基于深度学习的人脸表情识别研究综述

钟　源[1,3]　李鸿天[1,3]　袁家政[2]　刘宏哲[1,3]　徐　成[1,3]

1 北京联合大学北京市信息服务工程重点实验室　北京 100101

2 北京开放大学科学技术学院　北京 100081

3 北京联合大学脑与认知智能北京实验室　北京 100101

摘　要　图像、声音、文字和手势都可以用来表达人类的情感。面部表情是一种非语言交流，它揭示了一个人的内心感受和情感。随着计算机视觉和人工智能技术的发展，人脸表情识别已逐渐成为图像分类领域的研究热点。文中主要列举了几种基于深度学习的表情识别方法，包括卷积神经网络微调、多网络融合、多通道级联等，并详细阐述和比较了各种方法涉及的具体技术。最后，简要总结了人脸表情识别的研究现状，提出了人脸表情识别研究的难点和痛点，也对未来人脸表情识别领域进行了一些展望。

关键词：计算机视觉；人工智能；人脸表情识别；卷积神经网络；多网络融合；多通道级联

中图法分类号　TP391.4

Survey of Facial Expression Recognition Based on Deep Learning

ZHONG Yuan[1,3], LI Hong-tian[1,3], YUAN Jia-zheng[2], LIU Hong-zhe[1,3] and XU Cheng[1,3]

1 Beijing Key Laboratory of Information Service Engineering, Beijing Union University, Beijing 100101, China

2 School of Science and Technology, Beijing Open University, Beijing 100081, China

3 Institute for Brain & Cognitive Sciences, Beijing Union University, Beijing 100101, China

Abstract　Images, sounds, words, and gestures can all be used to express human emotions. Facial expressions are a form of non-verbal communication because they reveal a person's inner feelings and emotions. With the development of computer vision and artificial intelligence technology, facial expression recognition has gradually become a research hotspot in the field of image classification. This paper mainly lists several deep learning-based expression recognition methods, including slight adjustment of convolutional neural network, multi-network fusion, multi-channel cascade, elaborates and compares the specific methods involved in various methods. Finally, the research status of facial expression recognition is briefly summarized, the difficulties and pain points of facial expression recognition research are proposed, and some prospects for the future field of facial expression recognition are also put forward.

Keywords　Computer vision, Artificial intelligence, Facial expression recognition, Convolutional Neural Network, Multi-network fusion, Multi-channel cascade

1　引言

表情作为传递人类情感的重要标志之一，在人际沟通与交往的过程中起着非常重要的作用。学者Mehrabian曾通过实验提出，人们在情绪表达的过程中，面部表情所占的比重值高达55%[1]。因此人

基金项目：国家自然科学基金（61871028，62102033，61906017，62171042）；北京市重点科技项目（CIT&TCD20190313，KZ202211417048）；北京市科技项目（KM202111417001，KM201911417001）；协同创新中心（CYXC2203）

通信作者：袁家政；徐成

脸表情识别是一个重要的研究课题，在各大人机交互的系统中可以得到广泛的应用，例如各类服务机器人、辅助检测疲劳驾驶、监测课堂上学生的专注度情况[2]等。

心理学家 Ekman 和 Friesen 曾把基本表情划分为 6 种[3]，分别为开心、悲伤、惊讶、害怕、愤怒和厌恶，如何识别这 6 种基本表情成为了表情识别这一研究课题中最为关键的目标。

随着计算机视觉和人工智能技术的发展，现如今人脸表情识别的方法主要可以分为两类，分别是传统表情识别方法与基于深度学习的表情识别方法。本文将侧重于围绕基于深度学习的表情识别方法展开研究，比较目前主流的几类基于深度学习的人脸表情识别方法，并给出说明。

2 基于深度学习的表情识别方法

2.1 卷积神经网络微调

深度学习是当今人工智能领域机器学习最热门的方法，最近这几年的发展可以说是日新月异。而卷积神经网络（Convolution Neural Network，CNN）[4]作为深度学习算法中的一个重要组成部分，凭借着其强大的特征提取能力被广泛应用于计算机视觉领域。表情识别作为一种图像分类任务，也可以通过深度学习的方法来实现。因此许多深度学习领域经典的卷积神经网络模型，如 ResNet[5]，VGG[6]等，都可用于人脸表情识别任务。但是若直接套用这些模型用于人脸表情识别，其效果并不理想，因此需要在原有模型的基础上进行优化，根据表情识别这一具体任务的需求修改模型，从而获取更加准确的识别结果，这就是通过卷积网络微调的方式来进行人脸表情识别。

文献[7]在深度残差网络 ResNet18 的基础上，提出了一种改进的网络 C-ResNet18。改进之处是将原网络的最后一个平均池化层替换为一个卷积层和一个最大池化层，使网络提取更多的判别信息特征，防止过拟合，对损失函数进行加权，反复训练以得到一个更优化的神经网络模型。最后的实验结果表明，在 CK＋数据集和 RAF-DB 数据集上，C-ResNet18 人脸表情识别的准确率分别达到了 96.89％和 87.13％，与原模型相比，改进后的网络结构不仅减少了计算量，而且还提高了人脸表情识别的速度以及识别准确率。

文献[8]提出了一种基于改进的 VGG 模型的人脸表情识别方法，对 VGG-16 模型进行改进。主要方法是在原有模型中加入了 BN（Batch Normalization）模块，提高了对参数调整的效率。同时引入 PReLU 激活函数，该文提出，PReLU 激活函数在保留有部分小于 0 的信息的同时又达到了激活函数的目的，小于 0 的部分的斜率也是可学习的，PReLU 激活函数相比 ReLU 激活函数具有更好的激活效果。最终实验结果也表明，改进后的模型进行人脸表情识别的准确率比原模型要高，泛化能力也更强。

综上所述，在一些经典卷积神经网络上进行改进、微调已成为表情识别过程中比较常用的方法，相比直接使用原模型，改进后的模型的识别效果会更加精准。

2.2 多网络融合

人脸表情识别过程中最为关键的就是对人脸特征的提取，多网络融合恰好是采用足够多样的卷积神经网络来提取尽可能多的特征层，通过合适的集成方法来高效融合这些神经网络[9]。如文献[10]提出使用 Visual 和 Landmark 两个分支融合对输入数据进行特征提取，Visual 分支利用浅层神经网络提取图像序列中的中低层次特征，Landmark 分支处理更高层次的特征，提取脸部坐标的位置信息。实验结果表明，该融合网络在 CK＋数据集上的性能明显优于其他方法。文献[11]提出了两个相互协作的深度网络模型。第一个网络是基于多帧外观的 DTAN 网络，第二个网络是基于原始人脸坐标点提取有用的时间几何特征的 DTGN 网络，这两个模型是结合在一起的，为了提高人脸表情识别的性能，采用了一种新的集成方法。最终实验结果表明，该方法在相应数据集上取得了较好的识别效果。

多网络融合由于采用了多个子网络对面部特征进行提取，在特征提取的过程中更加细致，能够准确地将面部特征的变化考虑进去，提升了识别效果，尤其在 CK＋数据集上效果更优秀。

2.3 多通道级联

前文提到采用多网络集成来提取面部特征。基于此，我们也可以从通道数量的方向出发，利用多个平行的通道卷积网络从不同的面部区域学习整合的整体和局部特征，利用联合嵌入式特征学习来从多

个角度和姿势识别人脸表情[9]。文献[12]提出了一种三维人脸表情识别系统,其中考虑了两个通道的特征图像,即LBP和LDP图像。两个学习网络分别是预训练网络和浅层CNN,用于从特征图像中提取特征。然后使用CCA将这些特征融合在一起,将融合后的特征集输入多支持向量机(mSVM)分类器。实验在Bosphorus数据集上取得了不错的效果,结果表明,使用具有多项式核的mSVM分类器的平均准确率为87.69%,并证明了该系统通过表征面部表情的特性比其他方法表现得更好。文献[13]提出了一种有效的端到端可训练的网络DMI-Net,用于姿势感知和身份不变的人脸表情识别。DMI-Net由两个阶段组成。在第一阶段构造一个五元组,在第二阶段,DMI-Net首先使用多通道子CNN提取基于区域的融合特征,然后将融合特征映射到嵌入空间以进行多三元组度量学习。最后,DMI-Net基于多通道度量学习,通过最小化深度多度量损失、FER(Facial Expression Recognition)损失和动态学习损失权重的姿态估计损失来联合识别面部表情和估计姿态。最终实验结果表明,该方法在性能和鲁棒性方面优于现有的最先进的方法,在多视图FER和姿势估计中的最高准确率为93.5%和99.9%。文献[14]提出了一种联合多通道姿态感知卷积神经网络(MPCNN)的多视图人脸表情识别方法。在MPCNN中,多通道子CNN首先学习人脸、嘴巴和眼睛区域,进行多尺度卷积特征提取。然后,联合设计多尺度特征融合层,分层统一不同视图和尺度的高层融合特征表示。最后,姿态感知网络利用条件关节损失函数对头部姿态估计下的最终面部表情进行分类,从而抑制姿态方差的影响。

多通道级联法利用多个平行通道卷积网络从不同的面部区域学习整合的整体和局部特征,在性能和鲁棒性方面效果优异,优于其他方法。

3　分析比较

本节主要是对前文提到的3类主流的基于深度学习的表情识别方法进行分析比较,在比较之前先介绍相关的数据集。

3.1　数据集

(1)FER2013[15]。FER2013数据集由35886张人脸的不同表情图片构成,其中训练集有28708张,验证集和测试集各有3589张。该数据集中每张图片都是灰度图像,其大小为48*48,其中每个人的表情分为7种,分别是开心、悲伤、惊讶、害怕、愤怒、厌恶、正常。

(2)CK+[16]。CK+数据集包含123名受试者的593个图像序列。每个序列都包含从开始(中性帧)到峰值表达(最后一帧)的图像,而其中的峰值帧被FACS编码为面部动作单元。因此,每个图像序列都有面部动作单元的标签,而在这593个图像序列中,有327个序列有对应表情的标签。

3.2　模型对比

本节将对前文提及的各类基于深度学习的表情识别算法进行对比,分别比较这些方法在相应数据集上人脸表情识别的效果,最终的对比结果如表1所列。

表1　基于深度学习的不同表情识别方法间的对比

Table 1　Comparison of expression recognition methods based on deep learning

(单位:%)

方法	FER2013	CK+
卷积网络微调	73.5	96.89
多网络融合	75.1	97.6

3.3　讨论分析

从表1可以看出,不管是在FER2013数据集或者是在CK+数据集上,多网络融合的人脸表情识别方法都有着不错的识别效果,这是因为它采用的是多网络特征提取的方法,在识别效果上也会更加精准,符合逻辑。由于多通道级联的方法是针对多视角多姿态的多通道识别,在FER2013与CK+这两类数据集上进行实验的效果可能会不太理想,因此本文没有将此方法作为对比方法。如果是在包含大量多姿态脸部图片的这类数据集下进行实验,如BU-3DFE[17]数据集,其效果会比前两种方法更好。

结束语　本文主要列举了几类基于深度学习的表情识别方法,对各类方法所涉及到的具体算法进行了详细的阐述与对比分析。目前基于深度学习的人脸表情识别方法在特定数据集下的识别效果尚可,但是仍存在一些问题,如泛化能力不强,目前还不能在市场上大规模推广,只能停留于实验室阶段。以及在一些极端情况下,如低光照、非正脸等,表情

识别效果会显著下降，如何解决这一难题也会是未来的一大研究热点。再者，当拿到表情识别的结果后，如何对这些结果数据进行分析整理，成为了推动人机交互领域发展的关键信息，这将会是重中之重。综合上述问题，本文对人脸表情识别未来的发展进行了展望，未来会针对这些问题寻找合适的解决办法。

参 考 文 献

[1] MEHRABIAN A, RUSSELL A. An approach to environmental psychology [M]. Cambridge: MIT Press, 1974.

[2] LI S, DENG W H. Deep facial expression recognition: a survey [J]. Journal of Image and Graphics, 2020, 25(11): 2306-2320.

[3] EKMAN P. Strong evidence for universals in facialexpressions: a reply to russell's mistaken critique [J]. Psychological Bulletin, 1994, 115(2): 268-287.

[4] O'SHEA K, NASH R. An introduction to convolutional neural networks[J]. arXiv: 1511.08458, 2015.

[5] 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. 2016: 770-778.

[6] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. arXiv: 1409.1556, 2014.

[7] CHEN J, LUO X, MENG Z, et al. Research on Facial Expression Recognition Based on Improved Deep Residual Network Model [C] // Journal of Physics: Conference Series. IOP Publishing, 2021, 2010(1): 012139.

[8] 张士豹，王文韬. 基于改进 VGG 模型的人脸表情识别研究[J]. 现代信息科技，2021，5(23)：100-103.

[9] 洪惠群，沈贵萍，黄风华. 表情识别技术综述[J/OL]. 计算机科学与探索：1-16. [2022-08-10]. https://kns-cnki-net.webvpn.buu.edu.cn/kcms/detail/11.5602.TP.20220420.1147.002.html

[10] VERMA M, KOBORI H, NAKASHIMA Y, et al. Facial expression recognition with skip-connection to leverage low-level features [C] // Proceedings of 2019 IEEE International Conference on Image Processing (ICIP). Taipei, China, Piscataway: IEEE, 2019: 51-55.

[11] JUNG H, LEE S, YIM J, et al. Joint fine-tuning in deep neural networks for facial expression recognition[C] // ICCV. 2017: 2983-2991.

[12] RAMYA R, MALA K, SELVA NIDHYANANTHAN S. 3D facial expression recognition using multi-channel deep learning framework[J]. Circuits, Systems, and Signal Processing, 2020, 39(2): 789-804.

[13] LIU Y Y, DAI W, FANG F, et al. Dynamic multichannel metric network for joint pose-aware and identity-invariant facial expression recognition [J]. Information Sciences, 2021(578): 195-213.

[14] LIU Y, ZENG J, SHAN S, et al. Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition [C] // 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 458-465.

[15] GIANNOPOULOS P, PERIKOS I, HATZILYGEROUDIS I. Deep learning approaches for facial emotion recognition: A case study on FER-2013 [M] // Advances in Hybridization of Intelligent Methods. Cham: Springer, 2018: 1-16.

[16] 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] // 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-workshops. IEEE, 2010: 94-101.

[17] VENKATESH Y V, KASSIM A A, YUAN J, et al. On the simultaneous recognition of identity and expression from BU-3DFE datasets [J]. Pattern Recognition Letters, 2012, 33(13): 1785-1793.

XU Cheng, received the B. E. and M. A. Sc. degrees from the Beijing Key Lab oratory of Information Service Engineering, Beijing Union University, China, in 2012 and 2015, respectively, and the Ph.D degree from the Beijing University of Posts and Telecommunications (BUPT), China. He is currently a lecturer with Beijing Union University. His main research interests include the Internet of Vehicles, intelligent driving, data intelligent, and data security.