResNet-34网络[44]，接着在AU数据集上分别微调2个网络，并将2个网络预测的AU出现概率取平均作为最终的预测值. 预训练的数据集与当前数据集之间存在域(Domain)差异，且微调过程可能会丢失一些有用信息，因而限制了微调预训练模型的有效性.

另一个基于已有模型的思路是生成伪标签，即利用训练好的AU识别模型对图片自动地标注，这实质上是利用了AU识别模型中存储的训练数据的知识. Benitez-Quiroz等人[28]发布了一个从互联网上抓取的非受控场景人脸图片数据集EmotioNet，其中优化集具有准确的人工标签，而训练集只有受控场景图片上训练的模型所标注的伪标签. 考虑到自动标注模型的训练数据与被标注图片之间存在域差异，自动标注的伪标签并不准确. 为改进EmotioNet的伪标签，Werner等人[45]采用一个自训练方法，以多任务的形式同时在优化集和训练集上训练深度卷积神经网络(Deep Convolutional Neural Network，DCNN)，其中优化集对应的分类器分支作为最终分类器，然后利用训练好的模型对训练集图片重新标注伪标签，再重新训练网络，重复这一过程直至性能已收敛或已满足精度要求. 然而，这一自训练方法依赖优化集的人工标签.

3.1.2　基于已有标签的迁移学习

由于人工标注AU的成本高昂，很多情况下数据集中只有部分样本拥有完整的AU标签，而其余样本没有AU标签或只有一部分AU的标签. 这里极端的情况是所有样本都没有AU标签，而只有粗略的标签如整体表情标签是可用的，由于其对表情的描述没有AU精细，

因而标注成本很低.

由表1不难看出,AU与整体表情之间存在条件依赖关系.Peng等人[46]从多个AU数据集中统计出给定整体表情下某一AU出现的条件概率,并结合先验的AU间关系,从表情标签生成AU的伪标签.进一步地,Peng等人[47]基于全部样本的表情标签和部分样本的AU标签,提出一个对偶半监督的生成对抗网络(Generative Adversarial Network,GAN)[48],联合地学习AU分类器和人脸图片生成器.由于任务的对偶性,AU分类器的输入输出联合分布和人脸生成器应该是一致的,该方法通过对抗学习迫使输入输出联合分布收敛到AU-表情标注数据的真实分布.Zhang等人[49]将表情独立的和表情依赖的AU概率作为约束融入目标函数,促进AU分类器的训练.然而,将固定的先验知识应用于所有样本忽视了不同样本间AU动态变化的特性.

另一些方法在具有AU标签的样本基础上,引入大量无标签的样本.Wu等人[50]基于深度神经网络学习人脸特征,并利用受限玻尔兹曼机(Restricted Boltzmann Machine,RBM)从部分样本的完整AU标签中学习标签分布,然后通过最大化AU映射函数相对于所有无标签数据的标签分布的似然对数,同时最小化有标签数据的AU预测值和真实值之间的误差,来训练AU分类器.然而,从有限样本学习的标签分布可能并不适用于其他样本.Zhang等人[51]利用4种先验的AU约束来额外地监督训练过程:越临近帧的特征越相似、一段AU动作中强度随帧非递减、面部对称性、相对于中性表情外观的差异性.该方法在训练时要求图像序列中某一AU在一段动作过程中峰值和谷值所在帧具有该AU的标签,降低了适用性.

此外,相关联任务的已有标签也可以被利用来促进AU识别.Shao等人[52]采取多任务学习,基于CNN联合地实现人脸AU识别和人脸配准,利用任务间的关联性使得彼此相互促进,且配准分支学习的特征被传入AU识别任务,有利于提升AU识别精度.Jyoti等人[53]将整体表情识别网络所提取的特征传入AU识别网络,促进AU识别.Tu等人[54]采用底部层共享的人脸识别网络和AU识别网络,其中人脸识别网络学习身份特征,然后AU识别网络所提取的特征在减去身份特征后进一步回归AU预测值.这类方法的效果很大程度上依赖任务间的关联性强弱以及所设计多任务结构的有效性.

3.1.3 基于域映射的迁移学习

域映射指从一个域映射到另一个域,其中域包括图像、特征、标签等.近年来,一些工作通过域适应(Domain Adaptation)来提取源域知识,使其适应目标域,从而促进目标域任务的学习.一个常见做法是将目标图片的表情编辑为源图片的表情,从而将源图片的AU标签迁移到新生成的目标图片上,实现数据扩增.Liu等人[55]以源AU标签为条件,基于条件GAN[56]生成源表情参数,再与目标图片的其他人脸属性参数组合,利用3D可变模型(3D Morphable Model,3DMM)[57]生成具有源表情和目标图片纹理的新图片.Wang等人[58]在不依赖3DMM的情况下同时训练GAN和AU分类器,合成具有源图片AU属性且保留目标纹理的新图片.然而这2个工作针对的源图片和目标图片都仅来自受控场景.

除了域适应外,域映射的另一个应用是自监督学习,其从数据本身的结构推断出监督信号而不需要AU标签.Wiles等人[59]提出一个人脸属性网络,输入为来自同一视频的目标帧和源帧,首先编码器学习目标帧和源帧的人脸属性特征,两者被串联起来输入到解码器中生成具有源帧表情和目标帧姿态的新图像,其中解码器对生成图像上每一像素与源帧像素的位置对应关系进行预测,同时约束生成图像与目标帧相似,这里人脸属性特征包含了表情信息,因而可以用于AU识别.考虑到AU是面部肌肉动作,Li等人[60]将视频中2张不同帧之间的人脸变化视为动作,并以此为自监督信号来学习特征,具体采用一个双循环自编码器,将AU相关的动作和头部姿态相关的动作解耦出来,从而得到AU相关的特征.然而这些方法要求训练时输入的一对图像来自同一视频且具有相同的人脸身份,限制了其适用性.

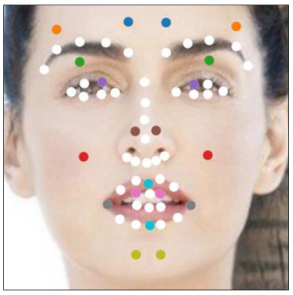
3.2 基于区域学习的方法

AU为人脸局部肌肉动作,因而提取其特征需要准确定位关联区域,每个AU的关联区域包括其所在部位以及存在一定关联的其他部位.

3.2.1 特征点辅助的区域学习

FACS基于客观的人脸解剖学来定义AU,每个AU的中心与人脸特征点之间有先验的位置关系,图5展示了一些常见AU的位置定义规则[61,52],因此可以通过特征点来准确确定AU的中心位置,从而提取与AU关联的局部特征.Jaiswal等人[62]利用特征点为每个AU预定义方形的感兴趣区域(Region of Interest,ROI)以及对应的二进制掩膜(Mask),其中掩膜上特征点形成的多边形区域内点的值为1而其他点的值为0,然后基于CNN从裁剪的ROI和掩膜提取每个AU的特征.Ali等人[63]先利用一个卷积层提取低层特征,然后根据特征点位置在这一特征图(Feature Map)上裁剪与AU的ROI对应的方块,并分别利用一个CNN从每个方块进一步提取特征.Ma等人[64]利用特征点为AU定义边界框(Bounding Box),将通用的物体检测问题融入AU识别,预测AU在哪个边界框出现,若某一AU不出现于当前人脸,则对于所有边界框都应被预测为不出现.这些方

法将ROI内所有位置视为相等的重要性,没有考虑到离AU中心越近的位置应该与AU越相关.



AU	中心位置
1（眉毛内侧抬起）	眉毛内侧向上1/2尺度
2（眉毛外侧抬起）	眉毛外侧向上1/3尺度
4（眉毛垂下来并拢） 5（上眼睑提起）	眉毛中心向下1/3尺度
6（脸颊提起）	眼睛底部向下1尺度
7（眼睑收紧）	眼睛
9（皱鼻）	鼻子底部向上1/2尺度
10（上唇提起）	上唇中心
12（拉动嘴角） 14（挤出酒窝） 15（嘴角下撇） 20（嘴唇拉伸）	唇角
17（下巴抬起） 26（下巴落下）	嘴唇下方1/2尺度
23（收紧嘴唇） 24（嘴唇相互按压） 25（嘴唇微张）	嘴唇中心

图5　常见AU的中心位置及可视化

注:其中每2个相同颜色的点表示某一AU的2个对称的中心."尺度"指2个内眼角之间的距离.白色的点表示49个人脸特征点,其中一些点被AU的中心覆盖

Li等人[61,65]为每个AU的ROI定义注意力(Attention)分布,ROI内离中心越近的位置其注意力权重越大,然后在AU识别网络中利用注意力图对特征图中的AU特征进行增强,并在网络的末端从特征图上裁剪每个AU的ROI方块.Sanchez等人[66]依据AU的标签将其注意力分布定义为高斯分布,特征点决定分布的中心位置而AU强度决定分布的振幅和大小,然后利用CNN从输入图像回归每个AU的注意力图来实现AU强度估计.考虑到AU会随人和表情变化而非刚性变化且不可加性导致AU的外观改变,Shao等人[52]在CNN中利用配准分支所预测的特征点定义初始注意力图,然后利用AU识别的监督信号自适应地优化每个AU的注意力图,从而更准确地捕捉AU关联区域.然而,上述方法均被特征点先验知识所约束,每个AU的注意力高亮区域集中在预定义ROI的附近,难以准确捕捉远离预定义ROI的关联区域.

3.2.2　自适应区域学习

当利用AU标签来有监督地训练深度神经网络时,网络在特征学习过程中会隐式地自适应捕捉AU的关联区域.Liu等人[67]迭代地在CNN学习的特征图上选择与目标表情标签相关性最高的特征,这些特征所在区域被期望为与AU关联的感受野,然后将这些感受野内的特征输入到RBM来实现表情分类.考虑到不同人脸区域的AU具有不同的结构和纹理属性,对不同区域应该采用独立而不是共享的滤波器,Zhao等人[68]引入分块卷积层,将特征图划分为相同大小的多个小块,在每一小块内部采用独立的卷积滤波器来提取特征,该特征图能够隐式地捕捉AU的关联区域.为了适应不同大小的AU,Han等人[69]提出自适应大小的卷积滤波器,在训练CNN时学习卷积层的滤波器大小和权重参数.然而,这些方法没有以显式的方式来自适应学习关联区域,因此只能粗略地确定AU的区域位置.

近年来,一些工作在网络中加入注意力学习模块,显式地捕捉AU关联区域.Shao等人[24]不依赖特征点的先验约束,直接通过AU识别的监督信号自适应地学习通道级注意力和空间注意力,同时利用全连接条件随机场(Conditional Random Field,CRF)捕捉像素级关系来优化空间注意力,从而选择和提取每个AU的关联特征.Ertugrul等人[70,71]分别采用一个CNN从裁剪的人脸块提取特征,接着利用注意力机制对各个块所提取的特征进行加权,实现AU识别.虽然上述工作能够较好地捕捉AU特征,但仍包含了一些不相关的信息,影响AU识别的精度.

3.3　基于关联学习的方法

人脸表情涉及多个局部位置的肌肉动作,因而像素位置间的关系可以被利用起来.表情中会时常出现多个AU,但不会所有AU都出现,因而除部分AU相互独立(不相关)外,多数AU之间并不独立,可能同时出现(正相关),也可能相互排斥(负相关).而且,在视频中AU是动态变化的,挖掘时域关联可以促进AU识别.

3.3.1　像素级关联学习

Shao等人[24]利用全连接CRF捕捉像素级关联关系,对每个AU的空间注意力进行优化,从而捕捉更准确的AU特征.Niu等人[72]首先利用CNN提取人脸特征,这一特征的空间上每一点沿通道的特征向量被作为一个局部特征,接下来利用长短期记忆(Long Short-Term Memory,LSTM)网络学习局部特征间的关系,由于不同AU涉及不同位置的肌肉动作,该方法对每个AU分别采用一个LSTM来学习不同局部特征的贡献.鉴于密集的人脸特征点可以描述人脸几何结构,Fan等人[73]利用图卷积网络(Graph Convolutional Network,GCN)从特征点空间位置形成的几何图结构中学习一个隐向量,该隐向量包含人脸形状模式以及特征点间的相互依赖关系,在特征学习过程中被用来增强表征能力.在这些工作中像素与AU的对应并不明确,使得像素级关系对AU识别的促进作用较有限.

3.3.2　AU级关联学习

考虑到AU的强度级别从0到5是有序的(Ordinal),Tran等人[74]引入变分有序高斯过程自编码器(Variational Ordinal Gaussian Process Auto-Encoder,VOGPAE),在学习隐特征时施加AU强度有序关系的约束.Benitez-Quiroz等人[75]提出一个全局-局部损失,其中局部损失分别促进每个AU的预测,而全局损失对2个或2个以上AU真实值均为出现即正相关的情况进行约束,促进对正相关AU的预测.Walecki等人[76]将CNN

和CRF组合在一个端到端的框架中,其中CRF的一元能量项捕捉AU强度的有序结构,二元能量项捕捉AU间的依赖关系. Corneanu等人[77]将CNN和循环神经网络(Recurrent Neural Network,RNN)组合成一个深度结构推理网络(Deep Structure Inference Network,DSIN),其中RNN由许多结构推理单元构成,采用门控策略控制每2个AU结点间的信息传递,从而推理AU之间的结构关系. Jacob等人[78]采用一个注意力网络来回归每个AU由特征点所预定义的注意力图,然后将注意力增强后的AU特征输入到一个变换器(Transformer)中,捕捉AU间的关系.

近年来,图神经网络(Graph Neural Network,GNN)开始被应用于AU关联学习. Li等人[79]从多个AU数据集统计出AU对的3种依赖关系,基于此构建有向的AU关系图,每个AU是一个结点,结点间的有向边类型包括正相关和负相关2种,AU间不相关则没有边相连,然后利用门控GNN[80]对AU关系建模. Liu等人[81]和Niu等人[82]首先基于数据集统计的依赖关系构建AU关系图,然后利用GCN建模AU间的关系. 由于AU间依赖可能随人和表情的变化而变化,另一些工作采用动态的关系图结构. Fan等人[83]提出一个语义对应卷积(Semantic Correspondence Convolution,SCC)模块,将前一层的每个特征图通道作为一个结点,构建K-近邻图,动态地计算通道间的语义对应,由于每个通道编码了AU的一个特定模式,这样可以学习AU间的关系. Song等人[84]提出不确定图卷积(Uncertain Graph Convolution),自适应地学习基于概率的掩膜来捕捉个体样本的AU间依赖以及不确定性. Song等人[85]提出一个混合信息传递神经网络,利用性能驱动的蒙特卡罗马尔可夫链采样方法来学习AU关系图,然后在信息传递过程中动态地组合不同类型信息使它们相互补充.

此外,为了抑制标签不均衡导致的预测偏置,许多工作通过调整采样率和权重来进行平衡. Li等人[61]在深度神经网络的训练过程中对训练集中出现频率较低的AU采用更大的随机采样率,使得每个小批量(Mini-Batch)中不同AU出现的频率较均衡. 另一些工作[24,52,77]在计算AU识别损失时,给每一AU所赋的权重与该AU出现的频率成反比,从而加强了出现频率较低的AU. 此外,为了平衡每个AU的出现频率和不出现频率,Li等人[79]对交叉熵损失中出现频率的熵项乘以训练集中该AU的不出现频率,而对不出现频率的熵项乘以该AU的出现频率,这样,若某一AU的不出现频率大于出现频率,其对应于出现的损失项被加强. Song等人[84]提出自适应加权损失函数,通过自适应地学习认知不确定性(Epistemic Uncertainty)来计算小批量中每个样本的权重,不确定性越高的样本被赋以越大的权重,从而抵消数据不均衡.

上述方法所学习的AU关联依赖训练数据集的AU标签分布,使得训练的AU识别模型难以适应跨数据集测试,泛化能力较低.

3.3.3 时域关联学习

当前采用时域关联学习的方法一般先提取视频中每帧人脸图像的空间特征,然后利用LSTM等时间序列模型对时域上帧间关联进行建模. Chu等人[86]采用CNN提取各帧空间特征,并用LSTM对帧间的时域信息进行建模,最后在CNN和LSTM的末端将时空特征进行融合. Bishay等人[87]设计一个三层级的框架:在第一层级利用CNN学习人脸外观特征,并利用多层感知机从人脸特征点学习几何特征;在第二层级利用RNN从连续帧学习时域上的关联;在第三层级将各网络的预测结果进行融合. He等人[88]将双向LSTM与RNN结合起来学习时域特征. Song等人[89]利用多个LSTM同时挖掘时域和空间域上的关联信息. Yang等人[90]采用2D的CNN对每帧图像提取特征,同时采用3D的CNN捕捉图像序列的时空信息,从而实现AU识别. Yang等人[91]利用单张图像及一张锚定图像来无监督地学习光流,从而捕捉时域信息,再将光流输入到AU识别网络进行AU预测,这里光流网络和AU识别网络被联合地训练,使得AU标签可以提供语义信息从而促进光流的学习. Zhang等人[92]利用注意力机制实现特征融合和标签融合,其中前者用于捕捉人脸局部块间的空间关系,而后者用于捕捉时域动态关系.

这些工作主要是将已有的时间序列模型应用于AU识别任务,并未明确地对AU在时域上动态非刚性变化的过程进行分析和处理,限制了时域关联学习的有效性.

4 代表性AU识别方法对比

4.1 实验设置

在这一节,本文选取近年来最广泛使用的AU数据集BP4D[26]和DISFA[27],将相同实验设置下基于深度学习的AU识别工作所报告的结果进行展示. AU识别包括AU检测和AU强度估计,下面分别介绍在2个数据集上具体的实验设置.

4.1.1 AU检测

在BP4D和DISFA上按照文献[68,61,52]的设置,采用3-折交叉验证(3-Fold Cross-Validation),每折包含的人脸身份无交叠,每次实验其中两折用于训练而剩余的一折用于测试,BP4D为在12个AU(1,2,4,6,7,10,12,14,15,17,23和24)上评估,DISFA为在8个AU(1,2,4,6,9,12,25和26)上评估. 评价指标采用基于

帧的F1分数(Frame-Based F1-Score),其定义为

$$\mathrm{F1}=\frac{2PR}{P+R} \tag{4}$$

其中P指精确率(Precision),R指召回率(Recall),F1分数能够可靠地度量AU标签出现和不出现频率不均衡情况下模型的性能.

4.1.2 AU强度估计

在BP4D和DISFA上按照文献[76,51,24]的设置,BP4D的训练集包含21个人脸身份,测试集包含20个人脸身份,在5个AU(6,10,12,14和17)上评估,DISFA的训练集包含18个人脸身份,测试集包含9个人脸身份,在12个AU(1,2,4,5,6,9,12,15,17,20,25和26)上评估. 评价指标采用组内相关系数(Intra-Class Correlation,ICC(3,1))[93].

4.2 性能对比

表4、表5分别对代表性的基于深度学习的AU检测和AU强度估计方法进行了总结和对比,从中可以观察到如下几方面的现象.

表4　基于深度学习的AU检测代表性方法总结

方法	基于已有模型的迁移学习	基于已有标签的迁移学习	基于域映射的迁移学习	特征点辅助的区域学习	自适应区域学习	像素级关联学习	AU级关联学习	时域关联学习	BP4D[26]/DISFA[27]
DRML[68]					√				0.483/0.267
EAC-Net[61]	√			√			√		0.559/0.485
R-T1[65]	√			√			√	√	**0.661**/0.513
DSIN[77]				√			√		0.589/0.536
LP-Net[72]	√			√		√			0.610/0.569
MLCR[82]		√					√		0.598/—
SRERL[79]	√			√			√		0.629/0.559
ARL[24]					√	√	√		0.611/0.587
PAttNet[70]				√					0.626/—
D-PAttNet[71]				√				√	0.641/—
TAE[60]			√		√				0.603/0.515
OF-Net[91]		√						√	0.597/0.537
AU R-CNN[64]		√		√					0.630/0.513
AU-GCN[81]				√			√		0.628/0.550
JÂA-Net[52]		√		√			√		0.624/**0.635**
UGN-B[84]	√						√		0.633/0.600
Transformer[78]	√			√			√		0.642/0.615
HMP-PS[85]	√						√		0.634/0.610

注:表中展示了每个方法原始文献所报告的在所有AU上的平均F1分数

表5　基于深度学习的AU强度估计代表性方法总结

方法	基于已有模型的迁移学习	基于已有标签的迁移学习	基于域映射的迁移学习	特征点辅助的区域学习	自适应区域学习	像素级关联学习	AU级关联学习	时域关联学习	BP4D[26]/DISFA[27]
DRML[68,76]					√				0.52/0.29
CCNN-IT[76]							√		0.63/0.45
2DC[74]							√		0.66/0.50
KBSS[51]		√						√	0.67/0.36
ARL[24]					√	√			0.66/0.48
SCC[83]				√			√		**0.72**/0.47
G2RL[73]	√			√		√			0.69/0.52
DPG[89]							√	√	**0.72/0.56**

注:这里展示了每个方法原始文献所报告的在所有AU上的平均ICC

(1)目前研究AU检测的工作多于AU强度估计,这是因为强度估计不仅需要判断每个AU是否出现,还需识别AU的强度,更具挑战性.

(2)大多数AU识别工作将迁移学习、区域学习和

关联学习中多种策略进行结合,而不是仅基于一种学习策略,这是因为实现高精度的AU识别需要同时解决标签稀缺性、特征难捕捉性和标签不均衡性的挑战.

(3)采用关联学习的工作如R-T1[65],D-PAttNet[71]和DPG[89]取得相比于其他工作更高的精度,表明AU间关联以及时域关联对AU识别具有重要意义.

(4)当前基于迁移学习的工作如MLCR[82]和TAE[60]并未取得相比于其他工作明显的性能优势,说明这类方法仍有较大的挖掘空间,需要进一步从AU的特性出发,提出有效的模型来充分利用已有的样本、标签、模型以及先验知识.

(5)与JÂA-Net[52]和G2RL[73]相比,R-T1[65],AU R-CNN[64],KBSS[51]和SCC[83]等工作无法在BP4D和DISFA上同时取得较高的精度,说明AU识别模型的可靠性和泛化能力也是需要着重研究的地方.

5 总结与展望

目前,表情AU识别技术已取得较大的发展,但其精度仍有很大的提升空间,无法很好地满足实际应用需求. 未来可从以下几方面进一步进行探索.

(1)已有基于迁移学习的工作尚无法有效地解决标签稀缺性挑战. 未来可以采取融合多种策略的方式:①将具有AU标签的样本作为源样本,利用GAN将无标签目标样本的表情编辑为源表情,则其具有源样本的AU标签,这些新生成的目标样本提高了训练数据的多样性;②利用最新的人脸配准开源库对样本标注特征点,同时结合具有整体表情标签的数据集,挖掘特征点、整体表情与AU间关联性,促进AU识别;③将自监督学习、有监督学习、域适应多种方法综合起来,利用自监督学习从无标签样本中学习AU本质属性的特征表示,利用有监督学习从具有AU标签的样本中学习AU识别模型,利用域适应使得其他域训练的模型可以被应用于当前域.

(2)当前的AU识别模型在对多个AU同时预测时仍易于偏向提升出现频率较高AU的精度,以及偏向将AU预测为不出现,标签不均衡性依然严重限制着AU识别的精度. 可选的解决方案为:①利用GAN进行数据扩增,尽量使所生成的数据集在每个AU的出现与不出现频率、不同AU间的出现频率方面保持均衡;②借鉴已有的处理长尾分布等不均衡数据的方法,对不均衡的AU标签分布进行建模,充分挖掘不同AU间的关联关系.

(3)现有的工作主要关注受控环境,更接近实际应用场景的非受控AU识别的相关研究仍较少. 未来可从以下角度切入非受控环境的研究:①研究受控域到非受控域的AU迁移方法,利用具有AU标签的受控域数据集生成新的非受控域样本,扩增非受控域训练数据;②提高方法对不同头部姿态的鲁棒性,可以定位3D的人脸特征点、构造UV映射、计算3D人脸表面的测地距离,这些辅助信息都可以加到深度神经网络中,在输入、中间的特征提取或者后置处理环节提升AU识别的精度;③利用特征解耦方法将光照、姿态、遮挡等信息从AU特征中分离,实现光照无关、姿态无关、遮挡无关的AU识别.

(4)当前的AU数据集具有样本规模小且多样性低、标签稀缺且不均衡、缺乏非受控样本等不足. 未来可以构建一个规模大、样本多样性丰富、AU标注全面的非受控环境数据集. 由于对AU进行人工标注的成本很高,在标注的过程中,可以基于主动学习(Active Learning)[94~96],从一个具有人工标注的小训练集开始,训练模型并对未标注样本进行预测,然后基于预测结果选择信息最丰富、存在出现频率较低AU的未标注样本进行人工标注,再将新标注的样本加入训练集并更新模型,重复上述步骤直至被训练的模型在测试集上的性能已收敛或已满足精度要求,这样可以保证有限的标注成本用在最需要的样本上.

参考文献

[1] 邱玉, 赵杰煜, 汪燕芳. 结合运动时序性的人脸表情识别方法[J]. 电子学报, 2016, 44(6): 1307-1313.
QIU Y, ZHAO J Y, WANG Y F. Facial expression recognition using temporal relations among facial movements [J]. Acta Electronica Sinica, 2016, 44(6): 1307-1313. (in Chinese)

[2] 孙晓, 潘汀. 基于兴趣区域深度神经网络的静态面部表情识别[J]. 电子学报, 2017, 45(5): 1189-1197.
SUN X, PAN T. Static facial expression recognition system using roi deep neural networks[J]. Acta Electronica Sinica, 2017, 45(5): 1189-1197. (in Chinese)

[3] 张瑞, 蒋晨之, 苏剑波. 基于稀疏特征挑选和概率线性判别分析的表情识别研究[J]. 电子学报, 2018, 46(7): 1710-1718.
ZHANG R, JIANG C Z, SU J B. Expression recognition based on sparse selection and plda[J]. Acta Electronica Sinica, 2018, 46(7): 1710-1718. (in Chinese)

[4] 孔德壮, 朱梦宇, 于江坤. 人脸表情识别在辅助医疗中的应用及方法研究[J]. 生命科学仪器, 2019, 18(2): 43-48.
KONG D Z, ZHU M Y, YU J K. Research on the application and method of facial expression recognition in assistive medical care[J]. Life Science Instruments, 2019, 18 (2): 43-48. (in Chinese)

[5] FRANK M G, EKMAN P. The ability to detect deceit generalizes across different types of high-stake lies[J]. Journal of Personality and Social Psychology, 1997, 72(6): 1429-

1439.

[6] EKMAN P, FRIESEN W V. Facial Action Coding System: A Technique for the Measurement of Facial Movement [M]. Palo Alto: Consulting Psychologists Press, 1978.

[7] EKMAN P, FRIESEN W V, HAGER J C. Facial Action Coding System[M]. Salt Lake City: Research Nexus, 2002.

[8] SCHERER K R, EKMAN P. Handbook of Methods in Nonverbal Behavior Research[M]. Cambridge: Cambridge University Press, 1982.

[9] FASEL B, LUETTIN J. Automatic facial expression analysis: Survey[J]. Pattern Recognition, 2003, 36(1): 259-275.

[10] 刘晓旻, 谭华春, 章毓晋. 人脸表情识别研究的新进展[J]. 中国图象图形学报, 2006, 11(10): 1359-1368.
LIU X M, TAN H C, ZHANG Y J. New research advances in facial expression recognition[J]. Journal of Image and Graphics, 2006, 11(10): 1359-1368. (in Chinese)

[11] KUMARI J, RAJESH R, POOJA K M. Facial expression recognition: A survey[J]. Procedia Computer Science, 2015, 58(1): 486-491.

[12] CORNEANU C A, SIMÓN M O, COHN J F, et al. Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(8): 1548-1568.

[13] LI S, DENG W. Deep facial expression recognition: A survey[EB/OL]. (2020-03-17)[2022-05-09]. https://doi.org/10.1109/TAFFC.2020.2981446.

[14] 贲晛烨, 杨明强, 张鹏, 等. 微表情自动识别综述[J]. 计算机辅助设计与图形学学报, 2014, 26(9): 1385-1395.
BEN X Y, YANG M Q, ZHANG P, et al. Survey on automatic micro expression recognition methods[J]. Journal of Computer-Aided Design and Computer Graphics, 2014, 26(9): 1385-1395. (in Chinese)

[15] 徐峰, 张军平. 人脸微表情识别综述[J]. 自动化学报, 2017, 43(3): 333-348.
XU F, ZHANG J P. Facial microexpression recognition: A survey[J]. Acta Automatica Sinica, 2017, 43(3): 333-348. (in Chinese)

[16] MARTINEZ B, VALSTAR M F, JIANG B, et al. Automatic analysis of facial actions: A survey[J]. IEEE Transactions on Affective Computing, 2019, 10(3): 325-347.

[17] ZHI R, LIU M, ZHANG D. A comprehensive survey on automatic facial action unit analysis[J]. The Visual Computer, 2020, 36(5): 1067-1093.

[18] PANTIC M, VALSTAR M, RADEMAKER R, et al. Web-based database for facial expression analysis[C]// Proceedings of the IEEE International Conference on Multimedia and Expo. Amsterdam: IEEE, 2005: 1-5.

[19] MATSUMOTO D, YOO S H, NAKAGAWA S. Culture, emotion regulation, and adjustment[J]. Journal of Personality and Social Psychology, 2008, 94(6): 925-937.

[20] YAN W J, WU Q, LIANG J, et al. How fast are the leaked facial expressions: The duration of micro-expressions[J]. Journal of Nonverbal Behavior, 2013, 37(4): 217-230.

[21] SHEN X, WU Q, FU X. Effects of the duration of expressions on the recognition of microexpressions[J]. Journal of Zhejiang University-Science B, 2012, 13(3): 221-230.

[22] DAVISON A K, MERGHANI W, YAP M H. Objective classes for micro-facial expression recognition[J]. Journal of Imaging, 2018, 4(10): 119.

[23] GUDI A, TASLI H E, DENUYL T M, et al. Deep learning based facs action unit occurrence and intensity estimation[C]//Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. Ljubljana: IEEE, 2015: 1-5.

[24] SHAO Z, LIU Z, CAI J, et al. Facial action unit detection using attention and relation learning[EB/OL]. (2019-10-23) [2022-05-09]. https://doi.org/10.1109/TAFFC.2019.2948635.

[25] KANADE T, COHN J F, TIAN Y. Comprehensive database for facial expression analysis[C]//Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. Grenoble: IEEE, 2000: 46-53.

[26] ZHANG X, YIN L, COHN J F, et al. Bp4d-spontaneous: A high-resolution spontaneous 3d dynamic facial expression database[J]. Image and Vision Computing, 2014, 32 (10): 692-706.

[27] MAVADATI S M, MAHOOR M H, BARTLETT K, et al. Disfa: A spontaneous facial action intensity database [J]. IEEE Transactions on Affective Computing, 2013, 4 (2): 151-160.

[28] BENITEZ-QUIROZ C F, SRINIVASAN R, MARTINEZ A M. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 5562-5570.

[29] KOLLIAS D, ZAFEIRIOU S., Expression affect, action unit recognition: Aff-wild2, multi-task learning and arcface[C]//Proceedings of the British Machine Vision Con-

ference. Cardiff: BMVA Press, 2019: 297.

[30] LUCEY P, COHN J F, KANADE T, et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. San Francisco: IEEE, 2010: 94-101.

[31] VALSTAR M, PANTIC M. Induced disgust, happiness and surprise: An addition to the mmi facial expression database[C]//Proceedings of the International Conference on Language Resources and Evaluation Workshops. Valletta: ELRA 2010: 65-70.

[32] SAVRAN A, ALYÜZ N, DIBEKLIOĞLU H, et al. Bosphorus database for 3D face analysis[C]//Proceedings of the European Workshop on Biometrics and Identity Management. Roskilde: Springer, 2008: 47-56.

[33] STRATOU G, GHOSH A, DEBEVEC P, et al. Effect of illumination on automatic expression recognition: A novel 3D relightable facial database[C]//Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. Santa Barbara: IEEE, 2011: 611-618.

[34] COSKER D, KRUMHUBER E, HILTON A. A facs valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling[C]//Proceedings of the IEEE International Conference on Computer Vision. Barcelona: IEEE, 2011: 2296-2303.

[35] ZHANG Z, GIRARD J M, WU Y, et al. Multimodal spontaneous emotion corpus for human behavior analysis [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 3438-3446.

[36] GIRARD J M, CHU W S, JENI L A,et al. Sayette group formation task (gft) spontaneous facial expression database[C]//Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017: 581-588.

[37] YAN W J, LI X, WANG S J, ZHAO G, et al. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation[J]. PloS One, 2014, 9(1): e86041.

[38] DAVISON A K, LANSLEY C, COSTEN N, et al. Samm: A spontaneous micro-facial movement dataset[J]. IEEE Transactions on Affective Computing, 2018, 9(1): 116-129.

[39] BEN X, REN Y, ZHANG J, et al. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms[EB/OL]. (2021-03-19)[2022-05-09]. https://doi.org/10.1109/TPAMI.2021.3067464.

[40] ZHOU Y, PI J, SHI B E. Pose-independent facial action unit intensity regression based on multi-task deep transfer learning[C]/Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017: 872-877.

[41] RUSSAKOVSKY O, DENG J, SU H, et al. Imagenet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[42] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]//Proceedings of the International Conference on Learning Representations. San Diego: OpenReview, 2015: 1-14.

[43] JI S, WANG K, PENG X, et al. Multiple transfer learning and multi-label balanced training strategies for facial AU detection in the wild[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020: 1657-1661.

[44] 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. Las Vegas: IEEE, 2016: 770-778.

[45] WERNER P, SAXEN F, AL-HAMADI A. Facial action unit recognition in the wild with multi-task cnn self-training for the emotionet challenge[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2020: 410-411.

[46] PENG G, WANG S. Weakly supervised facial action unit recognition through adversarial training[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2188-2196.

[47] PENG G, WANG S. Dual semi-supervised learning for facial action unit recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI Press, 2019: 8827-8834.

[48] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Proceedings of the Advances in Neural Information Processing Systems. Montreal: MIT Press, 2014: 2672-2680.

[49] ZHANG Y, DONG W, HU B G, et al. Classifier learning with prior probabilities for facial action unit recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5108-5116.

[50] WU S, WANG S, PAN B, et al. Deep facial action unit

recognition from partially labeled data[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3951-3959.

[51] ZHANG Y, DONG W, HU B G, et al. Weakly-supervised deep convolutional neural network learning for facial action unit intensity estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 2314-2323.

[52] SHAO Z, LIU Z, CAI J, et al. Jâa-net: Joint facial action unit detection and face alignment via adaptive attention [J]. International Journal of Computer Vision, 2021, 129 (2): 321-340.

[53] JYOTI S, SHARMA G, DHALL A. Expression empowered residen network for facial action unit detection[C]// Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Lille: IEEE, 2019: 1-8.

[54] TU C H, YANG C Y, HSU J Y. Idennet: Identity-aware facial action unit detection[C]//Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Lille: IEEE, 2019: 1-8.

[55] LIU Z, SONG G, CAI J, et al. Conditional adversarial synthesis of 3D facial action units[J]. Neurocomputing, 2019, 355: 200-208.

[56] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-06) [2021-05-18]. https://arxiv.org/abs/1411.1784.

[57] BLANZ V, VETTER T. A morphable model for the synthesis of 3D faces[C]//Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques of SIGGRAPH. Los Angeles: ACM, 1999: 187-194.

[58] WANG C, WANG S. Personalized multiple facial action unit recognition through generative adversarial recognition network[C]//Proceedings of the ACM International Conference on Multimedia. Seoul: ACM, 2018: 302-310.

[59] WILES O, KOEPKE A S, ZISSERMAN A. Self-supervised learning of a facial attribute embedding from video [C]//Proceedings of the British Machine Vision Conference. Newcastle: BMVA Press, 2018: 302.

[60] LI Y, ZENG J, SHAN S. Learning representations for facial actions from unlabeled videos[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44 (1): 302-317.

[61] LI W, ABTAHI F, ZHU Z, et al. Eac-net: Deep nets with enhancing and cropping for facial action unit detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2583-2596.

[62] JAISWAL S, VALSTAR M. Deep learning the dynamic appearance and shape of facial action units[C]//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Lake Placid: IEEE, 2016: 1-8.

[63] ALI A M, ALKABBANY I, FARAG A, et al. Facial action units detection under pose variations using deep regions learning[C]//Proceedings of the International Conference on Affective Computing and Intelligent Interaction. San Antonio: IEEE, 2017: 395-400.

[64] MA C, CHEN L, YONG J. AU R-CNN: Encoding expert prior knowledge into R-CNN for action unit detection[J]. Neurocomputing, 2019, 355: 35-47.

[65] LI W, ABTAHI F, ZHU Z. Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6766-6775.

[66] SANCHEZ E, TZIMIROPOULOS G, VALSTAR M. Joint action unit localisation and intensity estimation through heatmap regression[C]//Proceedings of the British Machine Vision Conference. Newcastle: BMVA Press, 2018: 233.

[67] LIU M, LI S, SHAN S, et al. AU-aware deep Networks for facial expression recognition[C]//Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. Shanghai: IEEE, 2013: 1-6.

[68] ZHAO K, CHU W S, ZHANG H. Deep region and multi-label learning for facial action unit detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 3391-3399.

[69] HAN S, MENG Z, LI Z, et al. Optimizing filter size in convolutional neural networks for facial action unit recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 5070-5078.

[70] ERTUGRUL I O, JENI L A, COHN J F. Pattnet: Patch-attentive deep network for action unit detection[C]//Proceedings of the British Machine Vision Conference. Cardiff: BMVA Press, 2019: 114.1-114.13.

[71] ERTUGRUL I O, YANG L, JENI L A, et al. D-pattnet: Dynamic patch-attentive deep network for action unit detection[J]. Frontiers in Computer Science, 2019, 1(11): 1-13.

[72] NIU X, HAN H, YANG S, et al. Local relationship learn-

ing with person-specific shape regularization for facial action unit detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 11917-11926.

[73] FAN Y, LIN Z. G2rl: Geometry-guided representation learning for facial action unit intensity estimation[C]//Proceedings of the International Joint Conference on Artificial Intelligence. Virtual Conference: IJCAI, 2020: 731-737.

[74] TRAN D L, WALECKI R, RUDOVIC O, et al. Deepcoder: Semi-parametric variational autoencoders for automatic facial action coding[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3190-3199.

[75] BENITEZ-QUIROZ C F, WANG Y, MARTINEZ A M. Recognition of action units in the wild with deep nets and a new global-local Loss[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 3970-3979.

[76] WALECKI R, RUDOVIC O, PAVLOVIC V, et al. Deep structured learning for facial action unit intensity estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3405-3414.

[77] CORNEANU C A, MADADI M, ESCALERA S. Deep structure inference network for facial action unit recognition[C]//Proceedings of the European Conference on Computer Vision. Munich: Springer, 2018: 309-324.

[78] JACOB G M, STENGER B. Facial action unit detection with transformers[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2021: 7680-7689.

[79] LI G, ZHU X, ZENG Y, et al. Semantic relationships guided representation learning for facial action unit recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019: 8594-8601.

[80] LI Y, TARLOW D, BROCKSCHMIDT M, et al. Gated graph sequence neural networks[C]//Proceedings of the International Conference on Learning Representations. San Juan: OpenReview, 2016: 1-16.

[81] LIU Z, DONG J, ZHANG C, et al. Relation modeling with graph convolutional networks for facial action unit detection[C]//Proceedings of the International Conference on Multimedia Modeling. Daejeon: Springer, 2020: 489-501.

[82] NIU X, HAN H, SHAN S, et al. Multi-label co-regularization for semi-supervised facial action unit recognition[C]//Proceedings of the Advances in Neural Information Processing Systems. Vancouver: Curran Associates 2019: 909-919.

[83] FAN Y, LAM J C K, LI V O K. Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020: 12701-12708.

[84] SONG T, CHEN L, ZHENG W, et al. Uncertain graph neural networks for facial action unit detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Conference: AAAI, 2021: 5993-6001.

[85] SONG T, CUI Z, ZHENG W, et al. Hybrid message passing with performance-driven structures for facial action unit detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2021: 6267-6276.

[86] CHU W S, DE LA TORRE F, COHN J F. Learning spatial and temporal cues for multi-label facial action unit detection[C]//Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017: 25-32.

[87] BISHAY M, PATRAS I. Fusing multilabel deep networks for facial action unit detection[C]//Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017: 681-688.

[88] HE J, LI D, YANG B, et al. Multi view facial action unit detection based on CNN and BLSTM-RNN[C]//Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Washington: IEEE, 2017: 848-853.

[89] SONG T, CUI Z, WANG Y, et al. Dynamic probabilistic graph convolution for facial action unit intensity estimation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual Conference: IEEE, 2021: 4845-4854.

[90] YANG L, ERTUGRUL I O, COHN J F, et al. FACS3D-NET: 3D convolution based spatiotemporal representation for action unit detection[C]//Proceedings of the International Conference on Affective Computing and Intelligent Interaction. Cambridge: IEEE, 2019: 538-544.

[91] YANG H, YIN L. Learning temporal information from a single image for au detection[C]//Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition. Lille: IEEE, 2019: 1-8.

[92] ZHANG Y, JIANG H, WU B, et al. Context-aware feature and label fusion for facial action unit intensity estimation with partially labeled data[C]//Proceedings of the IEEE International Conference on Computer Vision. Seoul: IEEE, 2019: 733-742.

[93] SHROUT P E, FLEISS J L. Intraclass correlations: Uses in assessing rater reliability[J]. Psychological Bulletin, 1979, 86(2): 420-428.

[94] LIN L, WANG K, MENG D, et al. Active self-paced learning for cost-effective and progressive face identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(1): 7-19.

[95] 胡小娟, 刘磊, 邱宁佳. 基于主动学习和否定选择的垃圾邮件分类算法[J]. 电子学报, 2018, 46(1): 203-209.
HU X J, LIU L, QIU N J. A novel spam categorization algorithm based on active learning method and negative selection algorithm[J]. Acta Electronica Sinica, 2018, 46(1): 203-209. (in Chinese)

[96] 姚拓中, 安鹏, 宋加涛. 基于历史分类加权和分级竞争采样的多视角主动学习[J]. 电子学报, 2017, 45(1): 46-53.
YAO T Z, AN P, SONG J T. Multi-view active learning based on weighted hypothesis boosting and hierarchical competition sampling[J]. Acta Electronica Sinica, 2017, 45(1): 46-53. (in Chinese)

作者简介

邵志文　男,1994年生,安徽马鞍山人. 2020年获得上海交通大学博士学位. 现为中国矿业大学计算机科学与技术学院准聘副教授. 曾主持国家自然科学基金青年项目、江苏省“双创博士”项目、中央高校基本科研业务费青年项目等. 研究方向为计算机视觉和深度学习,涵盖人脸表情识别、人脸表情合成、人脸配准等领域.
E-mail: zhiwen_shao@cumt.edu.cn

周　勇(通讯作者)　男,1974年生,江苏徐州人. 现为中国矿业大学计算机科学与技术学院教授、博士生导师. 曾主持国家自然科学基金面上项目、国家863计划子课题、江苏省“333人才工程”和“六大人才高峰”项目等. 研究方向为数据挖掘、机器学习和人工智能.
E-mail: yzhou@cumt.edu.cn