E-mail: jig@radi.ac.cn
Website: www.cjig.cn
Tel: 010-58887035

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
©中国图象图形学报版权所有

中图法分类号: TP391.41　**文献标识码:** A　**文章编号:** 1006-8961(2020)11-2306-15

论文引用格式: Li S and Deng W H. 2020. Deep facial expression recognition: a survey. Journal of Image and Graphics,25(11):2306-2320(李珊,邓伟洪. 2020. 深度人脸表情识别研究进展. 中国图象图形学报,25(11):2306-2320)[DOI: 10.11834/jig.200233]

深度人脸表情识别研究进展

李珊,邓伟洪

北京邮电大学人工智能学院,北京　100876

摘　要: 　随着人脸表情识别任务逐渐从实验室受控环境转移至具有挑战性的真实世界环境,在深度学习技术的迅猛发展下,深度神经网络能够学习出具有判别能力的特征,逐渐应用于自动人脸表情识别任务。目前的深度人脸表情识别系统致力于解决以下两个问题: 1) 由于缺乏足量训练数据导致的过拟合问题; 2) 真实世界环境下其他与表情无关因素变量(例如光照、头部姿态和身份特征) 带来的干扰问题。本文首先对近十年深度人脸表情识别方法的研究现状以及相关人脸表情数据库的发展进行概括。然后,将目前基于深度学习的人脸表情识别方法分为两类: 静态人脸表情识别和动态人脸表情识别,并对这两类方法分别进行介绍和综述。针对目前领域内先进的深度表情识别算法,对其在常见表情数据库上的性能进行了对比并详细分析了各类算法的优缺点。最后本文对该领域的未来研究方向和机遇挑战进行了总结和展望: 考虑到表情本质上是面部肌肉运动的动态活动,基于动态序列的深度表情识别网络往往能够取得比静态表情识别网络更好的识别效果。此外,结合其他表情模型如面部动作单元模型以及其他多媒体模态,如音频模态和人体生理信息能够将表情识别拓展到更具有实际应用价值的场景。

关键词: 人脸表情识别(FER); 真实世界; 深度学习; 综述

Deep facial expression recognition: a survey

Li Shan, Deng Weihong

School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Bejing 100876, China

Abstract: Facial expression is a powerful, natural, and universal signal for human beings to convey their emotional states and intentions. Numerous studies have been conducted on automatic facial expression analysis because of its practical importance in sociable robotics, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems. Various facial expression recognition (FER) systems have been explored to encode expression information from facial representations in the field of computer vision and machine learning. Traditional methods typically use handcrafted features or shallow learning for FER. However, related studies have collected training samples from challenging real-world scenarios, which implicitly promote the transition of FER from laboratory-controlled to in-the-wild settings since 2013. Meanwhile, studies in various fields have increasingly used deep learning methods, which achieve state-of-the-art recognition accuracy and remarkably exceed the results of previous investigations due to considerably improved chip processing abilities (e. g., GPU units) and appropriately designed network architectures. Moreover, deep learning techniques are increasingly utilized to handle challenging factors for emotion recognition in the wild because of the effective training of facial expression data. The transition of facial expression recognition from being laboratory-controlled to challenging in-the-wild conditions and the recent success of deep learning techniques in various fields have promoted the use of deep neural

收稿日期: 2020-05-31; **修回日期:** 2020-09-07; **预印本日期:** 2020-09-14

基金项目: 国家自然科学基金项目(61871052); 国家重点研发计划项目(2019YFB1406504)

Supported by: National Natural Science Foundation of China (61871052); National Key Research and Development Program of China (2019YFB1406504)

networks to learn discriminative representations for automatic FER. Recent deep FER systems generally focus on the following important issues. 1) Deep neural networks require a large amount of training data to avoid overfitting. However, existing facial expression databases are insufficient for training common neural networks with deep architecture, which achieve promising results in object recognition tasks. 2) Expression-unrelated variations are common in unconstrained facial expression scenarios, such as illumination, head pose, and identity bias. These disturbances are nonlinearly confounded with facial expressions and therefore strengthen the requirement of deep networks to address the large intraclass variability and learn effective expression-specific representations. We provide a comprehensive review of deep FER, including datasets and algorithms that provide insights into these intrinsic problems, in this survey. First, we introduce the background of fields of FER and summarize the development of available datasets widely used in the literature as well as FER algorithms in the past 10 years. Second, we divide the FER system into two main categories according to feature representations, namely, static image and dynamic sequence FER. The feature representation in static-based methods is encoded with only spatial information from the current single image, whereas dynamic-based methods consider temporal relations among contiguous frames in input facial expression sequences. On the basis of these two vision-based methods, other modalities, such as audio and physiological channels, have also been used in multimodal sentiment analysis systems to assist in FER. Although pure expression recognition based on visible face images can achieve promising results, incorporating it with other models into a high-level framework can provide complementary information and further enhance the robustness. We introduce existing novel deep neural networks and related training strategies, which are designed for FER based on both static and dynamic image sequences, and discuss their advantages and limitations in state-of-the-art deep FER. Competitive performance and experimental comparisons of these deep FER systems in widely used benchmarks are also summarized. We then discuss relative advantages and disadvantages of these different types of methods with respect to two open issues (data size requirement and expression-unrelated variations) and other focuses (computation efficiency, performance, and network training difficulty). Finally, we review and summarize the following challenges in this field and future directions for the design of robust deep FER systems. 1) Lacking training data in terms of both quantity and quality is a main challenge in deep FER systems. Abundant sample images with diverse head poses and occlusions as well as precise face attribute labels, including expression, age, gender, and ethnicity, are crucial for practical applications. The crowdsourcing model under the guidance of expert annotators is a reasonable approach for massive annotations. 2) Data bias and inconsistent annotations are very common among different facial expression datasets due to various collecting conditions and the subjectiveness of annotating. Furthermore, the FER performance fails to improve when training data is enlarged by directly merging multiple datasets due to inconsistent expression annotations. Cross-database performance is an important evaluation criterion of generalizability and practicability of FER systems. Deep domain adaption and knowledge distillation are promising trends to address this bias. 3) Another common issue is imbalanced class distribution in facial expression due to the practicality of sample acquirement. One solution is to resample and balance the class distribution on the basis of the number of samples for each class during the preprocessing stage using data augmentation and synthesis. Another alternative is to develop a cost-sensitive loss layer for reweighting during network work training. 4) Although FER within the categorical model has been extensively investigated, the definition of prototypical expressions covers only a small portion of specific categories and cannot capture the full repertoire of expressive behavior for realistic interactions. Incorporating other affective models, such as FACS (facial action coding system) and dimensional models, can facilitate the recognition of facial expressions and allow them to learn expression-discriminative representations. 5) Human expressive behavior in realistic applications involves encoding from different perspectives, with facial expressions as only one modality. Although pure expression recognition based on visible face images can achieve promising results, incorporating it with other models into a high-level framework can provide complementary information and further enhance the robustness. For example, the fusion of other modalities, such as the audio information, infrared images, and depth information from 3D face models and physiological data, has become a promising research direction due to the large complementarity of facial expressions and the good application value of human-computer interaction (HCI) applications.

Key words: facial expression recognition (FER); real world; deep learning; survey

0 引言

面部表情是人类表达情感和意图最为有效和普遍的方式之一(Darwin 和 Prodger,1998; Tian 等,2001)。自动人脸表情分析在日常生活中有众多应用,例如社交机器人、医疗服务、疲劳驾驶检测以及其他人机交互系统等。达尔文进化论发现人类脸部的丰富表情是自然选择的结果。在进化初期,古人类具有恐惧等简单表情,它协助增大了瞳孔的通光量,帮助他们在危险时成功逃脱; 在社交活动中,人们逐步进化出了微笑、内疚等复杂面部动作来表达内心情感。到了现代,面部肌肉可以组合出成百上千种动作,各种族进化出各异的情感表达方式。著名心理学家 Ekman 发现,有 6 类基本表情是全球通用的,各种族间可以相互识别,甚至连与世隔绝的部落和哺乳动物都有着相似的表情。据此,Ekman 和 Friesen(1971) 通过大量的跨文化研究(Ekman,1994) 定义了 6 类基础表情: 生气、厌恶、害怕、开心、悲伤和惊讶。由此,学术界普遍通过这 6 类基础表情的分类研究开始计算机自动表情识别的探索。

根据特征表示的不同,人脸表情识别系统主要分为两大类: 静态图片人脸表情识别和动态序列人脸表情识别。在静态识别方法中,仅有单幅图像的空间信息被编码于特征表示中。而在动态识别方法中,输入的人脸表情序列中连续帧之间的时间关系也被纳入考虑范围内。

传统方法大多运用手工设计特征或者浅层学习,例如局部二值模式(local binary pattern,LBP)(Shan 等,2009)、三正交平面的局部二值模式(local binary pattern from three orthogonal planes,LBP-TOP)(Zhao 和 Pietikainen,2007)、非负矩阵分解(nonnegative matrix factorization,NMF)(Zhi 等,2011) 和稀疏学习(Zhong 等,2012) 来进行人脸表情识别。2013 年起,表情识别比赛如 FER2013(the Facial Expression Recognition 2013)(Goodfellow 等,2013) 和 EmotiW(Dhall 等,2015,2016,2017) 从具有挑战性的真实世界场景中收集了相对充足的训练样本,促进了人脸表情识别从实验室受控环境到自然环境下的转换。从研究对象来看,表情识别领域正经历着从实验室摆拍到真实世界的自发表达、从长时间持续的夸张表情到瞬时出现的微表情、从基础表情分类到复杂表情分析的快速发展。

同时,由于芯片处理能力(如 GPU 单元) 的迅猛增长和网络体系结构的精心设计,各个领域的研究开始转向利用深度学习解决各种问题,并且也取得了远超先前方法的识别结果(Krizhevsky 等,2012; Simonyan 和 Zisserman,2014b; Szegedy 等,2015)。同样地,深度学习技术也逐渐运用到人脸表情识别中处理各种复杂的干扰因素。图 1 展示了自 2007 年以来人脸表情数据库以及表情识别算法的发展情况,如 DAE(deep autoencoder)、LP(locality-preserving) loss 和 IACNN(identity-aware CNN) 等。

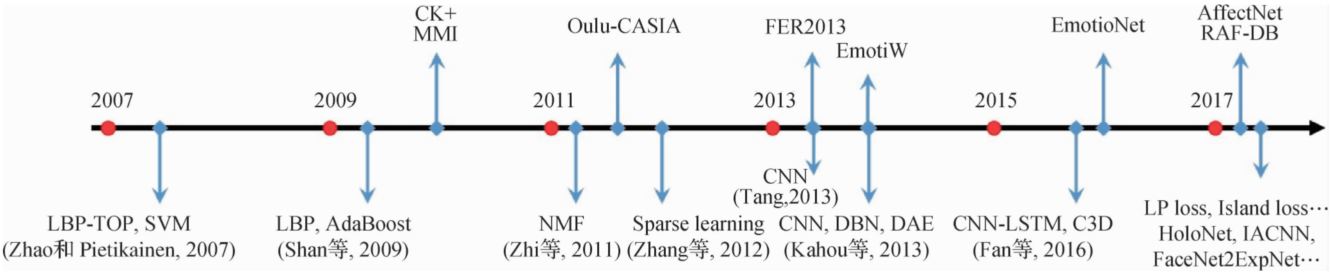

图 1 人脸表情识别算法和数据库发展图

Fig. 1 The evolution of facial expression recognition and facial expression dataset

尽管深度学习有着强大的特征表示能力,其在进行人脸表情识别时仍存在一些问题。首先,深度网络需要大量充足的训练数据来避免过拟合问题。然而,现存的人脸表情数据量不足以很好地训练目前在物体识别任务中取得良好效果的较大深度网络结构。其次,由于人物属性的不同,例如年龄、性别、种族背景和表达能力水平(Valstar 等,2012),不同对象间存在较大的差异性。除此之外,姿态、光照和阻挡等变量在不受约束的人脸表情场景中也十分常见。这些因素通常非线性地与人脸表情耦合在一起,从而加强了对利用深度网络来解决较大类内差距以及学习具有高效表情判别能力特征的需求。

1　深度人脸表情识别的研究现状

根据所处理数据类型的不同,深度人脸表情识别方法大致可以分为两大类: 基于静态图像的深度人脸表情识别网络和基于动态图像序列的深度人脸表情识别网络。图 2 展示了深度人脸表情识别系统分类的基本框架。本文将对现存的新颖人脸表情识别方法以及相关的网络训练技巧进行归纳总结。

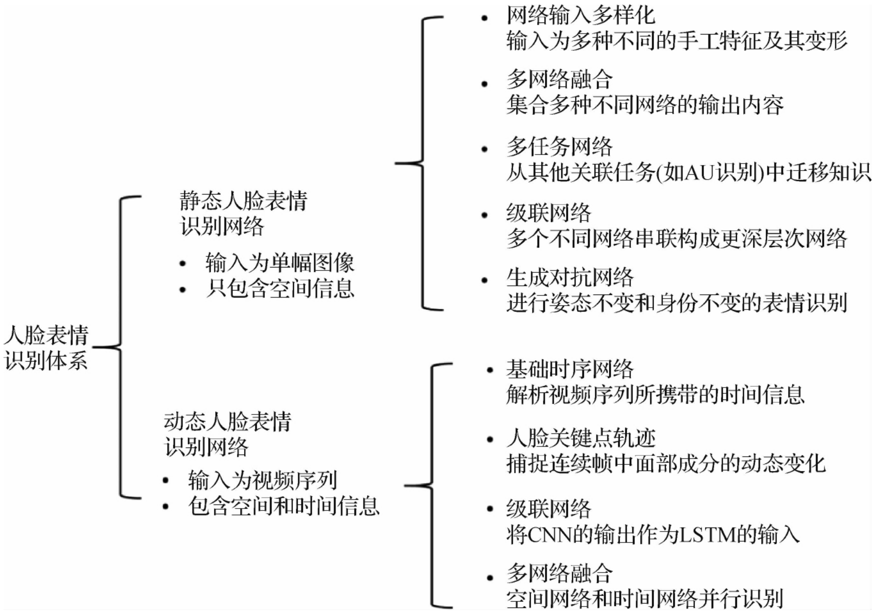
人脸表情识别体系
静态人脸表情识别网络
• 输入为单幅图像
• 只包含空间信息
• 网络输入多样化
输入为多种不同的手工特征及其变形
• 多网络融合
集合多种不同网络的输出内容
• 多任务网络
从其他关联任务(如AU识别)中迁移知识
• 级联网络
多个不同网络串联构成更深层次网络
• 生成对抗网络
进行姿态不变和身份不变的表情识别
动态人脸表情识别网络
• 输入为视频序列
• 包含空间和时间信息
• 基础时序网络
解析视频序列所携带的时间信息
• 人脸关键点轨迹
捕捉连续帧中面部成分的动态变化
• 级联网络
将CNN的输出作为LSTM的输入
• 多网络融合
空间网络和时间网络并行识别

图 2　人脸表情识别系统分类

Fig. 2　Different types of facial expression recognition system

1.1　基于静态图像的深度人脸表情识别网络

由于静态数据处理的便利性及其可得性,目前大量研究是基于不考虑时间信息的静态图像进行表情识别。

直接在相对较小的人脸表情数据库上进行深度网络的训练势必会导致过拟合问题。为了缓解这一问题,许多相关研究采用额外的辅助数据来从头预训练并自建网络,或者直接基于有效的预训练网络例如 AlexNet (Krizhevsky 等,2012), VGG (visual geometry group) (Simonyan 和 Zisserman,2014b), VGG-Face (Parkhi 等,2015) 和 GoogLeNet (Szegedy 等,2015) 进行微调。

大型人脸识别数据库如 CASIA WebFace (Yi 等,2014), CFW (Celebrity Face in the Wild) (Zhang 等,2012) 和 FaceScrub dataset (Ng 和 Winkler,2014),以及相对较大的人脸表情数据库例如 FER2013 (Goodfellow 等,2013) 和 TFD (The Toronto face database) 是较为合适的辅助训练数据。Kaya 等人(2017) 指出在人脸数据上进行预训练的 VGG-Face 模型比在物体数据上预训练的 ImageNet 模型更加适合于人脸表情识别任务。Knyazev 等人(2017) 也指出在大型的人脸数据库上进行预训练,然后进一步在额外的表情数据库上进行微调,能够有效地提高表情识别率。

除此之外,Ng 等人(2015) 提出了一个多阶段微调策略: 第 1 阶段利用额外的人脸表情数据库 FER2013 在已有的预训练模型上进行微调; 第 2 阶段则利用目标数据库(EmotiW) 的已知训练集来微调模型,使其更加适应于目标数据库。Ding 等人(2017) 提出 FaceNet2ExpNet 框架来排除预训练模型中所保留的人脸信息对表情识别任务带来的干扰。

1.1.1　多样化的网络输入

传统方法通常将整个对齐后人脸对应的 RGB 图像作为网络输入。然而,这类未经处理的数据缺乏对重要信息的提取,比如同质或常规的纹理信息以及图像缩放、旋转、阻挡和光照不变性。因此,一些方法结合多种不同的手工特征及其变形作为网络的输入来避免这个问题。

低阶特征表示将给定的 RGB 图像分为多个小块依次进行特征编码,然后对这些带有局部直方图

的特征进行聚类和池化,从而对光照变化和细微的人脸配准误差具有鲁棒性。Levi 和 Hassner(2015)提出映射的 LBP 特征用于光照不变性的表情识别。Zhang 等人(2016)则采用了尺度不变特征转换(scale-invariant feature transform,SIFT)特征(Lowe,1999)来进行多姿态的人脸表情识别。其他相关研究(Zeng 等,2018; Luo 等,2017)进一步从轮廓、纹理、角度和颜色几方面来结合不同的描述算子提高深度网络的性能。

基于局部的特征表示根据目标任务的不同来提取不同特征,从而将不重要的部分从整幅图像中剔除,同时挖掘出对目标任务较为敏感的关键部分。(Chen 等,2018a)指出 3 类兴趣区(region of interest,ROI),即眉毛、眼睛和嘴巴,与人脸表情变化有着极强的关联,并将这 3 类区域作为网络的输入。其他相关研究(Mavani 等,2017; Wu 和 Lin,2018)则提出自动学习人脸表情关键区域。

1.1.2 多网络融合

先前的研究表明集合多种不同网络可以取得比单个网络更好的性能(Ciregan 等,2012)。在进行网络集成时,需要考虑以下两个关键因素: 1) 足够多样的子网络来保证互补性; 2) 合适的集合方法来高效地融合各种子网络。

针对第 1 个因素,采用不同类型的训练数据或者不同的网络参数和结构能够增强子网络的多样性。其中一些预处理方法(Kim 等,2016),例如变形和归一化,能够生成不同的数据从而训练出多样的子网络。另外,通过改变滤波器的大小、神经元的个数和网络的层数,以及对权重初始化实施不同的随机因子,子网络的多样性也能得到提高(Kim 等,2015; Pons 和 Masip,2018b)。

针对第 2 个因素,可以从两个不同的层面来集成子网络: 特征层面和决策层面。在特征层面,最常见的策略是将从不同子网络学习到的特征进行串联(Bargal 等,2016; Liu 等,2016)。而在决策层面,通常有 3 种集成方式: 多数表决、简单平均和权重平均。由于权重平均考虑了不同子网络的重要性和置信度区别,许多研究提出运用权重平均方法来寻找网络集成的最优权值集。Kahou 等人(2013)提出一个随机搜索方法对每一类表情进行预测结果的加权。Yu 和 Zhang(2015)运用对数似然损失和铰链损失来自适应地对每个网络分配不同的权重。Kim 等人(2015)提出了一种基于验证精度的指数加权平均来强调各种子网络。Pons 和 Masip(2018b)则利用卷积网络自主学习每个子模型的权重。

1.1.3 多任务网络

目前许多网络致力于单一的表情识别任务,并且在特征学习时没有考虑表情与其他潜在因素之间的相互作用。然而,在真实世界中人脸表情是和各种不同因素交织在一起的,例如头部姿态、光照和人物身份。考虑到这一点,相关研究引入多任务学习从其他关联任务中迁移知识来解耦干扰因素。Reed 等人(2014)构造了一个高阶玻尔兹曼机(disentangling Boltzmann machines,disBM)来学习与表情相关因素的流形空间坐标,同时使得表情相关隐藏单元具有人脸形态不变性。其他工作提出在人脸表情识别任务的同时进行其他任务,如人脸特征点定位(Devries 等,2014)、人脸面部肌肉单元检测(Pons 和 Masip,2018a)和人脸认证任务(Zhang 等,2017),能辅助提高表情识别性能。

1.1.4 级联网络

在级联网络中,各种子模块的输出依次串联,从而构成一个更深层次的网络。相关研究表明,在组合不同模块来学习特征的层次结构时,与表情无关的干扰因素能够逐渐被剔除。通常情况下,不同的子网络依次单独地结合在一起,每一个子网络都有着不同且层次分明的贡献。Lyu 等人(2014)提出的深度置信网络(deep belief network,DBN)被用于检测人脸和人脸中与表情相关的区域,然后这些被解析过的人脸区域被输入到一个堆栈自编码器(stacked autoencoder,SAE)中进行分类。Rifai 等人(2012)首先提出一个多尺度的压缩卷积网络(contractive convolutional network,CCNET)来获取局部转换不变性特征,然后将其输入到压缩自编码器(contractive autoencoder,CAE)用于从人物身份和姿态中分离出与表情无关的因素。除此之外,区别于简单地串联各种子网络,Liu 等人(2014)提出了一个提升深度置信网络(boosted deep belief network,BDBN),该网络在一个统一的循环状态中迭代地进行特征表示、特征选择和分类器构造。相比较之前所提及的无反馈式串联,该循环框架能够反向传播分类误差用于交替地初始化特征选择直至收敛。由此,表情识别的判别能力能够在该迭代过程中得到

很好的提高。

1.1.5 生成对抗网络

生成对抗网络(generative adversarial networks，GAN)成功运用于图像合成任务中来生成逼真的人脸、数字和其他各种各样的图像类型，从而有效地扩大了训练数据量。一些相关工作提出了基于GAN的方法来进行姿态不变的人脸表情识别以及身份不变的人脸表情识别。

针对姿态不变的人脸表情识别任务，Lai和Lai(2018)提出了一个基于GAN的人脸正面化框架，其中生成器在保留人物身份信息和表情特性的同时将输入图像正面化，判别器则用来区分真实图像和生成的正面化人脸图像。Zhang等人(2018a)则提出了一个GAN模型生成在任意姿态下的不同表情用于多角度的表情识别。针对身份不变的人脸表情识别，Yang等人(2018b)提出了一个包含两部分的身份适应生成模型，第1部分利用条件生成对抗网络(conditional GAN，cGAN)分别为每一个人生成不同的表情图像，第2部分依次为每一个人单独地进行人脸表情识别，从而避免了不同身份信息带来的干扰。Yang等人(2018a)提出了一个表情残差学习框架(de-expression residue learning，DeRL)来挖掘出与表情相关的信息，该信息在表情中立化阶段被过滤出但仍然保留在生成器内。然后该模型从生成器中直接提取出该信息用于减轻人物身份变量带来的干扰，从而提高表情识别率。

1.1.6 讨论

表1展示了目前有代表性的基于静态图像的深度人脸表情识别算法在常见数据库上的性能比较。接下来，本文对不同类别网络的优缺点进行讨论。多网络融合模型在特征层次或者决策层次融合了不同子网络的优点，并广泛应用于表情识别竞赛中来

表1 基于静态图像的深度人脸表情识别算法在常见数据库上的性能比较

Table 1 Performance summary of representative methods for static-based deep facial expression recognition on the most widely evaluated datasets

数据库	文献	网络类型	识别率/%
CK+(Lucey等，2010)	Ding等人(2017)	卷积网络微调	6类:(98.6) 8类:(96.8)
	Zeng等人(2018)	自编码网络	7类:95.79(93.78) 8类:89.84(86.82)
	Meng等人(2017)	多任务网络	7类:95.37(95.51)
	Liu等人(2017)	损失函数	7类:97.1(96.1)
	Yang等人(2018a)	对抗生成网络	7类:97.30(96.57)
	Zhang等人(2018b)	多任务网络	6类:98.9
MMI(Pantic等，2005)	Liu等人(2017)	损失函数	6类:78.53(73.50)
	Li等人(2017)	损失函数	6类:78.46
	Yang等人(2018a)	对抗生成网络	6类:73.23(72.67)
	Liu等人(2019)	损失函数	6类:81.13(79.33)
FER2013(Goodfellow等，2013)	Guo等人(2016)	损失函数	测试集:71.33
	Kim等人(2016)	多网络融合	测试集:73.73
	Georgescu等人(2019)	多网络融合	测试集:75.42
SFEW2.0(Dhall等，2015)	Li等人(2017)	损失函数	验证集:51.05
	Ding等人(2017)	微调	验证集:55.15(46.6)
	Liu等人(2017)	损失函数	验证集:54.19(47.97)
	Meng等人(2017)	多任务网络	验证集:50.98(42.57)

注:1)括号内的数值为平均精确度，由混淆矩阵对角线均值得出。2)7类:生气、蔑视、厌恶、害怕、开心、伤心和惊讶;8类:生气、蔑视、厌恶、害怕、开心、伤心、惊讶和中立。CK+为the extended Cohn-Kanade database，MMI为Maja Pantic，Michel Valstar Initiative。

提高最终的识别率。然而，设计不同类型用于互补的子网络也大大增加了模型计算量和储存空间。此外，由于不同子网络的权重通常是基于现有的训练集或者验证集学习得到的，这导致学习而来的参数易在测试集上过拟合。多任务网络则在训练表情识别任务的同时也考虑了其他与表情相关联的任务，例如面部特征点定位、面部肌肉单元检测和人脸识别，从而排除了与表情无关的干扰因素的影响。该方法的主要缺陷是其要求更多的与其他任务相关的标签参与训练，并使得训练量更大。级联网络则将不同子网络串联在一起，逐步加强了模型的判别能力。该方法能够有效避免过拟合问题并排除与表情无关因素的干扰。此外，生成对抗网络因其可生成高质量目标样本的优点也逐渐应用于表情识别领域中，进行姿态不变的表情识别或者增加训练样本的数量和多样性。

1.2 基于动态图像序列的深度人脸表情识别网络

一个输入序列中连续帧之间的时间相关信息有利于人脸表情识别。本节着重介绍基于动态序列的深度时空表情识别网络，该网络将一段时间窗口中一定范围内的帧作为一个单独的输入，考虑了视频序列中时空运动模态，同时运用空间信息和时间信息来捕捉出更加细微的表情。

1.2.1 循环神经网络和长短期记忆网络

由于连续数据的特征向量是相互依存地连接在一起的，循环神经网络（recurrent neural network，RNN）能够从该序列中捕捉到有效信息。长短期记忆网络（long short-term memory，LSTM）则能够在保证较低计算成本的同时处理可变长的序列数据。

1.2.2 3维卷积网络

与RNN相比，卷积神经网络（convolutional neural networks，CNN）更加适合于处理图像。其衍生3维卷积网络（3D convolutional neural network，C3D）则广泛运用于基于动态序列的人脸表情识别中。不同于传统CNN中的2维卷积核，C3D运用在时间轴具有相同权重的3维卷积核来捕捉时间信息。

1.2.3 人脸关键点轨迹

相关心理学研究表明，人脸表情是由某些面部部位（如眼睛、鼻子和嘴巴）的动态运动所激发的，这些部位包含了最具表情描述性的信息。为了在表情识别中获取更加准确的面部动作，相关工作提出了人脸关键点轨迹模型来捕捉连续帧中面部的动态变化。

提取关键点轨迹信息最直接的方法是将帧间的人脸关键点坐标在归一化后沿时间轴串联起来，从而为每一个输入序列生成一个1维的轨迹信号（Jung等，2015），或者构建一个类似图的向量作为CNN的输入（Yan等，2016）。除此之外，还可以利用连续帧中各关键点的相对距离变化来捕捉时间信息（Kim等，2017）。另外，基于局部信息的模型（Zhang等，2017）根据面部物理结构将人脸坐标划分为若干部分，然后将这些部分分层地输入到网络中，从而编码出同时具有局部低阶信息和全局高阶信息的特征。

1.2.4 级联网络

通过结合CNN强大的视觉感知表征能力和LSTM能够变长输入/输出的优点，Donahue等人（2015）提出了一个深度时空模型，该模型将CNN的输出（通常为全连接层）与LSTM进行级联，从而完成一系列涉及时变输入和输出的视觉任务。与该混合网络类似，许多级联网络被提出来用于人脸表情识别（Kim等，2019；Fan等，2016；Vielzeuf等，2017；Jain等，2017）。除了将LSTM与CNN的全连接层连接起来，Kankanamge等人（2017）提取CNN中最后一个卷积层的特征作为LSTM的输入，以在获得更大范围依赖的同时保证全局一致性。Ouyang等人（2017）则采用了一个更加灵活的ResNet-LSTM网络，它允许CNN中较低层的节点直接与LSTMs连接，从而捕捉时空信息。除了CNN外，Baccouche等人（2012）使用了卷积稀疏自编码器来学习稀疏和移位不变的特征；然后将其输入到LSTM进行分类。除了使用LSTM，Hasani和Mahoor（2017）使用条件随机场（conditional random fields，CRFs）模型来区分输入序列中的时间关系。

1.2.5 多网络融合

针对视频动作识别，Simonyan和Zisserman（2014a）提出了一个双流卷积网络，其中一个支流将多帧密集光流作为输入来捕捉时间信息，另一个支流直接输入静态图像来学习面部空间特征，然后将两个支流的输出融合在一起。受此架构的启发，一些工作提出了多网络融合模型用于人脸表情识别。Sun等人（2019）提出了一个多通道网络，该网络从人脸表情图像中提取空间信息，并利用一个序列中表情图像和中立图像之间的差值来提取空间信

息。Zhang 等人(2017)则融合了时间网络 PHRNN(part-based hierarchical bidirectional recurrent neural network)和空间网络 MSCNN(multisignal CNN)来同时提取局部/整体、几何/外观和静态/动态的表情特征。传统的融合方式是在各个网络独立训练完成后给网络的输出赋予不同的权重。与之不同，Jung 等人(2015)提出了一个联合微调的方式，该方法能够根据识别情况同时训练多个不同的子网络，从而取得更好的效果。

1.2.6 讨论

表 2 展示了目前有代表性的基于动态图像的深度人脸表情识别算法在常见数据库上的性能比较。接下来，对不同类别网络的优缺点进行讨论。RNN 及其扩展 LSTM 作为基本的时序网络结构广泛运用于视频序列的学习。然而其网络结构使其难以捕捉到有效的图像卷积特征。而 3 维卷积网络则能更好地学习图像特征，但其中的 3 维滤波结构往往只覆盖了短时间内的序列而忽略了长范围内的动态变化。人脸关键点轨迹则是依据人脸生理结构捕捉人脸形状特征在时间序列内的动态变化。该方法计算量小而且不受光照等无关因素的干扰。但是其对面部特征点定位精确度的要求比较高。级联网络则是首先提取出有表情判别能力的空间特征，然后将该信息依次输入到时序网络中进行时序信息的编码。

表 2　基于动态图像的深度人脸表情识别算法在常见数据库上的性能比较

Table 2　Performances of representative methods for dynamic-based deep facial expression recognition on the most widely evaluated datasets

数据库	文献	网络类型	识别率/%
CK+(Lucey 等，2010)	Sun 等人(2019)	多网络融合	6 类：97.28
	Kumawat 等人(2019)	3 维卷积网络	7 类：97.38(96.65)
	Jung 等人(2015)	多网络融合	7 类： 97.25(95.22)
	Zhang 等人(2017)	多网络融合	7 类： 98.50(97.78)
MMI(Pantic 等，2005)	Hasani 和 Mahoor(2017)	人脸关键点轨迹	6 类：77.50(74.50)
	Zhang 等人(2017)	多网络融合	6 类：81.18(79.30)
	Wang 等人(2020)	人脸关键点轨迹	6 类：82.21
	Sun 等人(2019)	多网络融合	6 类：91.46
Oulu-CAISA(Zhao 等，2011)	Jung 等人(2015)	多网络融合	6 类：81.46(81.49)
	Kumawat 等人(2019)	3 维卷积网络	6 类：82.41(82.41)
	Zhang 等人(2017)	多网络融合	6 类：86.25(86.25)
AFEW 6.0(Dhall 等，2016)	Yan 等人(2016)	VGG16-LSTM	7 类：44.46
	Yan 等人(2016)	人脸关键点轨迹	7 类：37.37
	Fan 等人(2016)	VGG16-LSTM	验证集：45.43(38.96)
	Fan 等人(2016)	3 维卷积网络	验证集：39.69(38.55)
	Yan 等人(2016)	多模态融合	测试集：56.66(40.81)
	Fan 等人(2016)	多模态融合	测试集：59.02(44.94)
AFEW 7.0(Dhall 等，2017)	Ouyang 等人(2017)	VGG-LSTM	验证集：47.4
	Ouyang 等人(2017)	3 维卷积网络	验证集：35.2
	Vielzeuf 等人(2017)	VGG16-LSTM	验证集：48.6
	Vielzeuf 等人(2017)	多模态融合	测试集：58.81(43.23)

注：括号内的数值为平均精确度，由混淆矩阵对角线均值得出。Oulu-CAISA 为 Oulu-CAISA facial expression database。

而多网络融合则是同时训练两个子网络分别用于捕捉时序信息和空间信息，然后将其输出结果加权融合。

2 机遇和挑战

2.1 真实世界人脸表情数据库

考虑到人脸表情识别是一个依靠数据驱动的任务，训练一个足够深的网络来捕捉与表情相关的细微形变需要大量的相关数据。因此，在数量和质量上均较为匮乏的数据库是当今深度人脸表情识别系统面临的主要挑战。由于不同年龄段、不同种族和不同性别的人表达和解析面部表情的方式也不同，一个理想的表情数据集应该包含除了表情标签之外，各种丰富且精确的其他面部属性标签，例如年龄、性别和种族。除此之外，虽然面部遮挡和多姿态问题在深度人脸识别领域得到了广泛的研究，但其在深度人脸表情识别中受到的关注仍较少。主要原因是缺乏具有遮挡类型和头部姿态标注的大型面部表情数据集。另外，对大量携带复杂自然场景变化的数据进行精确标注的难度很大。一个可靠的解决方式是在专家的指导下对数据进行多人的众包标注。Benitez-Quiroz 等人（2016）提出了一个基于专家校准的工具用于表情的自动标注。2017 年，本课题组从 flicker 上下载了 3 万多幅表情丰富的人脸图像，每幅图像由 40 个不同的标注者进行了独立标注，利用鲁棒估计算法得到可靠的标签分布。该工作以前所未有的精度标注了互联网环境下的真实表情，组成了 RAF-DB（Real-world affective face database）（Li 等，2017；Li 和 Deng，2018）。此外，目前已有的真实世界大范围人脸表情的大型数据库还有 EmotiNet（Benitez-Quiroz 等，2016）和 AffectNet（Mollahosseini 等，2019），但由于它们采用单人标注，标签的主观性较大。随着技术的发展和互联网的大范围普及，期望能有越来越多的人脸表情数据集能被建立，但如何获得准确的表情标签仍然是一个值得研究的问题。

2.2 丰富的表情模型

虽然基于 7 类基本表情的分类模型在人脸表情识别领域得到了广泛的认同和研究，但是该原型表情并不能涵盖现实交互中所表达的全部情感行为。面部肌肉可以组合出上千种动作，基本表情只涵盖了小部分动作类别。其他模型则可以涵盖更大范围的表情类型：面部动作单元编码模型（facial action coding system，FACS）（Ekman 和 Rosenberg，1997）中不同面部肌肉相互结合用来描述表情的面部变化；持续时间较短的微表情对于分析人类的真实情感有独特的作用，中国科学院心理研究所在此方面做了大量开创性工作（Yan 等，2013，2014）。此外，维度模型（Gunes 和 Schuller，2013；Russell，1980）中两个连续值变量，即效价和唤醒度，能够连续地编码出表情强度的微小变化。其他新颖的定义还有 Du 等人（2014）提出的复合表情，该研究认为一些面部表情实际上是由多种基础表情组合而成。这些模型均在一定程度上改进了对面部表情特性的描述并且能够与分类模型进行互补。RAF-DB（Li 和 Deng，2019b）和 RAF-ML（Real-world affective face multiLabel）（Li 和 Deng，2019a）两个新数据库，包括 7 类基本表情、12 类复合表情和 30 余种混合表情。Lu 等人（2018）研究了混合表情的多标签向量空间与 Wordnet 语义空间的对齐，探索了更加丰富的表情标签描述。针对如何有效地利用这些模型，本文总结出以下思路：首先在设计网络参数时可以针对面部不同区域对表情的贡献值来赋予不同权重；其次，也可以基于注意力机制来强调与面部肌肉单元最相关的区域，从而使模型能够学到具有表情判别性的特征表示。

2.3 数据集偏差和不平衡分布

由于收集条件的不同和标注的主观性，数据偏差和不一致的标注问题在不同人脸表情数据库中也十分常见。研究者通常在一个确切的数据集内来评估算法，从而能够获得令人满意的性能。然而，最新的跨库实验表明（Li 和 Deng，2020a），由于不同数据库之间存在明显差异，通过在数据库内进行评估的算法往往缺乏对未知测试数据的普适性，其性能将会在跨库实验中明显恶化。深度领域自适应和知识蒸馏则是解决这一偏差问题的有效方法（Wei 等，2018；Li 和 Deng，2020a）。通常的做法是学习一个转换空间，使不同数据库在转换后的特征空间上分布的区分度尽可能相似。另一个常见的问题则是类别不平衡问题，该问题主要与数据采集过程中的实际情况有关：诱发并标注一个笑脸是十分容易的，但是捕捉厌恶、生气等其他更加不常见的表情则十分具有挑战性。针对这个问题，一种解决方案是在预

处理阶段使用数据扩充或者合成手段来平衡类别分布。另一种选择是在网络训练阶段设计一个代价敏感损失层，针对稀少类样本给予更大的权重来平衡常见类和稀少类所占比重。在一定的表情模型下，小样本和不平衡分类问题在表情识别任务中将长期存在，如何引入机器学习的新技术将是非常值得研究的课题。

2.4　多模态表情识别

在现实应用中人们有着多种情感表达方式，面部表情只是其中的一种模态。尽管基于可视人脸图像的表情识别能够取得不错的效果，但与其他模态结合到一个高层框架中能够提供互补信息，从而进一步增强模型的鲁棒性。例如，可将音频模态作为次重要的因素与图像信息相融合来进行多模态的情感识别。此外，红外图像、3 维人脸模型的深度信息、人体生理信息以及手势姿态也可以作为面部表情的互补数据来辅助情感识别。在脸部远程光电容积脉搏波(remote photoplethysmography，rPPG) 信号分析上的最新进展 RhythmNet (Niu 等，2019) ，也可能为表情分析带来新的模态。本文认为表情识别结合语音、文字甚至是脑电信号(Li 等，2019c) 的多模态表达问题是非常值得研究的问题，它将使机器可以读懂人类的内心，人机交互将变得更加自然流畅，疲劳驾驶监控、犯罪心理测试技术和自闭症医疗服务等实际应用可以得到落实。

最后，自监督面部动作识别方法(Li 等，2019b) 为解决表情标注难题开辟新的方向。基于局部快注意力的表情识别机制(Li 等，2019a) 提高了遮挡和多姿态条件下的识别鲁棒性。时空模式的融合对表情进行了动态描述，取得了较大突破(Wang 等，2020; Li 等，2019b) 。3 维等多模态信息(Chen 等，2018b) 有效地提高了识别的稳定性。此外，域自适应(Zheng 等，2018) 、小样本学习(Wang 等，2018) 和半监督学习(Liu 等，2020) 的进展也推动了表情识别的发展。Li 和 Deng(2020b) 更详细地介绍了表情识别领域的最新进展以及探讨未来的研究方向。

3　结　语

本文首先介绍了人脸表情识别的相关背景知识，并对表情识别领域自 2007 年以来数据库和算法的演化和发展进行了概述，指出利用深度学习进行表情识别已经成为该领域的主流框架。接着将基于深度学习的表情识别算法分为两大类(静态表情识别网络和动态表情识别网络) 进行了介绍和讨论。通过大量基于目前常见表情数据库的算法性能对比，发现动态表情网络往往能取得比静态表情网络更好的识别效果。考虑到表情实际上是处于动态变化过程的这一性质，而基于动态序列的深度表情识别算法能够很好地捕捉这一有效时序信息且能利用同一序列中人物身份信息这一特点，空间—时间深度网络将成为表情识别领域的一大趋势。最后，对该领域的机遇和挑战进行了总结并指出了未来研究方向。从数据集角度出发，收集大量多样的且具有准确表情标签的训练数据可以从根本上提高表情识别率。从算法角度而言，结合其他表情模型，例如面部动作单元模型和愉悦度—唤醒度维度模型，以及其他多模态，例如音频模态、3 维人脸深度信息和人体生理信息，可以使表情识别更具备实际运用价值。

参考文献(References)

Baccouche M，Mamalet F，Wolf C，Garcia C and Baskurt A. 2012. Spatio-temporal convolutional sparse auto-encoder for sequence classification//Proceedings of the British Machine Vision Conference. Surrey: BMVA Press: 1-12 [DOI: 10.5244/C.26.124]

Bargal S A，Barsoum E，Ferrer C C and Zhang C. 2016. Emotion recognition in the wild from videos using images//Proceedings of the 18th ACM International Conference on Multimodal Interaction. New York: ACM: 433-436 [DOI: 10.1145/2993148.2997627]

Benitez-Quiroz C F，Srinivasan R and Martinez A M. 2016. EmotioNet: an accurate，real-time algorithm for the automatic annotation of a million facial expressions in the wild//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE: 5562-5570 [DOI: 10.1109/CVPR.2016.600]

Chen L F，Zhou M T，Su W J，Wu M，She J H and Hirota K. 2018a. Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction. Information Sciences，428: 49-61 [DOI: 10.1016/j.ins.2017.10.044]

Chen Z X，Huang D，Wang Y H and Chen L M. 2018b. Fast and light manifold CNN based 3D facial expression recognition across pose variations//Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM: 229-238 [DOI: 10.1145/3240508.3240568]

Ciregan D，Meier U and Schmidhuber J. 2012. Multi-column deep neural networks for image classification//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE: 3642-3649 [DOI: 10.1109/CVPR.2012.6248110]

Darwin C and Prodger P. 1998. The Expression of the Emotions in Man and Animals. Oxford: Oxford University Press, 1998

Devries T, Biswaranjan K and Taylor G W. 2014. Multi-task learning of facial landmarks and expression//Proceedings of 2014 Canadian Conference on Computer and Robot Vision. Montreal: IEEE: 98-103 [DOI: 10.1109/CRV.2014.21]

Dhall A, Goecke R, Ghosh S, Joshi J, Hoey J and Gedeon T. 2017. From individual to group-level emotion recognition: EmotiW 5.0//Proceedings of the 19th ACM International Conference on Multimodal Interaction. New York: ACM: 524-528 [DOI: 10.1145/3136755.3143004]

Dhall A, Goecke R, Joshi J, Hoey J and Gedeon T. 2016. EmotiW 2016: video and group-level emotion recognition challenges//Proceedings of the 18th ACM International Conference on Multimodal Interaction. New York: ACM: 427-432 [DOI: 10.1145/2993148.2997638]

Dhall A, Murthy O V R, Goecke R, Joshi J and Gedeon T. 2015. Video and image based emotion recognition challenges in the wild: emotiW 2015//Proceedings of 2015 ACM on International Conference on Multimodal Interaction. New York: ACM: 423-426 [DOI: 10.1145/2818346.2829994]

Ding H, Zhou S K and Chellappa R. 2017. facenet2expnet: regularizing a deep face recognition net for expression recognition//Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition. Washington: IEEE: 118-126 [DOI: 10.1109/FG.2017.23]

Donahue J, Hendricks L A, Guadarrama S, Rohrbach M, Venugopalan S, Darrell T and Saenko K. 2015. Long-term recurrent convolutional networks for visual recognition and description//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 2625-2634 [DOI: 10.1109/CVPR.2015.7298878]

Du S C, Tao Y and Martinez A M. 2014. Compound facial expressions of emotion. Proceedings of the National Academy of Sciences of the United States of America, 111(15): E1454-E1462 [DOI: 10.1073/pnas.1322355111]

Ekman P. 1994. Strong evidence for universals in facial expressions: a reply to Russell´s mistaken critique. Psychological Bulletin, 115(2): 268-287 [DOI: 10.1037/0033-2909.115.2.268]

Ekman P and Friesen W V. 1971. Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17(2): 124-129 [DOI: 10.1037/h0030377]

Ekman P and Rosenberg E L. 1997. What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS). New York: Oxford University Press, 1997

Fan Y, Lu X J, Li D and Liu Y L. 2016. Video-based emotion recognition using CNN-RNN and C3D hybrid networks//Proceedings of the 18th ACM International Conference on Multimodal Interaction. New York: ACM: 445-450 [DOI: 10.1145/2993148.2997632]

Georgescu M I, Ionescu R T and Popescu M. 2019. Local learning with deep and handcrafted features for facial expression recognition. IEEE Access, 7: 64827-64836 [DOI: 10.1109/ACCESS.2019.2917266]

Goodfellow I J, Erhan D, Carrier P L, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y C, Thaler D, Lee D H, Zhou Y B, Ramaiah C, Feng FX, Li R F, Wang X J, Athanasakis D, Shawe-Taylor J, Milakov M, Park J, Ionescu R, Popescu M, Grozea C, Bergstra J, Xie J J, Romaszko L, Xu B, Chuang Z and Bengio Yoshua. 2013. Challenges in representation learning: a report on three machine learning contests//Proceedings of the 20th International Conference on Neural Information Processing. Daegu: Springer: 117-124 [DOI: 10.1007/978-3-642-42051-1_16]

Gunes H and Schuller B. 2013. Categorical and dimensional affect analysis in continuous input: current trends and future directions. Image and Vision Computing, 31(2): 120-136 [DOI: 10.1016/j.imavis.2012.06.016]

Guo Y N, Tao D P, Yu J, Xiong H, Li Y T and Tao D C. 2016. Deep neural networks with relativity learning for facial expression recognition//Proceedings of 2016 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Seattle: IEEE: 1-6 [DOI: 10.1109/ICMEW.2016.7574736]

Hasani B and Mahoor M H. 2017. Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields//Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition. Washington: IEEE: 790-795 [DOI: 10.1109/FG.2017.99]

Jain D K, Zhang Z and Huang K Q. 2017. Multi angle optimal pattern-based deep learning for automatic facial expression recognition. [EB/OL]. [2020-05-22]. https://www.sciencedirect.com/science/article/pii/S0167865517302313

Jung H, Lee S, Yim J, Park S and Kim J. 2015. Joint fine-tuning in deep neural networks for facial expression recognition//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE: 2983-2991 [DOI: 10.1109/ICCV.2015.341]

Kahou S E, Pal C, Bouthillier X, Froumenty P, Gülçehre Ç, Memisevic V, Vincent P, Courville A, Bengio Y, Ferrari R C, Mirza M, Jean S, Carrier P L, Dauphin Y, Boulanger-Lewandowski N, Aggarwal A, Zumer J, Lamblin P, Raymond J P, Desjardins G, Pascanu R, Warde-Farley D, Torabi A, Sharma A, Bengio E, Côte M, Konda K R and Wu Z Z. 2013. Combining modality specific deep neural networks for emotion recognition in video//Proceedings of the 15th ACM on International Conference on Multimodal Interaction. New York: ACM: 543-550 [DOI: 10.1145/2522848.2531745]

Kankanamge S, Fookes C and Sridharan S. 2017. Facial analysis in the wild with LSTM networks//Proceedings of 2017 IEEE International Conference on Image Processing. Beijing: IEEE: 1052-1056 [DOI: 10.1109/ICIP.2017.8296442]

Kaya H, Gürpınar F and Salah A A. 2017. Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image and Vision Computing, 65: 66-75 [DOI: 10.1016/j.imavis.2017.01.012]

Kim B K, Dong S Y, Roh J, Kim G and Lee S Y. 2016. Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas: IEEE: 1499-1508 [DOI: 10.1109/CVPRW.2016.187]

Kim B K, Lee H, Roh J and Lee S Y. 2015. Hierarchical committee of deep CNNs with exponentially-weighted decision fusion for static facial expression recognition//Proceedings of 2015 ACM on International Conference on Multimodal Interaction. New York: ACM: 427-434 [DOI: 10.1145/2818346.2830590]

Kim D H, Baddar W J, Jang J and Ro Y M. 2019. Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Transactions on Affective Computing, 10(2): 223-236 [DOI: 10.1109/TAFFC.2017.2695999]

Kim D H, Lee M K, Choi D Y and Song B C. 2017. Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild//Proceedings of the 19th ACM International Conference on Multimodal Interaction. New York: ACM: 529-535 [DOI: 10.1145/3136755.3143005]

Knyazev B, Shvetsov R, Efremova N and Kuharenko A. 2017. Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video. [EB/OL]. [2020-05-31] https://arxiv.org/pdf/1711.04598.pdf

Krizhevsky A, Sutskever I and Hinton G E. 2012. ImageNet classification with deep convolutional neural networks//Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook: ACM: 1097-1105

Kumawat S, Verma M and Raman S. 2019. LBVCNN: local binary volume convolutional neural network for facial expression recognition from image sequences//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Long Beach: IEEE: 207-216 [DOI: 10.1109/CVPRW.2019.00030]

Lai Y H and Lai S H. 2018. Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition//Proceedings of the 13th IEEE International Conference on Automatic Face and Gesture Recognition. Xi'an: IEEE: 263-270 [DOI: 10.1109/FG.2018.00046]

Levi G and Hassner T. 2015. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns//Proceedings of 2015 ACM on International Conference on Multimodal Interaction. New York: ACM: 503-510 [DOI: 10.1145/2818346.2830587]

Li S and Deng W H. 2018. Deep emotion transfer network for cross-database facial expression recognition//Proceedings of the 24th International Conference on Pattern Recognition. Beijing: IEEE: 3092-3099 [DOI: 10.1109/ICPR.2018.8545284]

Li S and Deng W H. 2019a. Blended emotion in-the-wild: multi-label facial expression recognition using crowdsourced annotations and deep locality feature learning. International Journal of Computer Vision, 127(6): 884-906 [DOI: 10.1007/s11263-018-1131-1]

Li S and Deng W H. 2019b. Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing, 28(1): 356-370 [DOI: 10.1109/TIP.2018.2868382]

Li S and Deng W H. 2020a. A deeper look at facial expression dataset bias. [EB/OL]. [2020-05-01]. https://arxiv.org/pdf/1904.11150.pdf

Li S and Deng W H. 2020b. Deep facial expression recognition: a survey. [EB/OL]. [2020-05-31]. https://ieeexplore.ieee.org/document/9039580

Li S, Deng W H and Du J P. 2017. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE: 2584-2593 [DOI: 10.1109/CVPR.2017.277]

Li Y, Zeng J B, Shan S G and Chen X L. 2019a. Occlusion aware facial expression recognition using CNN with attention mechanism. IEEE Transactions on Image Processing, 28(5): 2439-2450 [DOI: 10.1109/TIP.2018.2886767]

Li Y, Zeng J B, Shan S G and Chen X L. 2019b. Self-supervised representation learning from videos for facial action unit detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE: 10924-10933 [DOI: 10.1109/CVPR.2019.01118]

Li Y, Zheng W M, Wang L, Zong Y and Cuio Z. 2019c. From regional to global brain: a novel hierarchical spatial-temporal neural network model for EEG emotion recognition. IEEE Transactions on Affective Computing: #2922912 [DOI: 10.1109/TAFFC.2019.2922912]

Liu K, Zhang M M and Pan Z G. 2016. Facial expression recognition with CNN ensemble//Proceedings of 2016 International Conference on Cyberworlds. Chongqing: IEEE: 163-166 [DOI: 10.1109/CW.2016.34]

Liu P, Han S Z, Meng Z B, and Tong Y. 2014. Facial expression recognition via a boosted deep belief network//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE: 1805-1812 [DOI: 10.1109/CVPR.2014.233]

Liu P, Wei Y C, Meng Z B, Deng W H, Zhou J T and Yang Y. 2020. Omni-supervised facial expression recognition: a simple baseline. [EB/OL]. [2020-05-22]. https://arxiv.org/pdf/2005.08551.pdf

Liu X F, Kumar B V K V, Jia P and You J. 2019. Hard negative generation for identity-disentangled facial expression recognition. Pattern

Recognition, 88: 1-12 [DOI: 10.1016/j.patcog.2018.11.001]

Liu X F, Kumar B V K V, You J and Jia P. 2017. Adaptive deep metric learning for identity-aware facial expression recognition//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu: IEEE: 522-531 [DOI: 10.1109/CVPRW.2017.79]

Lowe D G. 1999. Object recognition from local scale-invariant features//Proceedings of the International Conference on Computer Vision. Washington: ACM: 1150-1157

Lu Z J, Zeng J B, Shan S G and Chen X L. 2018. Zero-shot facial expression recognition with multi-label label propagation//Proceedings of the 14th Asian Conference on Computer Vision. Perth: Springer: 19-34 [DOI: 10.1007/978-3-030-20893-6_2]

Lucey P, Cohn J F, Kanade T, Saragih J, Ambadar Z and Matthews I. 2010. The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. San Francisco: IEEE: 94-101 [DOI: 10.1109/CVPRW.2010.5543262]

Luo Z J, Chen J H, Takiguchi T and Ariki Y. 2017. Facial expression recognition with deep age//Proceedings of 2017 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). Hong Kong, China: IEEE: 657–662 [DOI: 10.1109/ICMEW.2017.8026251]

Lv Y D, Feng Z Y and Xu C. 2014. Facial expression recognition via deep learning//Proceedings of 2014 International Conference on Smart Computing. Hong Kong, China: IEEE: 303-308 [DOI: 10.1109/SMARTCOMP.2014.7043872]

Mavani V, Raman S and Miyapuram K P. 2017. Facial expression recognition using visual saliency and deep learning//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE: 2783-2788 [DOI: 10.1109/ICCVW.2017.327]

Meng Z B, Liu P, Cai J, Han S Z and Tong Y. 2017. Identity-Aware Convolutional Neural Network for Facial Expression Recognition//Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition. Washington: IEEE: 558-565 [DOI: 10.1109/FG.2017.140]

Mollahosseini A, Hasani B and Mahoor M H. 2019. AffectNet: a database for facial expression, valence, and arousal computing in the wild. IEEE Transactions on Affective Computing, 10(1): 18-31 [DOI: 10.1109/TAFFC.2017.2740923]

Ng H W, Nguyen V D, Vonikakis V and Winkler S. 2015. Deep learning for emotion recognition on small datasets using transfer learning//Proceedings of the 17th ACM International Conference on Multimodal Interaction. New York: ACM: 443-449 [DOI: 10.1145/2818346.2830593]

Ng H W and Winkler S. 2014. A data-driven approach to cleaning large face datasets//Proceedings of 2014 IEEE International Conference on Image Processing. Paris: IEEE: 343-347 [DOI: 10.1109/ICIP.2014.7025068]

Niu X S, Shan S G, Han H and Chen X L. 2019. RhythmNet: end-to-end heart rate estimation from face via spatial-temporal representation. IEEE Transactions on Image Processing, 29: 2409-2423 [DOI: 10.1109/TIP.2019.2947204]

Ouyang X, Kawaai S, Goh E G, Shen S M, Ding W, Ming H P and Huang D Y. 2017. Audio-visual emotion recognition using deep transfer learning and multiple temporal models//Proceedings of the 19th ACM International Conference on Multimodal Interaction. New York: ACM: 577-582 [DOI: 10.1145/3136755.3143012]

Pantic M, Valstar M, Rademaker R and Maat L. 2005. Web-based database for facial expression analysis//Proceedings of 2015 IEEE International Conference on Multimedia and Expo. Amsterdam: IEEE: #1521424 [DOI: 10.1109/ICME.2005.1521424]

Parkhi O M, Vedaldi A and Zisserman A. 2015. Deep face recognition//Proceedings of the British Machine Vision Conference. [s. l.]: BMVA Press: 41.1-41.12 [DOI: 10.5244/c.29.41]

Pons G and Masip D. 2018a. Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition. [EB/OL]. [2020-05-22]. https://arxiv.org/pdf/1802.06664.pdf

Pons G and Masip D. 2018b. Supervised committee of convolutional neural networks in automated facial expression analysis. IEEE Transactions on Affective Computing, 9(3): 343-350 [DOI: 10.1109/TAFFC.2017.2753235]

Reed S, Sohn K, Zhang Y and Lee H. 2014. Learning to disentangle factors of variation with manifold interaction//Proceedings of the 31st International Conference on International Conference on Machine Learning. New York: ACM: 1431-1439

Rifai S, Bengio Y, Courville A, Vincent P and Mirza M. 2012. Disentangling factors of variation for facial expression recognition//Proceedings of the 12th European Conference on Computer Vision. Florence: Springer: 808-822 [DOI: 10.1007/978-3-642-33783-3_58]

Russell J A. 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6): 1161-1178 [DOI: 10.1037/h0077714]

Shan C F, Gong S G and McOwan P W. 2009. Facial expression recognition based on local binary patterns: a comprehensive study. Image and Vision Computing, 27(6): 803-816 [DOI: 10.1016/j.imavis.2008.08.005]

Simonyan K and Zisserman A. 2014a. Two-stream convolutional networks for action recognition in videos//Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: ACM: 568-576

Simonyan K and Zisserman A. 2014b. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2020-05-22]. https://arxiv.org/pdf/1409.1556.pdf

Sun N, Li Q, Huan R Z, Liu J X and Han G. 2019. Deep spatial-temporal feature fusion for facial expression recognition in static images. Pattern Recognition Letters, 119: 49-61 [DOI: 10.1016/j.patrec.2017.10.022]

Szegedy C, Liu W, Jia Y Q, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A. 2015. Going deeper with convolutions//Proceedings of 2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE: 1-9 [DOI: 10.1109/CVPR.2015.7298594]

Tian Y I, Kanade T and Cohn J F. 2001. Recognizing action units for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2): 97-115 [DOI: 10.1109/34.908962]

Tang Y C. 2015. Deep learning using linear support vector machines. [EB/OL]. [2020-05-01]. https://arxiv.org/pdf/1306.0239.pdf

Valstar M F, Mehu M, Jiang B H, Pantic M and Scherer K. 2012. Meta-analysis of the first facial expression recognition challenge. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 42(4): 966-979 [DOI: 10.1109/TSMCB.2012.2200675]

Vielzeuf V, Pateux S and Jurie F. 2017. Temporal multimodal fusion for video emotion classification in the wild//Proceedings of the 19th ACM International Conference on Multimodal Interaction. New York: ACM: 569-576 [DOI: 10.1145/3136755.3143011]

Wang S F, Zheng Z Q, Yin S, Yang J J and Ji Q. 2020. A novel dynamic model capturing spatial and temporal patterns for facial expression analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(9): 2082-2095 [DOI: 10.1109/TPAMI.2019.2911937]

Wang S J, Li B J, Liu Y J, Yan W J, Ou X Y, Huang X H, Xu F and Fu X L. 2018. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. Neurocomputing, 312: 251-262 [DOI: 10.1016/j.neucom.2018.05.107]

Wei X F, Li H B, Sun J and Chen L M. 2018. Unsupervised domain adaptation with regularized optimal transport for multimodal 2D+3D facial expression recognition//Proceedings of 13th IEEE International Conference on Automatic Face and Gesture Recognition. Xi'an: IEEE: 31-37 [DOI: 10.1109/FG.2018.00015]

Wu B F and Lin C H. 2018. Adaptive feature mapping for customizing deep learning based facial expression recognition model. IEEE Access, 6: 12451-12461 [DOI: 10.1109/ACCESS.2018.2805861]

Yan J W, Zheng W M, Cui Z, Tang C G, Zhang T, Zong Y and Sun N. 2016. Multi-clue fusion for emotion recognition in the wild//Proceedings of the 18th ACM International Conference on Multimodal Interaction. New York: ACM: 458-463 [DOI: 10.1145/2993148.2997630]

Yan W J, Li X B, Wang S J, Zhao G Y, Liu Y J, Chen Y H and Fu X L. 2014. CASME II: an improved spontaneous micro-expression database and the baseline evaluation. PLoS One, 9(1): e86041 [DOI: 10.1371/journal.pone.0086041]

Yan W J, Wu Q, Liu Y J, Wang S J and Fu X L. 2013. CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces//Proceedings of 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition. Shanghai: IEEE: 1-7 [DOI: 10.1109/FG.2013.6553799]

Yang H Y, Ciftci U and Yin L J 2018a. Facial expression recognition by de-expression residue learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 2168-2177 [DOI: 10.1109/CVPR.2018.00231]

Yang H Y, Zhang Z and Yin L J. 2018b. Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks//Proceedings of 13th IEEE International Conference on Automatic Face and Gesture Recognition. Xi'an: IEEE: 294-301 [DOI: 10.1109/FG.2018.00050]

Yi D, Lei Z, Liao S C and Li S Z. 2014. Learning face representation from scratch. [EB/OL]. [2020-05-31]. https://arxiv.org/pdf/1411.7923.pdf

Yu Z D and Zhang C. 2015. Image based static facial expression recognition with multiple deep network learning//Proceedings of 2015 ACM on International Conference on Multimodal Interaction. New York: ACM: 435-442 [DOI: 10.1145/2818346.2830595]

Zeng N Y, Zhang H, Song B Y, Liu W B, Li Y R and Dobaie A M. 2018. Facial expression recognition via learning deep sparse autoencoders. Neurocomputing, 273: 643-649 [DOI: 10.1016/j.neucom.2017.08.043]

Zhang F F, Zhang T Z, Mao Q R and Xu C S. 2018a. Joint pose and expression modeling for facial expression recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE: 3359-3368 [DOI: 10.1109/CVPR.2018.00354]

Zhang K H, Huang Y Z, Du Y and Wang L. 2017. Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Transactions on Image Processing, 26(9): 4193-4203 [DOI: 10.1109/TIP.2017.2689999]

Zhang T, Zheng W M, Cui Z, Zong Y, Yan J W and Yan K Y. 2016. A deep neural network-driven feature learning method for multi-view facial expression recognition. IEEE Transactions on Multimedia, 18(12): 2528-2536 [DOI: 10.1109/TMM.2016.2598092]

Zhang X, Zhang L, Wang X J and Shum H Y. 2012. Finding celebrities in billions of web images. IEEE Transactions on Multimedia, 14(4): 995-1007 [DOI: 10.1109/TMM.2012.2186121]

Zhang Z P, Luo P, Loy C C and Tang X O. 2018b. From facial expression recognition to interpersonal relation prediction. International Journal of Computer Vision, 126(5): 550-569 [DOI: 10.1007/s11263-017-1055-1]

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

915-928 [DOI: 10.1109/TPAMI.2007.1110]

Zhao G Y, Huang X H, Taini M, Li S Z and Pietikäinen M. 2011. Facial expression recognition from near-infrared videos. Image and Vision Computing, 29(9): 607-619 [DOI: 10.1016/j.imavis.2011.07.002]

Zheng W M, Zong Y, Zhou X Y and Xin X M. 2018. Cross-domain color facial expression recognition using transductive transfer subspace learning. IEEE Transactions on Affective Computing, 9(1): 21-37 [DOI: 10.1109/TAFFC.2016.2563432]

Zhi R C, Flierl M, Ruan Q Q and Kleijn W B. 2011. Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(1): 38-52 [DOI: 10.1109/TSMCB.2010.2044788]

Zhong L, Liu Q S, Yang P, Liu B, Huang J Z and Metaxas D N. 2012. Learning active facial patches for expression analysis//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence: IEEE: 2562-2569 [DOI: 10.1109/CVPR.2012.6247974]

作者简介

李珊,1995 年生,女,博士研究生,主要研究方向为人脸表情识别以及深度学习。
E-mail: ls1995@bupt.edu.cn

邓伟洪,通信作者,男,教授,博士生导师,主要研究方向为模式识别与计算机视觉、人脸识别、表情识别、行人再识别、细粒度图像识别等。
E-mail: whdeng@bupt.edu.cn