JOURNAL OF IMAGE AND GRAPHICS

主办：中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

中国图象图形学报

2022
06
VOL.27

ISSN1006-8961
CN11-3758/TB

中国图象图形学报
刊名题字：宋健 | 月刊（1996年创刊）

第27卷第6期（总第314期）
2022年6月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊
中文核心期刊

版权声明
凡向《中国图象图形学报》投稿，均视为同意在本刊网站及CNKI等全文数据库出版，所刊载论文已获得著作权人的授权。本刊所有图片均为非商业目的使用，所有内容，未经许可，不得转载或以其他方式使用。

Copyright
All rights reserved by Journal of Image and Graphics, Institute of Remote Sensing and Digital Earth, CAS. The content (including but not limited text, photo, etc) published in this journal is for non-commercial use.

主管单位 中国科学院
主办单位 中国科学院空天信息创新研究院
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 吴一戎
编辑出版 《中国图象图形学报》编辑出版委员会
通信地址 北京市海淀区北四环西路19号
邮　　编 100190
电子信箱 jig@aircas.ac.cn
电　　话 010-58887035
网　　址 www.cjig.cn

广告发布登记号 京朝工商广登字20170218号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics
Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Aerospace Information Research Institute, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief Wu Yirong
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
Address No. 19, North 4th Ring Road West, Haidian District, Beijing, P. R. China
Zip code 100190
E-mail jig@aircas.ac.cn
Telephone 010-58887035
Website www.cjig.cn

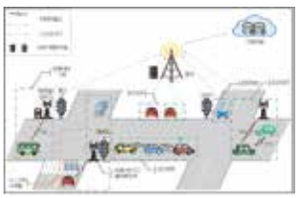
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 60.00元

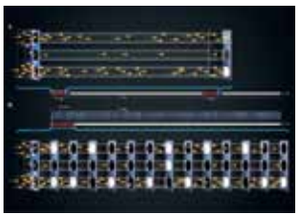
2022年6月 第27卷 第6期（总第314期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次



面向智慧交通的图像处理与边缘计算(第1743页)

视觉理解与计算成像



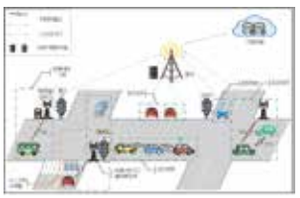
脉冲视觉研究进展(第1823页)

数据挖掘和信息交互

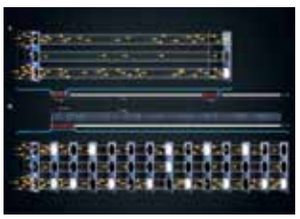
移动在线实时绘制技术研究综述(第1877页)

Volume 27, Number 6
Published June 16, 2022

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



The review of image processing and edge computing for intelligent transportation system(P1743)



Advances in spike vision(P1823)

A review of real-time rendering technology based on mobile internet platforms(P1877)

©中国图象图形学报版权所有

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2022)06-1723-20

论文引用格式:Ma L Z, Wu F, Mao Q R, Wang P J and Chen Y L. 2022. Visual recognition technologies for complex scenarios analysis. Journal of Image and Graphics,27(06):1723-1742(马利庄,吴飞,毛启容,王鹏杰,陈玉珑. 2022. 面向复杂场景的人物视觉理解技术. 中国图象图形学报,27(06):1723-1742)[DOI:10.11834/jig.220157]

面向复杂场景的人物视觉理解技术

马利庄[1*],吴飞[2],毛启容[3],王鹏杰[4],陈玉珑[1]

1. 上海交通大学,上海　200240;2. 浙江大学,杭州　310058;
3. 江苏大学,镇江　212013;4. 大连民族大学,大连　116600

摘　要:面向复杂场景的人物视觉理解技术能够提升社会智能化协作效率,加速社会治理智能化进程,并在服务人类社会的经济活动、建设智慧城市等方面展现出巨大活力,具有重大的社会效益和经济价值。人物视觉理解技术主要包括实时人物识别、个体行为分析与群体交互理解、人机协同学习、表情与语音情感识别和知识引导下视觉理解等,当环境处于复杂场景中,特别是考虑"人物—行为—场景"整体关联的视觉表达与理解,相关问题的研究更具有挑战性。其中,大规模复杂场景实时人物识别主要集中在人脸检测、人物特征理解以及场景分析等,是复杂场景下人物视觉理解技术的重要研究基础;个体行为分析与群体交互理解主要集中在视频行人重识别、视频动作识别、视频问答和视频对话等,是视觉理解的关键行为组成部分;同时,在个体行为分析和群体交互理解中,形成综合利用知识与先验的机器学习模式,包含视觉问答对话、视觉语言导航两个重点研究方向;情感的识别与合成主要集中在人脸表情识别、语音情感识别与合成以及知识引导下视觉分析等方面,是情感交互的核心技术。本文围绕上述核心关键技术,阐述复杂场景下人物视觉理解领域的研究热点与应用场景,总结国内外相关成果与进展,展望该领域的前沿技术与发展趋势。

关键词:复杂场景;视觉理解;人物识别;深度学习;行为分析

Visual recognition technologies for complex scenarios analysis

Ma Lizhuang[1*], Wu Fei[2], Mao Qirong[3], Wang Pengjie[4], Chen Yulong[1]

1. Shanghai Jiao Tong University, Shanghai 200240, China; 2. Zhejiang University, Hangzhou 310058, China;
3. Jiangsu University, Zhenjiang 212013, China; 4. Dalian Nationalities University, Dalian 116600, China

Abstract: Public security and social governance is essential to national development nowadays. It is challenged to prevent large-scale riots in communities and various city crimes for spatial and timescaled social governance in corona virus disease 2019 (Covid-19) likehighly accurate human identity verification, highly efficient human behavior analysis and crowd flow track and trace. The core of the challenge is to use computer vision technologies to extract visual information in complex scenarios and to fully express, identify and understand the relationship between human behavior and scenes to improve the degree of social administration and governance. Complex scenarios oriented visual technologies recognition can improve the efficiency of social intelligence and accelerate the process of intelligent social governance. The main challenge of human recognition is composed of three aspects as mentioned below: 1) the diversity attack derived from mask occlusion attack,

收稿日期:2022-02-28;**修回日期**:2022-03-07;**预印本日期**:2022-03-15

*通信作者:马利庄　ma-lz@cs.sjtu.edu.cn

基金项目:国家自然科学基金项目(61972157);上海市科委项目(21511101200)

Supported by: National Natural Science Foundation of China (61972157); Shanghai Municipal Science and Technology Commission (21511101200)

affecting the security of human identity recognition; 2) the large span of time and space information has affected the accuracy of multiple ages oriented face recognition (especially tens of millions of scales retrieval); 3) the complex and changeable scenarios are required for the high robustness of the system and adapt to diverse environments. Therefore, it is necessary to facilitate technologies of remote human identity verification related to the high degree of security, face recognition accuracy, human behavior analysis and scene semantic recognition. The motion analysis of individual behavior and group interaction trend are the key components of complex scenarios based human visual contexts. In detail, individual behavior analysis mainly includes video-based pedestrian re-recognition and video-based action recognition. The group interaction recognition is mainly based on video question-and-answer and dialogue. Video-based network can record the multi-source cameras derived individuals/groups image information. Multi-camera based human behavior research of group segmentation, group tracking, group behavior analysis and abnormal behavior detection. However, it is extremely complex that the individual behavior/group interaction is recorded by multiple cameras in real scenarios, and it is still a great challenge to improve the performance of multi-camera and multi-objective behavior recognition through integrated modeling of real scene structure, individual behavior and group interaction. The video-based network recognition of individual and group behavior mainly depends on visual information in related to scene, individual and group captured. Nonetheless, complex scenarios based individual behavior analysis and group interaction recognition require human knowledge and prior knowledge without visual information in common. Specifically, a crowdsourced data application has improved visual computing performance and visual question-and-answer and dialogue and visual language navigation. The inherited knowledge in crowdsourced data can develop a data-driven machine learning model for comprehensive knowledge and prior applications in individual behavior analysis and group interaction recognition, and establish a new method of data-driven and knowledge-guided visual computing. In addition, the facial expression behavior can be recognized as the human facial micro-motions like speech the voice of language. Speech emotion recognition can capture and understand human emotions and beneficial to support the learning mode of human-machine collaboration better. It is important for research to get deeper into the technology of human visual recognition. Current researches have been focused on human facial expression recognition, speech emotion recognition, expression synthesis, and speech emotion synthesis. We carried out about the contexts of complex scenarios based real-time human identification, individual behavior and group interaction understanding analysis, visual speech emotion recognition and synthesis, comprehensive utilization of knowledge and a priori mode of machine learning. The research and application scenarios for the visual ability is facilitated for complex scenarios. We summarize the current situations, and predict the frontier technologies and development trends. The human visual recognition technology will harness the visual ability to recognize relationship between humans, behavior and scenes. It is potential to improve the capability of standard data construction, model computing resources, and model robustness and interpretability further.

Key words: complex scenes; visual understanding; human recognition; deep learning; behavior analysis

0 引 言

公共安全与社会治理是国家发展的核心需求，习近平总书记指出疫情中“始终把人民群众的生命安全放在首位”。疫情的爆发使得社会治理面临更为严峻的挑战:需要攻克高精度的人物身份核实、高效的人物行为分析以及人群跨时空流动的跟踪溯源等技术难题,以防止社区大规模骚乱与城市中各类犯罪。其核心是利用计算机视觉技术对复杂场景中视觉信息进行提取,并对其中的“人物—行为—场景”及三者的关联关系进行充分的视觉表达、识别与理解,对于提高社会管理与治理水平,促进行业健康有序发展,具有重要作用。

复杂场景实时人物识别主要包括人物的身份检索与核实、人群跨时空流动的跟踪溯源以及大规模复杂场景实时人物识别等,是对复杂场景中人类活动进行视觉理解的重要基础。人物识别的挑战主要在于面具遮挡攻击等多样性攻击,会影响身份识别安全;时空信息跨度大,会影响跨年龄人脸识别精度(特别是千万级规模的检索);场景复杂多变、要求系统的高鲁棒性和适应多样性环境等问题,需要研究高安全的远程核身、超精准的人脸识别技术,以及高效的行为分析和场景语义理解等技术。

对个体行为进行分析,并理解群体交互规则是复杂场景人物视觉领域的关键组成部分。其中,个体行为分析主要包括视频行人重识别、视频动作识别,群体交互理解主要包括视频问答、视频对话。视频网络可记录个体/群体在多源摄像机中的影像信息,因此多相机环境下的群体分割、群体跟踪、群体行为分析和异常行为检测等研究是人物行为理解的关键,已经成为当前国际国内的热点学术问题。但是,真实场景中多相机所记录的个体行为/群体交互异常复杂,对真实场景结构、个体行为和群体交互进行联合建模来提高多相机、多目标行为理解性能,仍然具有极大的挑战性。

视频网络中个体和群体行为理解主要依赖于摄像机所捕获的场景、个体和群体等视觉信息。然而在复杂场景下,个体行为分析和群体交互理解往往需要视觉信息以外的人类知识与先验常识。特别是,随着互联网中用户产生数据日渐增多,如何利用众包数据来提升视觉计算性能也吸引了众多学者,以此产生了视觉问答与对话和视觉语言导航两个重点任务。这类任务对众包数据中内隐知识进行辨识,在个体行为分析和群体交互理解中形成综合利用知识与先验的数据驱动机器学习模式,建立数据驱动和知识指导的视觉计算新方法,具有广阔的应用前景。

此外,人物表情可以理解为人的脸部微动作,表情识别能够实现人物的情感捕捉与理解,从而更好地支持人机协同的学习模式,是人物视觉理解技术的重要研究方向。情感计算在人工智能与人机交互相关研究中的地位日益凸显,目前,国内外已经在人脸表情识别、表情合成等方面取得了初步成果。

本文重点围绕复杂场景实时人物识别、个体行为分析与群体交互理解、视觉语音情感识别与合成、综合利用知识与先验的机器学习模式,深入阐述面向复杂场景的任务视觉理解技术及应用,汇总国内外的相关成果,并对该领域的前沿进展进行总结与展望。

1　人脸检索与分析及复杂场景实时人物识别

面向复杂场景的人物视觉理解技术是实现社会治理智慧化的核心技术。针对复杂场景中的大规模视觉媒体数据,需要充分识别感知、分析理解其中的人物、行为和场景,挖掘其内在关联,探索“人物—行为—场景”的三位一体视觉表达与理解的科学问题。其中,面向大规模复杂场景的人物视觉理解面临数据量更大、场景更为复杂以及效果需求更高等技术挑战。人物的人脸检索与分析,以及场景中的人物分析则是核心和基础。因此,本文围绕远程核身中高精度人脸验证、鲁棒的活体检测和快速的系统响应等核心问题,从人脸预处理、检测与配准、人脸验证和人脸活体检测 4 个方面进行现状综述,分析了人物视觉理解技术应用的社会影响,并进行了相关风险评估。

1.1　人脸预处理

受限于实际应用中移动设备的图像采集能力,人脸图像的画质可能十分低下,造成人脸采集的光照环境、人脸姿态和表情等的不可控。这些因素都有可能造成人脸识别系统性能的急剧下降。现有的研究分别从人脸图像增强、光照处理、姿态矫正和表情归一化等方面对人脸进行预处理,提升识别质量。

1)人脸图像画质增强。针对人脸识别系统中由模糊产生的人脸图像的降质问题,主要的难点在于推测表示模糊过程的点扩散函数(point spread function,PSF),而从单幅图像推测点扩散函数是一个不适定的问题,因此可以从一个包含多个人物的模糊人脸图像训练集学习得到先验信息来得到 PSF 函数,并利用该 PSF 对输入图像进行去模糊。在此基础上,结合模糊不变的描述子来进一步处理人脸图像的模糊问题,能够提升人脸识别系统的准确率;基于集合论的特征方法,可以解决人脸识别中的模糊和光照变化问题;通过估计人脸图像中的运动模糊和大气模糊,可以自动嵌入到实时的人脸识别系统中;利用基于样例的个人照片增强方法,可以自动地对输入图片进行全局和特定人脸的矫正;同时进行人脸图像的盲卷积和识别,通过追求识别时人脸表示的稀疏性,迭代地求解图像去模糊,实现图像的复原和人脸的识别。

2)人脸光照预处理。人脸识别一大挑战是不同光照环境会造成图像差异,大的光照变化,如阴阳脸,会严重影响人脸识别系统的性能。对光照处理的现有工作大致可以分为两类:主动式和被动式。前者通过一些硬件设备来获取对光照不敏感的图像或 3D 信息;后者通过各种方法来减小或消除不同

光照的影响,包括对光照进行建模、提取光照不变的特征以及对图像进行光照平衡处理等。

基于模型的方法要求光照条件已知或者对象的形状和反射特性已知,理论性强,需要通过数学理论结合光度学理论,给光照变化建立统一的模型,如Shashua 和 Riklin-Raviv(2001)在假定人脸为朗伯体模型且不存在阴影的情况下,引入商图像(quotient image, QI)的概念,以消除图像中的光照变化。提取对光照不敏感特征的方法,需要在光照条件变化不大的情况下才能获得较好的识别效果。该方法需在目标识别的特征提取阶段找到光照不敏感特征或图像表达,并以此作为特征矢量进行目标识别,如Ahonen 等人(2006)将局部二值模式(local binary pattern, LBP)引入人脸识别中,通过提取不同区域的局部特征和直方图统计特征来进行人脸识别,一定程度上降低了光照对识别率的影响。基于图像处理技术的方法包括直方图均衡化、对数变换、Gamma 灰度校正(Shan 等,2003)、自商相位(Wang 等,2004)和相位图(Savvides 等,2004)等,这些方法以其简单有效性在实际中广泛应用。

1.2 人脸检测与配准

人脸检测与配准在计算机视觉技术中有着广泛的应用价值。人脸检测的困难主要来自两个方面:杂乱背景中人脸视觉上的显著变化;人脸所有可能的位置和大小对应的解空间巨大。前者要求人脸检测算法可以准确地解决二分类问题,而后者对应于时间效率要求。人脸配准目前在视频方面的相关研究仍较少,视频前后帧中人脸特征点定位的抖动现象较严重。人脸姿态的多样性、表情变化将引起人脸特征点的变化、人脸光照的变化以及人脸遮挡问题,也增加了人脸特征点定位的难度、降低人脸 3 维重建的精度以及影响人脸识别的准确率。深度学习是近年人工智能领域取得的最重要突破之一,为解决不同人脸姿态、光照变化和人脸遮挡等问题,建立准确性更高、更鲁棒的人脸配准算法具有重要的意义。

1)人脸检测。人脸检测是人脸识别系统的重要步骤。目前人脸检测算法可以准确地检测正面人脸图像,但在非受控环境下的人脸检测中,姿态变化、表情夸张和极端光照条件等,都会导致人脸图像视觉上的巨大改变,从而显著地降低人脸检测的鲁棒性。传统的人脸检测方法主要基于人工设计的特征。自从具有开创意义的 Viola-Jones 人脸检测方法(Wang,2014)提出以来,便出现了许多用于实时人脸检测的方法。利用树结构模型来进行人脸检测,可以同时实现姿态估计和人脸特征点定位;基于部分的可变形模型(deformable part-based model)(Yan 等,2014),可以实现较高的人脸检测准确率。与这些基于模型的方法不同,也可以通过图像检索来检测人脸,形成一个增强的基于范例的人脸检测子,并达到很好的结果。

基于深度学习的人脸检测,深度卷积神经网络提供了强大的特征提取能力,可以获取体现人脸本质的特征表示。基于卷积神经网络(convolutional neural network, CNN)的检测方法之一是 R-CNN(region CNN)(He 等,2017;Girshick,2015),采用“基于区域的识别”模式,在 VOC(visual object classes) 2012 上实现了最好的结果。通过单个深度卷积神经网络检测所有可能方向的人脸,即多视角人脸检测;利用级联深度卷积神经网络,分别以不同分辨率对输入图像进行处理,在低分辨率阶段快速地拒绝多数非人脸图像块,最终在高分辨率阶段准确地判断是否为人脸;根据人脸空间结构获取人脸各子块的响应分数,从而进行人脸检测;利用可变形部分模型(deformable part models)(Yan 等,2014)和深度金字塔提取有效的人脸特征,可以很好地检测非受控条件下各种大小和姿态的人脸。

2)人脸配准。人脸配准是人脸检索与分析中的一个关键技术之一。人脸配准主要完成人面部特征点的定位,包括面部整体轮廓关键点,以及面部五官轮廓的位置关键点,如眼角、嘴角、眼球中心和鼻尖等。传统人脸检测的代表性的方法是 Cootes 等人(1995)提出的主动形状模型(active shape model,ASM)。在 ASM 算法提出后,很多研究人员也对该模型进行了改进,提出了很多改进方法,如利用混合高斯模型对变形参数进行建模,来处理非线性的形状变化;通过核主成分分析(kernel principal component analysis, Kernel PCA)和支持向量机(support vector machine,SVM)来处理非线性的模型变化;基于 Lucas-Kanade 算法的反向组合 AAM(active appearance model)算法(Cootes 等,2001);基于 SDM(supervised descent method)算法将人脸配准视为非线性最小二乘法的优化问题(Xiong 和 de la Torre,2013)以及基于尺度不变特征变换(scale invariant

feature transform,SIFT)特征采用线性回归来预测形状增量(Lindeberg,2012)。传统方法性能的好坏很大程度上取决于初始形状或参数的选取,对未见样例的泛化能力较弱。

基于深度学习的人脸配准方法,在人脸特征点定位的准确性上比传统方法大大提升。与 ASM、AAM 相比,CLM (constrained local model)(Cristinacce 和 Cootes,2006)算法综合考虑了人脸关键点之间的位置关系。采用级联的多个卷积网络来估计人脸关键点的位置;通过使用级联自编码网络提升判别能力来进行人脸配准(Cao 等,2012)。为了实现复杂场景下的多姿态人脸配准,伍凯等人(2017)在级联回归的基础上提出了与初始形状无关的改进的级联回归算法。

1.3　人脸验证

近年来,由于广泛的社会实际需求和人脸识别数据集 LFW(labeled faces in the wild)的发布,非受控条件下的人脸验证技术得到大量研究,并取得了可喜成就。仅在过去的一两年中,人脸验证的准确率就获得大幅度提高,在 LFW 上测试的准确率从 95% 左右提高到 99% 左右,达到乃至超过了人类自身的表现(97.53%)。目前,人脸验证较优的算法分为两类:广度模型(wide model)和深度模型(deep model)。好的模型必须有足够的容量来表示人脸复杂的变化模式。高维 LBP 是一个典型的广度模型,其通过将人脸变换到非常高维的空间使复杂的人脸流形变平。CNN 是目前最先进的深度模型,广泛应用于人脸识别和图像分析。

1)广度模型。许多人脸验证方法用高维的、超完备的人脸描述子表示人脸。将每幅人脸图像编码为 26 K 的基于学习的描述子,然后对 LE(learning-based)描述子用 PCA 降维,再计算 LE 描述子之间的 L2 范数距离;在多尺度下对密集的人脸关键特征点提取了 100 K 的局部二值模式描述子,然后再用 PCA 降维后采用联合贝叶斯进行人脸验证;在尺度和空间上密集地计算 1.7 M 的尺度不变特征变换(SIFT)描述子,将密集的 SIFT 特征编码为 Fisher 向量,并学习以区分性的降维为目的的线性映射;将 1.2 M 的协方差矩阵对象描述子和软直方图局部二值模式描述子组合,学习稀疏的马氏距离。一些研究人员针对身份关联的低层次特征进行了深入研究。利用属性和微笑分类器来检测人脸属性,并度量与参照人脸的相似度;通过 SVM 分类器对来自于不同的两个人的脸进行分类,学习好的分类器的输出为特征。SVM 为浅的结构,且提取的特征是低层次的。

2)深度模型。尽管可以从广度和深度两个方向增加模型的复杂度,但是在同样数目参数的情况下,深度模型比广度模型更有效。普通电脑对广度模型提取的高维特征处理起来较困难,而深度模型每一层的特征维数相对而言小得多,使得其内存消耗是可以接受的。而且,广度模型为人工的设计特征,是非常费力、启发式的,依赖经验和运气,且调节需要大量的时间。而深度模型为无监督特征学习,自动提取特征,不需要人参与。一些深度模型被用来进行人脸验证或人脸辨认。采用暹罗网络(siamese network)进行深度度量学习,采用两个完全相同的子网络分别对两个输入提取特征,并对两个子网络的输出计算距离作为差异度,其子网络为深度卷积神经网络;采用卷积深度神经网络学习特征,然后使用信息论度量学习和线性 SVM 进行人脸验证;采用多个深度卷积神经网络来学习高层次的人脸相似特征,并训练受限玻尔兹曼机分类器进行人脸验证。在过去的一两年中,基于深度学习的人脸验证技术突飞猛进。Facebook 提出了“DeepFace”(Taigman 等,2014),将 3D 模型和姿势变换用于预处理,采用 SFC(social face classification)数据库训练深度卷积神经网络,对于单个网络在 LFW 上的测试,准确率达到 97.00%;采用多尺度网络的方式,训练 7 个神经网络,准确率达到 97.35%,已经与人类自身的表现 97.53% 非常接近。DeepID 通过多尺度、多个神经网络提取高维特征,结合联合贝叶斯对人脸对进行分类,准确率为 97.45%;DeepID2 利用辨认信号和验证信号,两者共同对神经网络参数进行更新,最后用 SVM 将 7 个联合贝叶斯似然比融合,进行分类,准确率达到 99.15%,使人脸验证技术上了一个新的台阶。

1.4　活体检测

关于频域和纹理的方法均是基于单帧的活体检测方法;利用直接拍摄真人与拍摄照片在多方面存在的差别,来区分真人与照片;利用摄像头的焦距变化进行多次拍摄,由于各部分的深度不同,在一定范围内,不同的焦距拍出照片的清晰部位会有所不同来进行活体检测;通过分析 3 维物体与 2 维平面所

产生的光流场属性差异,来进行人脸活体检测;利用背景一致性检测,对于防止视频伪装攻击也是非常重要的;通过图像失真分析的活体检测方法,解决活体检测算法泛化能力差的问题;利用从单幅图像提取 14 种图像质量特征,可应用于实时场景的低复杂度的活体检测算法,来区分真实和假冒的人脸图像;通过寻找由数字化网格的重叠产生的莫列波纹来进行活体检测。

基于运动分析的活体检测试图区分 3D 和 2D 人脸之间的运动模式。其假设为真实的(活的)人脸是 3 维结构,而伪造攻击的人脸是 2 维图像。这些图像可以打印在纸上或在屏幕上显示。运动分析通常依赖于从视频序列中计算出光流。通过用 SVM 对唇动进行分类和分析唇读来进行活体检测。传统的人工设计特征,如 LBP、LBP-TOP(local binary pattern histograms from three orthogonal planes)等,在抵御伪造的图像或视频的攻击方面取得了一定进展,但这些特征还无法捕获真实人脸与假脸间最具有判别力的信息。利用深度卷积神经网络以有监督的方式学习具有强判别力的特征。结合一些数据预处理操作,算法可大幅提升人脸活体检测系统的性能。相比当时最先进的算法,新算法在中国科学院自动化研究所(Institute of Automation, Chinese Academy of Sciences, CASIA)数据库和 REPLAY-ATTACK 数据库上的半错误率相对下降 70%。同时,在这两个数据库上的交叉测试结果,表明该方法具有很好的泛化能力。

2 视频个体行为分析

视频个体行为分析主要包括视频行人重识别、视频动作识别,下面分别介绍相关研究和进展。

2.1 视频行人重识别

视频行人重识别(person re-identification)是利用计算机视觉技术判断视频序列中是否存在目标行人的技术。给定一行人视频及其中目标人物,提供多个监控设备拍摄得到的视频序列,检索跨设备下含目标行人的视频。行人重识别技术可以弥补单一摄像头的视觉局限,在现实场景有着众多应用,如失踪者寻找,嫌疑人跟踪等。

Wojciech 等人(2005)最早进行跨摄像头的多目标跟踪研究,旨在解决当某一摄像头视频丢失特定行人目标后,如何在其他摄像头中再次查找该目标的问题。在此基础上,Gheissari 等人(2006)首次定义视频行人重识别概念。

2.1.1 处理过程及方法

视频行人重识别问题一般按照 3 个阶段流程进行处理。首先对视频数据进行预处,包括提取视频帧的图像特征、采用行人检测模型对人物进行边界框标注和处理光照变化等噪音问题。然后对指定行人进行特征提取,得到行人外观的稳定目标特征。最后找到一种有效的距离度量方法,使视频中同目标更相似的行人在特征空间中距目标特征更近。

1)视频数据预处理。已有的众多方法依赖视频帧的颜色特征,因此场景中光照变化导致的图像颜色变化会严重影响模型的性能。针对这一问题,可从多个方向进行研究,如提取对光照变化具有鲁棒性的图像特征(Farenzena 等,2010);研究正常图像和光照变化图像间的联系,过滤光照变化的影响(Ma 等,2014);采用合适的视频预处理方法,使视频帧颜色变化平缓(Anjum 等,2019)。

2)特征提取。随着深度学习的发展,视频特征提取也从早期的手工标记变成使用深度学习模型提取,模型主要提取两类特征。(1)时空特征,时间特征为视频帧序间的关联,空间特征为每一视频帧中不同位置的图像特征。(2)局部特征,早期研究对每一视频帧只提取一个全局图像特征,不考虑局部区域特征。随着研究的发展,行人识别数据集越来越复杂,因此需要引入视频帧中复杂局部特征。实际研究中行人数据集会存在不可避免的遮挡问题,行人身体的每个区域均可能被其他行人或环境物体(如车和指示牌)遮挡,这将导致行人外观的巨大变化。针对这一问题,最简单的做法是丢弃遮挡帧,如 Li 等人(2018)选择使用时间注意模型从其他所有未丢弃帧中学习有用信息。但丢弃帧会影响视频的时间特征,并且被丢弃的帧中可能包含其他有用信息,因此,Hou 等人(2019)对于部分遮挡的视频帧提出 STCnet(spatial-temporal completion network)方法进行恢复,充分利用视频每一帧的信息。

3)距离度量。找到一个合适的度量函数类计算行人特征向量间的距离,使模型经训练后能将行人特征投影到一个最优的表征空间,其中具有相同行人特征的视频间距尽可能小,不同行人视频间距尽可能大。实际处理的视频为流式数据,视频帧源

源不断加入现有数据当中,因此,距离度量函数不仅需计算出最终特征向量间的距离,还需在新数据输入时,对现有距离进行更新。针对这一点,Navaneet 等人(2019)对新加入的数据提出排名损失,既保证现有距离不断更新,也防止质量差的视频影响模型的性能。

2.2　视频动作识别

识别视频中的动作是视频理解任务中一个充满挑战而又具有较高实际应用价值的任务。视频内容和背景更加复杂多变,不同的动作类别之间具有相似性,而相同的类别在不同环境下又有着不同的特点。此外,由于拍摄造成的遮挡、抖动和视角变化等也为动作识别带来了困难。在实际应用中,精确的动作识别有助于舆情监控、广告投放以及很多其他视频理解相关的任务。

学习视频中帧与帧之间的时序关系,尤其是长距离的时序关系,本身就比较难。不同类型的动作变化快慢和持续时长有所不同,不同的人做同一个动作的方式也存在不同,同时相机拍摄角度和相机自身的运动也会对识别带来挑战。此外,不是视频中所有的帧对于动作识别都有相同的作用,有许多帧存在信息冗余。

2.2.1　基于人工特征的视频动作识别

早期的动作识别主要基于兴趣点的检测和表示。早期主要采用梯度直方图、时空兴趣点检测(Laptev,2005)以及光流直方图(Laptev 等,2008)等方法,都是用于提取图像和时序的特征表示。与图像相比,视频蕴含了大量的运动信息,为了更好地利用运动信息,Wang 和 Schmid(2013)提出密集轨迹的动作识别视频表示方法,提取和追踪密集光流中每个像素特征,编码后进行分类。然而,当面临大规模数据集时,这些特征缺乏一定的灵活性和可扩展性。

2.2.2　3D 卷积的动作识别

视频是由一系列图像帧组成的,图像分类模型已经相对成熟。如何进行视频分类?一种直观的想法是将图像分类的模型直接运用到视频分类中。先把视频各帧提取出来,每帧图像各自前馈(feedforward)一个图像分类模型,不同帧的图像分类模型之间相互共享参数。得到每帧图像的特征之后,对各帧图像特征进行汇合(pooling),例如采用平均汇合,得到固定维度的视频特征,最后经过一个全连接层和 Softmax 激活函数进行分类以得到视频的类别预测。

另一种直观的想法是先把视频逐帧拆分为图像,每帧图像各自用一个图像分类模型得到帧级别的特征,然后用某种汇合方法从帧级别特征得到视频级别特征,最后进行分类预测,其中的汇合方法包括:平均汇合、NetVLAD(net vector of local aggregated descriptors)、NetFV(net Fisher vector)和 RNN3D(3D recurrent neural network)卷积等。另外,也可以借助一些传统算法来补充时序关系,例如,双流法利用光流显式地计算帧之间的运动关系,TDD(trajectory-pooled deep-convolutional descriptor)利用 iDT(improved dense trajectories)计算的轨迹进行汇合等。基于 2D 卷积的动作识别方法一个优点是可以快速吸收图像分类领域的最新成果,通过改变骨架网络,新的图像分类模型可以十分方便地迁移到基于 2D 卷积的动作识别方法中。

2.2.3　基于 3D 卷积的动作识别

4 维视频比 3 维图像多了 1 维,图像使用的是 2D 卷积,则视频使用的是 3D 卷积。因此可以设计对应的 3D 卷积神经网络,从视频片段中同时学习图像特征和相邻帧之间复杂的时序特征,最后利用学到的高层级特征进行分类。相比于 2D 卷积,3D 卷积可以学习到视频帧之间的时序关系。

Tran 等人(2015)首次提出了在视频动作识别中使用 3 维神经网络 C3D(3-dimensional convolutional networks)代替 2 维的神经网络。由于 ResNet 在图像识别任务中取得的较好效果,可以将 2D 卷积神经网络扩展为对应的 3D 卷积神经网络,Hara 等人(2018)提出了基于三维网络的 ResNet。deep mind 团队提出了 I3D(inflated 3D ConvNets)(Carreira 和 Zisserman,2017),具体方法是利用 2D 网络权重展开作为 3D 网络的预训练权重,同时借助大规模的 Kinetics 数据集进行预训练,在基准数据集上效果得到明显提升。

3D 卷积 + RNN、ARTNet(appearance-and-relation network)、Non-Local 和 SlowFast 等从不同角度学习视频帧之间的时序关系。此外,多网格训练和 X3D 等对 3D 卷积神经网络的超参数进行调整,使网络更加精简和高效。

2.2.4　基于双流的神经网络

直接将用于图像分类的神经网络用于视频分类

会忽略视频的时序特征,而时序特征对于视频分类尤为重要。鉴于此,研究者提出了基于双流的动作识别方法。

Simonyan 和 Zisserman(2014)提出了一个融合网络,首次将视频分成空间和时间两个部分,分别将RGB 图像和光流图像送入两支神经网络并融合最终分类结果。利用双流神经网络,可以同时得到视频中人或物体外表和运动的信息,在当时各个基准数据集上取得了领先的识别水平。尽管该方法取得了不错的效果,但仍存在以下缺点:1)视频的预测还是依据从视频中抽取的部分样本,对于长视频来说,在特征学习中还是会损失时序信息;2)在训练时,从视频中抽取片段样本时由于是均匀抽取,存在错误标签的现象(即指定动作并不存在该样本片段中);3)在光流使用前,需要对视频预先做光流的抽取操作。

此外,仍有很多研究者在探索其他更有效的视频动作识别方法,如基于长短期记忆网络(long short term memory, LSTM)的识别框架,基于对抗神经网络(generative adversarial network, GAN)的框架等。虽然目前动作识别已经取得了快速的发展,但距离人类识别水平仍有很大的差距,在实际应用中也面临着各种复杂的问题。期待在今后的研究中能够出现更具有可扩展性、鲁棒性的算法和框架。

3 知识引导下机器学习模型

视频网络中个体和群体行为理解主要依赖于摄像机所捕获的场景、个体和群体等视觉信息,从这些视觉信息出发,进行语义理解。然而在复杂场景下,个体行为分析和群体交互理解往往需要视觉信息以外的人类知识与先验常识。因此建立融入知识的数据驱动机器学习模型,建立数据驱动和知识指导相互结合的视觉计算新方法成为一个新的研究热点。

本节着重介绍将先验知识和知识图谱引入数据驱动机器学习的视觉分析任务。

3.1 融入知识的视觉问答

视频问答(VideoQA)根据视频内容自动回答自然语言问题,广泛应用于在线教育、场景分析和视频内容检索等场景。具体地,视频问答通过理解视频和文本问题中的语义信息,以及它们的语义相关性,预测给定问题的正确答案。视频问答是一项十分复杂的任务,应用了许多人工智能技术,包括对象检测(Lin 等,2017)和分割(Maninis 等,2019)、特征提取(Wong 等,2017)、内容理解(Lu 等,2020)和分类(Anjum 等,2019)等。视频问答打破了视觉和语言的语义鸿沟,从而促进了视觉理解和人机交互。

视觉问答是一个复杂的问题,因为其推理过程中往往额外需要视频帧中不存在的信息,例如常识或有关视频帧的特定知识,因此,一系列工作探索如何将知识、先验融入到视觉问答任务中。

Wu 等人(2016)提出了一种视觉问答方法,该方法将图像内容表示与知识图谱中的信息相结合,以回答基于图像的问题。相比基于神经网络的主要方法,该方法能回答比以前更复杂的问题,即使图像本身不包含整个答案。具体地,该方法通过卷积神经网络(CNN)构建图像表示,并将其与来自知识图谱的文本信息融合。融合信息和查询问题通过循环神经网络,产生视觉问答答案。伍凯等人(2017)进一步将该思想扩展到图像描述任务,并在基准数据集上也达到了最优效果。当前的视觉问答数据集以及基于它们构建的模型专注于仅通过直接分析问题和图像即可回答的问题。Wang 等人(2018a)介绍基于事实的视频问答数据集(fact-based visual question answering, FVQA),主要包含需要外部信息才能回答的问题,需要并支持更深层次的推理。FVQA 通过附加的〈图像, 问题, 答案, 支持事实〉元组扩展了传统的〈图像, 问题, 答案〉三元组视觉问答数据集。基于 FVQA 数据集,Ramnath 和 Hasegawa-Johnson(2021)提出了一种新颖的问答架构,能够对不完整的知识图谱进行推理。该方法使用知识图谱嵌入进行图谱补全,用图像即图谱表示视频帧,采用协同注意力进行知识融合。为了视觉问答推理的可解释性,Wang 等人(2017)描述了一种视觉问答方法,能够根据从大规模知识库中提取的信息对图像进行基于语义结构化解析的可解释推理。通过引入与问题对象及图像对象相关的开放领域知识,Zhang 等人(2021a)提出了一个融合知识 ConceptNet 的视觉问答网络。Marino 等人(2021)利用两种类型的知识表示和推理:一是来自基于 Transformer 模型无监督语言预训练的隐性知识;二是在知识库中编码的显式符号知识。现有的可解释和显式的视觉推理方法只能根据视觉证据进行推理,很少考虑视觉场景之外的知识。为了解决视觉推理方

法和现实世界图像的语义复杂性之间的知识差距，Zhang 等人(2021b)提出了第 1 个结合外部知识的显式视觉推理方法。具体来说，该方法提出了一个知识注入网络帮助显式推理，该网络包含来自外部的实体和谓词的新图节点，用以丰富场景图语义的知识库。GraphRelate 模块随后在该场景图进行高阶关系推理。VQA(visual question answering)模型仅根据人工标注的样本进行训练，很容易对特定的问题样式或被询问的图像内容过拟合，使得 VQA 模型无法学习到问题的多样性。现有方法解决这个问题主要通过引入一个辅助任务，例如视觉基础、循环一致性或去偏差。Kil 等人(2021)发现 VQA 的许多“未知”其实已经隐式暗含在数据集中。例如，询问不同图像中同一物体的问题很可能是同一句子的改写；图像中检测到或标注的对象的数量已经提供了回答“多少”的问题。基于这些发现，提出了一个简单的数据增强方法。该方法将这些“已知”知识转化为 VQA 的训练样本，实验显示这些增强样本可以显著提高 VQA 模型的性能。以上知识分别来自单一模态，嵌入到统一的语义空间需要通过联合学习。为了缓解这一困难，Zhu 等人(2015)提出了多模态数据库，首先构建了一个大规模的多模态知识库，该知识库结合了视觉、文本和结构化数据，以及它们之间的各种关系。FVQA(fact-based visual question answering)现有解决方案在没有细粒度选择的情况下联合嵌入了各种信息，这会引入意想不到的噪音，影响最后推理的答案。Zhu 等人(2020b)通过包含视觉、语义和事实特征的多层多模态异构图来描绘图像。基于该多模态图表示，一种模态感知异构图卷积网络也被提出，用以从不同层中捕捉到与给定问题最相关的证据。为了鼓励开发面向后者的模型，Agrawal 等人(2018)提出了一个新的设置。其中对于每个问题类型，训练集和测试集都有不同的答案先验分布。在这个新设置下，现有 VQA 模型的性能显著下降。为此，Agrawal 等人(2018)同时提出了一种新颖的视觉问答模型(grounded visual question answering model，GVQA)，该模型专门设计架构中的归纳偏差，用于克服训练数据中的先验来防止模型“作弊”，使模型能够更稳健地概括不同的答案分布。随着 BERT(bidirectional encoder representation from Transformers)在文本预训练中的成功，视觉问答中也逐渐开始采用这种预训练—微调范式。Gardères 等人(2020)提出一种基于图像视觉、预训练文本表示以及知识图谱(knowledge graph，KG)表示的多模态概念感知算法 ConceptBert，学习联合概念—视觉—语言的统一嵌入，用以回答需要常识或来自外部结构化事实的问题。

3.2　融入知识的视觉对话

视觉对话旨在根据图像和对话历史生成每个问题的答案(Chen 等，2020b)。尽管最近取得了进展，对于需要先验及事实知识的逻辑推理，现有复杂场景下的视觉对话方法仍然有不足之处。基于此，一系列工作尝试将先验及事实知识融入到视觉对话中。Qi 等人(2020)通过引入因果知识改进视觉对话系统。通过检查模型和数据背后的因果关系，Qi 等人(2020)发现研究者忽略了视觉对话中的两个因果关系。原则 1 建议：应该删除对话历史对答案模型的直接输入，否则会引入有害的捷径偏差；原则 2 建议：历史、问题和答案存在未观察到的混杂因子，导致训练数据产生虚假相关性。视觉对话模型的标准训练范式是最大似然估计(maximum likelihood estimation，MLE)。然而，基于 MLE 的生成模型往往会产生安全和通用的回复，例如，“我不知道”。相比之下，判别式对话模型在回复的自动度量、多样性和信息量方面的表现优于生成式对话模型。为了联合生成模型的实用性和判别式对话模型的强大性能，Lu 等人(2021)训练端到端生成视觉对话模型，其中生成式对话模型接收来自判别式对话模型的梯度作为从生成式对话模型采样的序列的感知(而非对抗性)损失，实现从判别式对话模型到生成式对话模型的知识转移。预训练—微调范式也开始应用到视觉对话领域。Murahari 等人(2020)采用最近提出的 ViLBERT(vision-and-language bidirectional encoder representation from Transformers)模型，在图像描述(Wu 等，2018)和视觉问答(Wu 等，2016；Wang 等，2018a)数据集上进行了预训练，并在 VisDial 数据集上进行微调，使得视觉对话能够利用相关视觉语言数据集蕴含的知识。为了促进人机协同学习，Vries 等人(2016)提出视觉对话猜谜游戏 GuessWhat?!。GuessWhat?! 包含生成问题的提问者和回答图像中有关目标对象的问题的预言(oracle)。根据发问者和 Oracle 之间的对话历史，猜测者对目标对象做出最终猜测。之前的工作仅在 GuessWhat?! 上学习 3 方智能体的单独视觉语言编

码,为了弥补这些差距,Tu 等人(2021)利用预训练的视觉语言模型 VilBERT 学习共享和先验的视觉语言表示知识。

3.3 融入知识的视觉语言导航

视觉语言导航(visual language navigation, VLN)(Wu 等,2021)将自然语言与视觉联系起来,在非结构化、看不见的环境中进行导航任务,吸引了越来越多来自计算机视觉(computer vision, CV)和自然语言处理(natural language processing, NLP)领域研究人员的兴趣。在复杂开放环境中进行视觉语言导航,同样需要额外的常识、事实知识等。Gao 等人(2021)针对真实场景下的远程物体定位导航任务(REVERIE),提出了一种新颖的跨模态知识推理(cross-modal knowledge reasoning, CKR)模型。CKR 基于 Transformer 架构,学习生成场景记忆标记并利用这些信息丰富的历史线索进行环境探索。通过结合常识知识,一个基于知识的实体关系推理模块可以用来学习房间和对象实体之间的内外部相关性,以便智能体在每个视点采取适当的行动。Hong 等人(2020)认为人类能够在看不见的环境中进行导航并定位目标对象,主要是由于先验知识(或经验)和视觉线索的结合。因此,Hong 等人(2020)建议通过构建神经图网络,将外部学习的对象关系先验知识集成到视觉导航模型中,具体地,他们对 actor-critic 强化学习算法中的价值函数进行分解,以一种降低模型复杂性并提高模型泛化的新方式将先验合并到 critic 中。视觉语言导航中一个关键挑战是将当前指令与智能体感知的当前视觉信息进行对齐。大多数现有的工作使用软注意力对单个词来定位下一个动作所需的指令。然而,不同的词在句子中具有不同的功能(例如,修饰语传达属性、动词传达动作)。短语结构等语法信息可以帮助智能体定位指令的重要部分。因此,Mahdi 等人(2020)提出从依存树派生的语法信息来增强指令与当前视觉场景之间的对齐。预先定义位置的视觉对话导航需要昂贵的对话标注,并且不方便真实的人机交流与协作。视觉语言导航的多模态训练数据通常是有限的且标注代价高,因此, Zhu 等人(2021)提出了第 1 个用于视觉语言导航(visual language navigation, VLN)任务的预训练—微调方法。通过对大量图像—文本—动作三元组进行自监督学习方式的训练,视觉语言导航预训练模型提供视觉环境和语言指令的通用表示,并且可以很容易地用做现有的 VLN 框架的插件。大多数视觉语言导航方法采用指令中的单词以及和每个离散的全景视图作为编码的最小单位。然而,这需要模型根据相同的输入视图匹配不同的名词(例如,电视、桌子)。Qi 等人(2021)提出了一个对象感知的顺序 BERT,以相同的细粒度层次编码视觉感知和语言指令。该模型能够识别每个可导航位置的相对方向(例如,左/右/前/后)以及当前和最终导航目标的房间类型(例如卧室、厨房)。多模态 BERT 已经应用到许多视觉语言任务。然而,它在视觉语言导航任务中的应用仍然有限。原因之一是难以将需要依赖历史的注意力和决策的 BERT 架构应用到 VLN 的部分可观察的马尔可夫决策过程。为此,Hong 等人(2021)提出了一个时间感知的循环 BERT 模型。具体来说,该 BERT 模型具有循环函数并且保留智能体的跨模态状态信息。

3.4 复杂场景人脸表情识别与合成

情感是人类日常人际交往的重要组成部分,在人机交互、行为分析等方面起着至关重要的作用。通过对用户情感进行正确认知并做出快速、正确的反馈,实现计算机更“拟人化”地应用于日常生活。下面分别介绍表情识别和合成的相关研究。

3.4.1 复杂场景表情识别

人脸表情识别的定义是:计算机捕获面部相关样本信息,设计算法提取人脸情感表征,再进行情感分析和分类。在计算机视觉和机器学习领域,人脸表情识别有着广泛的应用,并涉及计算机图形学、心理学等多个研究领域的知识,吸引了国内外学者的关注并投入研究。人脸表情识别的流程一般分为 3 个部分:人脸图像预处理、人脸情感表征提取和表情识别。

1)人脸图像预处理方法。图像预处理主要分为 3 个模块:人脸检测、数据增强和人脸归一化。人脸检测的目的是从样本中定位出人脸区域并剔除人脸无关区域,常用的方法包含刚性模型和形变模型(Zafeiriou 等,2015)。数据增强的目的是扩充带标签的样本数量参与到模型训练,主要包括旋转、缩放、位移、加噪声以及颜色抖动等方式。姿态以及光照是复杂环境下人脸表情样本中存在的普遍影响因素。人脸归一化包含对尺度、姿态以及光照的处理。在预处理过程中,将图像中大幅度变化的非正脸姿

态归一化为标准姿态空间中的正脸姿态(Zhang 等,2020a)。光照归一化的目的是一定程度上减轻人脸表情样本的类内差异,具体做法一般将检测到的人脸区域进行统一光照处理。

2)人脸情感表征提取方法。人脸情感表征通常分为两类:学习型特征和手工特征(Li 等,2020)。学习型特征一般通过深度网络提取后与情感分类集成在统一模型中。手工特征的提取和情感分类是两个单独进行的过程,即在提取情感特征之后,再将提取到的特征作为识别模型的输入进行情感识别。

复杂环境下光照和姿态变化是影响分类性能的两大主要障碍(Tan 等,2021),学习并控制鲁棒的情感表征是人脸表情识别领域面临的重要挑战。近年已有大量研究者设计不同的算法解决此类任务(Tang 等,2020,2021;Shao 等,2021b)。其中,马利庄团队(Shao 等,2021a)提出了自适应调整人脸区域重要性的区域注意力网络,设计区域偏置损失,建模局部区域与全局面部信息,学习显著情感特征。区域注意力网络由 3 部分组成:情感特征提取模块、自注意力机制模块和关系注意力机制模块。

3)表情识别方法。表情识别的目的是通过表情识别算法,对提取的人脸情感表征进行理解与分析,获取样本对应的情感类别。根据不同的情感表征提取方法,表情识别的方法通常分为两类:基于传统机器学习和深度学习。基于传统机器学习的人脸表情识别常用模型包括支持向量机 SVM、贝叶斯模型,回归模型和 K 最近邻(K-nearest neighbors,KNN)等。

3.4.2　人脸表情合成

人脸表情合成是指利用人脸情感表征,合成大量任意表情下的人脸表情图像。生成对抗网络(generative adversarial network, GAN)(Otberdout 等,2022)作为主流模型广泛用于合成大量人脸表情图像。Yan 等人(2020)提出了一种将表情合成和表情识别集成到一个统一框架中的方法。首先,对人脸表情合成生成对抗网络(facial expression synthesis generative adversarial network, FESGAN)进行预训练,合成具有不同表情的人脸样本。为了增加训练图像的多样性,FESGAN 首先从先验知识中学习并合成具有新的身份标识信息的人脸图像。其次,将表情合成与表情识别集成到统一框架中,结合预先训练得到的 FESGAN 共同训练人脸表情识别网络。Zhang 等人(2020b)提出了基于多任务协同分析的鲁棒人脸表情识别方法。该方法联合了人脸关键点检测、人脸合成以及人脸表情识别 3 个任务,共享情感特征、几何特征以及生成数据。具体地,为了生成具有任意姿态和表情的人脸图像,从人脸图像中分离出属性(姿态和表情),从而得到足够的训练样本来辅助人脸表情识别和人脸合成任务。同时,人脸表情识别任务促使生成的人脸图像看起来更接近真实样本以及人脸对齐可以为人脸合成提供有效的几何特征。

4　国内研究进展

4.1　人脸检索和分析与复杂场景实时人物识别

画质增强处理技术的研究主要集中在分辨率增强和去运动模糊,学者们提出了一些有效的画质增强算法。卿来云等人(2006)基于球面谐波(Basri 和 Jacobs,2003)理论,提出了新的光照补偿算法;Zhang 等人(2021a)利用视频的前后帧相关性,引入视频加速的一些手段,以提高视频处理时画质增强算法的效率。

在人脸图像预处理的高效方法方面,已有的人脸光照预处理方法包括:直方图均衡化、基于小波的归一化和基于离散余弦变换的归一化等。其中直方图均衡化是这类方法中使用最多的手段,中心思想是把原始图像的灰度直方图从比较集中于某个区间转变成在全部范围内均匀分布。直方图均衡化就是对图像进行非线性拉伸,重新分配图像像素值,使一定灰度范围内的像素数量大致相同。经过实验表明,这种方法的优点是它对所有环境下的图片,即使是已经控制过光照条件的图像数据,均有提升效果。

早期国内对人脸检测问题的研究很多,清华大学(Lyu 等,2000;卢春雨 等,1999;周杰 等,2000;梁路宏 等,1999;Ai 等,2000),北京工业大学(Miao 等,1999;邢昕 等,2000),中国科学院计算技术研究所(刘明宝 等,1998)和中国科学院自动化研究所(Wang 和 Tan,2000)都有人员从事人脸检测相关的研究。马利庄团队 2014 年 11 月的人脸检测技术在 FDDB(face detection dataset and benchmark)评测数据库上达到世界领先水平(见图 1)。但这些方法或者为浅层模型,或者是基于人工设计的特征,对调参过程要求很高,而且泛化能力较弱。

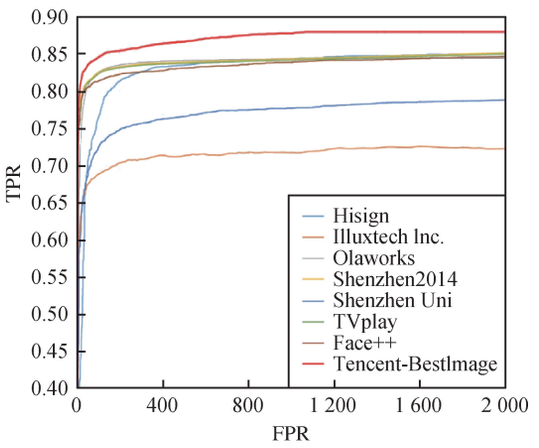


图 1　不同人脸检测算法在 FDDB 数据集上的性能比较(来自 2014 年 11 月 FDDB 官网榜单数据)

Fig. 1　Quantitative comparison results on FDDB face detection dataset(results from FDDB benchmark in Nov. 2014)

在人脸验证领域,马利庄团队研发的人脸验证技术 Tencent-BestImage 在 LFW 上取得了 99.65% 的识别率(2015 年 6 月),多次刷新世界纪录(见图 2)。目前百度、Google 等公司更是将识别准确率推升到 99.7% 以上的新高度。围绕人脸检测、人物特征理解以及场景分析所展开的研究,在显著性检测、图像增强、人脸识别与验证、人脸配准、3 维人体姿态估计和超高清渲染等技术均达到国际一流水平。

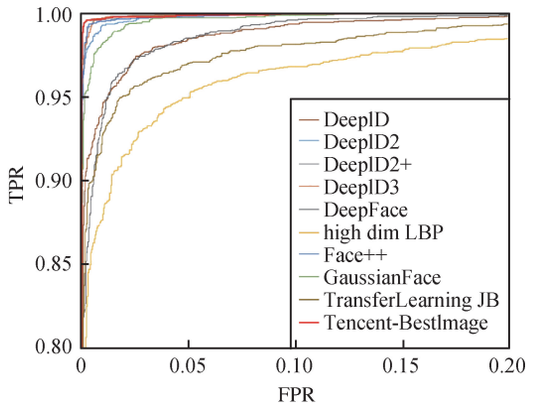


图 2　不同算法在 LFW 数据集上的 ROC 曲线

Fig. 2　ROC curves of different algorithms on the LFW dataset

在活体检测领域,进入到深度学习时代,许多方法(Feng 等,2016;Li 等,2016;Yang 等,2014)将活体检测问题视为二分类问题并利用卷积神经网络解决问题。为了避免过拟合问题(Liu 等,2018;Shao 等,2019;Liu 等,2019;Yang 等,2019;Yu 等,2020)可以使用额外的监督,例如深度图,反射图或者 rPPG(remote photoplethysmography)信号,来提升网络效果。Liu 等人(2018)首次使用深度图作为真实人脸和特征攻击的判别特征。基于辅助信息,已有方法从解耦的角度对特征进行了进一步的规整(Zhang 等,2020c;Liu 等,2020)。

4.2　从个体行为分析到群体交互理解

4.2.1　视频行人重识别

1)距离度量方法。Zhu 等人(2018)提出了 SI^{2}DL(simultaneous intra-video and inter-video distance learning)方法,对于单个视频内的视觉特征使相互间距离尽可能小,对于不同视频,使同类视觉特征间距离尽可能小,使不同类特征间距离尽可能大,以此进行不同行人视频的分类。Zhang 等人(2019)引入均值—体(mean-body),定义一个新的视频内的特征差异损失来处理同一视频内时空特征间的变化。

2)行人不对齐和姿态变化。由于背景杂波和位置不对齐导致的图像不对齐现象普遍存在于现有行人重识别数据集中,此外,由于拍摄角度变化、路径变化以及行为变化等原因会导致行人姿态变化,这两个问题会严重影响模型性能。Chen 等人(2019)提出 STSN(pose-guided spatial transformer sub-network)方法。对于图像不对齐问题,将输入图像的 Transformer 参数回归后,经仿射变换(affine transformation)转换为对齐的图像。为减轻姿态变化影响,挑选具有最大 Transformer 贡献值的帧作为关键帧来训练模型。姿态估计对齐方法需要额外的姿态标注,Wu 等人(2019)引入姿态估计模型对行人重识别数据集进行处理,利用半监督方法避免人工标注。

3)遮挡问题。一般情况下,视频中只会有部分时间存在人物遮挡问题,Zhou 等人(2017)提出通过时间注意模型来选择视频中特征最稳定、最具区别性的帧,并基于此进行特征学习。

4.2.2　视频动作识别

1)基于 3D 卷积的动作识别。由于 3D 卷积神经网络的参数量和计算量比 2D 卷积神经网络大了很多,不少研究工作专注于对 3D 卷积进行低秩近似,TSM(temparal shift module)对 2D 卷积进行改造以近似 3D 卷积的效果(Lin 等,2019),Qiu 等人(2017)也提出了用 2 维模拟 3 维神经网络的伪 3D 网络(P3D)。

2)基于双流的神经网络。在 Simonyan 和 Zis-

serman(2014)提出了双流网络之后,有许多研究针对双流网络这种框架进行了一些改进,例如 TSN(temporal segment network)(Wang 等,2016)是一种可以捕捉较长时序的网络结构。Xu 等人(2019)提出了基于密集扩张网络的框架,并探讨了空间和时间分支的不同融合方式。

4.3　人脸表情识别与合成

人脸表情识别在情感分析与识别中具有关键作用。在人类交流过程中,面部表情传达信息的比重高达 55% 。随着深度学习技术的快速发展,人脸表情识别领域引起了广泛的关注并产生了大量研究成果,然而,该领域仍然存在大量挑战。比如带可靠标签的表情样本较少、复杂环境下出现大幅度面部遮挡、非正脸姿态问题以及情感标签存在不确定性等问题。为了缓解表情数据库规模不大的问题,研究者们提出的方案主要包括:利用迁移学习方法将物体识别模型或者人脸识别模型迁移到表情识别任务中(Zhi 等,2019),利用半监督方法对数据库中没有标签的表情进行标注(Liu 等,2021b),以及利用生成对抗网络方法生成更多样本(Xie 等,2021)等。为了缓解面部区域遮挡和大幅度姿态对人脸表情识别产生的影响,借助局部块注意力机制来学习情感显著局部信息是比较高效的方法(Zhao 等,2021c;Wang 等,2020b;Wang 等,2021c),或者利用多任务学习促进人脸表情的特征学习(Chen 等,2021;Zhang 等,2020b)。表情标签不确定的问题主要表现为:存在模棱两可的人脸表情、低质量的表情图片,以及标注者的主观性导致在标注情感标签时存在歧义。为了解决此类问题,研究者们尝试在多个数据库上利用深度学习模型预测情感标签分布,辅助训练挖掘潜在标签来提升模型的鲁棒性,以及结合注意力机制与重新标注样本来抑制表情标签不确定的样本(Chen 等,2020a;Wang 等,2020a)。

随着人工智能的发展,类人机器人富有情感表现力的表情合成也成为情感计算领域的研究热点之一。面部表情合成即是利用计算机技术生成带有表情的人脸图像。由于面部表情的多样性以及类人机器人硬件设计的复杂性,如何实现类人机器人对人类表情自然而真实的模拟仍然是类人机器人领域所面临的难点之一。目前国内主流的方法是借助辅助信息在生成对抗网络中进行面部表情合成(Zhao 等,2021b;Wang 等,2019;Yu 等, 2021)。其中,辅助信息包括但不限于面部运动单元信息(Zhao 等,2021b)、表情识别标签信息(Wang 等,2019)以及身份信息(Yu 等,2021)等。

4.4　综合利用知识与先验的机器学习模式

4.4.1　融入知识的视觉问答

外部知识融入的 VQA 模型训练严重依赖于作为监督信息的真实知识事实,在训练过程中遗漏这些真实知识事实将导致无法产生正确的答案。为了解决这一问题,Li 等人(2020)提出了一种知识图增强模型,该模型不需要额外监督的真实知识事实,即可对外部知识图进行上下文感知知识聚合。具体地,该模型能够检索给定视觉图像和文本问题的上下文感知知识子图,并学习聚合有用的图像和问题相关知识,然后利用该知识来提高回答视觉问题的准确性。视觉问答(VQA)需要对图像和自然语言问题的联合理解,其中许多问题无法直接或明确地回答视觉内容,但需要结构化的人类推理从视觉内容确认的知识。Su 等人(2018)提出了视觉知识记忆网络(visual knowledge memory networks,VKMN),将结构化的人类知识和深度视觉特征无缝地整合到端到端学习框架中的记忆网络中。与其他利用外部知识的 VQA 方法相比,VKMN 首先将视觉内容与知识事实联合嵌入到视觉知识特征中。其次,VKMN 从问答数据中扩展出多个知识,并使用标签数据将其联合嵌入存储对记忆网络中。类似地,Yu 等人(2020)也将 FVQA 表示为多层多模态异构图,并将基于知识的视觉问答,形式化为从多模态信息中获取补充证据的循环推理过程。该推理过程由一系列基于记忆的推理步骤组成,每个步骤包含基于图的读取、更新和控制模块,对视觉和语义信息进行并行推理。通过多次堆叠模块,联合考虑所有概念来推断全局最优答案。除了上述人类知识,研究者也对先验知识对视觉问答的影响进行研究。许多研究发现,当今的视觉问答模型在很大程度上受到训练数据中表面相关性的驱动,并且缺乏足够的图像基础。Jing 等人(2020)提出了一种新的基于语言注意力的 VQA 方法,学习解耦的问题语言学表示,并利用这些表示推理克服语言先验的答案。Lao 等人(2021)指出目前的视觉问答研究主要挑战之一是模型对语言先验的过度依赖(以及对视觉模态的忽视)。为了缓解这个问题,通过重新调整标准交叉熵损失函数,Lao 等人(2021)提出了一种新颖的基于语言先

验的损失函数(LP-focal loss)。具体来说,LP-focal loss 仅使用问题分支来捕获每个答案候选者的语言偏见。在计算训练损失时,LP-focal loss 动态地为有偏见的答案分配较低的权重,从而减少训练数据中有偏样本的贡献。

4.4.2 融入知识的视觉对话

Zhao 等人(2021)提出结构化知识感知网络(structured knowledge-aware network, SKANet),该网络包含多模态融合模块、图像知识感知模块和描述知识感知模块。图像和描述知识感知模块从 ConceptNet 构建常识知识图,用以应对复杂的场景。Jiang 等人(2020)认为对话问题、视觉知识和文本知识的拼接整合操作,信息检索能力有限,无法缩小跨模态信息之间的异构语义鸿沟。为此,Jiang 等人(2020)提出知识桥图网络(knowledge-bridge graph network, KBGN)模型,通过使用图网络在细粒度上桥接视觉和文本知识之间的跨模态语义关系鸿沟,并通过自适应信息选择模式检索所需的知识。此外,从模态内实体和模态间桥梁中可以清楚地得出视觉对话的推理线索。Wang 等人(2020c)提出利用预训练 BERT 语言模型,建立一个简单而有效的统一框架视觉对话 Transformer,即 VD-BERT。该模型的统一之处在于其使用单流 Transformer 编码器捕获图像和多轮对话之间的所有交互,以及同时支持通过相同的架构进行答案排序和答案生成。

4.4.3 融入知识的视觉语言导航

Gao 等人(2021)认为场景之间的关系、物体对象以及方向线索对于智能体解释复杂指令并正确感知环境至关重要。为了捕捉和利用这些关系,他们提出了一种新颖的语言和视觉实体关系图来建模文本和视觉之间的模态间关系,以及模态内视觉实体之间的关系。同时,一种用于传播的图中语言元素和视觉实体之间的信息的消息传递算法也被结合起来确定智能体下一步要采取的行动。Li 等人(2021)提出了一种自我激励的交流智能体,该智能体能自适应地学习是否与人类交流以及与人类交流什么以获得指导信息,以实现对话无注释导航和增强现实世界中看不见的环境的可转移性。传统视觉语言导航方法只利用交叉模态中的视觉和语言特征,却忽略了环境中包含的丰富的语义信息(如隐含的导航图或子轨迹语义)。为此,Zhu 等人(2020a)提出具有 4 个自监督的辅助推理导航框架(AuxRN),该框架能利用来自环境中的额外语义信息进行训练。AuxRN 有 4 个推理目标:解释前面的动作、估计导航进度、预测下一个方向以及评估轨迹一致性。这些额外的训练信号有助于智能体获得语义表示知识,以便推理其活动并建立深入的环境感知。

5 发展趋势与展望

在面向复杂场景的人物视觉理解技术及应用的相关研究中,大规模场景实时人物识别、个体行为分析与群体交互理解、视觉语音情感识别与合成、知识引导和数据驱动是实现数字化、智能化生活与信息化服务不可或缺的重要环节,对于维护社会治理与公共安全、提升产业效率、促进智慧城市建设具有重要作用。其中,人脸检索和分析与大规模场景实时人物识别是面向公共安全、互联网金融和社交网络等领域的关键基础问题,近年来取得极大进展,但仍存在着具有面具遮挡攻击等多样性,影响身份识别安全;时空信息跨度大,影响跨年龄人脸识别精度;场景复杂多变,要求系统的高鲁棒性、适应多样性环境等问题。为进一步进行技术推广、促进产业升级,仍需要针对训练数据稀缺、深度学习难解释以及复杂环境存在各种非受控因素等问题进行深入研究,从而高效和鲁棒地实现人脸检索和分析与大规模场景实时人物识别。

在个体行为分析与群体交互理解方面,虽然近几年视频行人重识别取得了重大发展,但还是面临着诸多挑战。例如在真实场景下,行人重识别会遇到跨摄像头导致的姿态变化、视角变化等问题,导致行人外观的巨大变化;此外,视频行人重识别方法虽然在一定程度上解决了部分遮挡的问题,但是丢弃遮挡图像的解决思路并不理想;光照变化会进一步降低行人重识别模型的性能。虽然目前动作识别已经取得了长足的发展,但距离人类识别水平仍有很大的差距,在实际应用中也面临着各种复杂的问题。其中,训练视频模型所需的计算资源远超图像,使得视频模型的训练时长和训练所需的硬件资源开销巨大,导致模型的验证和迭代速度减慢。因此,将数据驱动机器学习和知识引导逻辑推理方法进行结合,研究泛化能力更强的算法和框架是未来重要的研究方向。此外,数据集规模制约了动作识别领域的发展,仍需要进一步完善。

与此同时，人工智能的发展带动着情感计算逐步达到更高水平，然而同其他高端科技一样，在到达一定阶段后，情感计算也迎来了技术的“瓶颈期”。比如，在表情识别中，真实世界人脸表情数据标注不足、表情数据类别不平衡、数据偏差大以及标注不一致等问题成为制约表情识别的主要因素。针对以上问题，未来发展除了要讨论方法的精度也要关注方法的耗时以及存储消耗。如何引入新技术解决小样本和不平衡分类问题、如何有效利用多类表情模型协同工作以及如何将表情信息与其他模态信息结合到一个高层框架中提供互补信息来增强模型的鲁棒性是表情识别领域未来的重点研究方向；如何构建情感表现力更丰富、情感控制度量更标准的数据库，如何利用深度学习方法（如少样本甚至单样本、零样本学习方法）来缓解可靠数据问题，如何在端到端的神经网络中融合更多的个性化、场景化的信息以合成更拟人化的情感信息是人脸表情合成领域的重要研究方向。

综上所述，面向复杂场景的人物视觉理解技术及应用在服务人类社会的经济活动、建设智慧城市等方面具有重大意义。期待人物视觉理解技术在人物—行为—场景 3 要素关联的视觉理解方面取得进展，同时在标准数据建设、模型计算资源以及模型鲁棒可解释性方面进一步完善。

致　谢　本文由中国图象图形学学会动画与数字娱乐专业委员会组织撰写，该专委会更多详情请见链接：http://www.csig.org.cn/detail/2387。

参考文献（References）

Agrawal A, Batra D, Parikh D and Kembhavi A. 2018. Don’t just assume; look and answer: overcoming priors for visual question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 4971-4980 [DOI: 10.1109/CVPR.2018.00522]

Ahonen T, Hadid A and Pietikainen M. 2006. Face description with local binary patterns: application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12): 2037-2041 [DOI: 10.1109/TPAMI.2006.244]

Ai H Z, Liang L H and Xu G Y. 2000. A general framework for face detection//Proceedigns of the 3rd International Conference on Multimodal Interfaces. Beijing, China: Springer: 119-126 [DOI: 10.1007/3-540-40063-X_16]

Anjum A, Abdullah T, Tariq M F, Baltaci Y and Antonopoulos N. 2019. Video stream analysis in clouds: an object detection and classification framework for high performance video analytics. IEEE Transactions on Cloud Computing, 7(4): 1152-1167 [DOI: 10.1109/TCC.2016.2517653]

Basri R and Jacobs D W. 2003. Lambertian reflectance and linear subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(2): 218-233 [DOI: 10.1109/TPAMI.2003.1177153]

Cao X D, Wei Y C, Wen F and Sun J. 2012. Face alignment by explicit shape regression//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, USA: IEEE: 2887-2894 [DOI: 10.1109/CVPR.2012.6248015]

Carreira J and Zisserman A. 2017. Quo vadis, action recognition? A new model and the kinetics dataset//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE: 4724-4733 [DOI: 10.1109/CVPR.2017.502]

Chen B Y, Guan W L, Li P X, Ikeda N, Hirasawa K and Lu H C. 2021. Residual multi-task learning for facial landmark localization and expression recognition. Pattern Recognition, 115: #107893 [DOI: 10.1016/j.patcog.2021.107893]

Chen S K, Wang J F, Chen Y D, Shi Z C, Geng X and Rui Y. 2020a. Label distribution learning on auxiliary label space graphs for facial expression recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 13981-13990 [DOI: 10.1109/CVPR42600.2020.01400]

Chen X F, Lao S Y and Duan T. 2020b. Multimodal fusion of visual dialog: a survey//Proceedings of the 2nd International Conference on Robotics, Intelligent Control and Artificial Intelligence. Shanghai, China: ACM: 302-308 [DOI: 10.1145/3438872.3439098]

Chen Y Z, Huang T D, Niu Y Z, Ke X and Lin Y Y. 2019. Pose-guided spatial alignment and key frame selection for one-shot video-based person re-identification. IEEE Access, 7: 78991-79004 [DOI: 10.1109/ACCESS.2019.2922679]

Cootes T F, Edwards G J and Taylor C J. 2001. Active appearance models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6): 681-685 [DOI: 10.1109/34.927467]

Cootes T F, Taylor C J, Cooper D H and Graham J. 1995. Active shape models-their training and application. Computer Vision and Image Understanding, 61(1): 38-59 [DOI: 10.1006/cviu.1995.1004]

Cristinacce D and Cootes T F. 2006. Feature detection and tracking with constrained local models//Proceedings of the British Machine Vision Conference. Edinburgh, UK: BMVA Press: #95 [DOI: 10.5244/C.20.95]

Farenzena M, Bazzani L, Perina A, Murino V and Cristani M. 2010. Person re-identification by symmetry-driven accumulation of local features//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). San Francisco, USA: IEEE: 2360-2367 [DOI: 10.1109/CVPR.2010.5539926]

Feng L T, Po L M, Li Y M, Xu X Y, Yuan F, Cheung T C H and

Cheung K W. 2016. Integration of image quality and motion cues for face anti-spoofing: a neural network approach. Journal of Visual Communication and Image Representation, 38: 451-460 [DOI: 10.1016/j.jvcir.2016.03.019]

Gao C, Chen J Y, Liu S, Wang L T, Zhang Q and Wu Q. 2021. Room-and-object aware knowledge reasoning for remote embodied referring expression//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3063-3072 [DOI: 10.1109/CVPR46437.2021.00308]

Gardères F, Ziaeefard M, Abeloos B and Lecue F. 2020. ConceptBert: concept-aware representation for visual question answering[s. l.]: Association for Computational Linguistics: 489-498 [DOI: 10.18653/v1/2020.findings-emnlp.44]

Gheissari N, Sebastian T B and Hartley R. 2006. Person reidentification using spatiotemporal appearance//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE: 1528-1535 [DOI: 10.1109/CVPR.2006.223]

Girshick R. 2015. Fast R-CNN//Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE: 1440-1448 [DOI: 10.1109/ICCV.2015.169]

Hara K, Kataoka H and Satoh Y. 2018. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? //Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6546-6555 [DOI: 10.1109/CVPR.2018.00685]

He K M, Gkioxari G, Dollár P and Girshick R. 2017. Mask R-CNN//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2980-2988 [DOI: 10.1109/ICCV.2017.322]

Hong Y C, Rodriguez-Opazo C, Qi Y K, Wu Q and Gould S. 2020. Language and visual entity relationship graph for agent navigation//Proceedings of the 34th Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 7685-7696

Hong Y C, Wu Q, Qi Y K, Rodríguez-Opazo C and Gould S. 2021. VLN↻BERT: a recurrent vision-and-language BERT for navigation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1643-1653 [DOI: 10.1109/CVPR46437.2021.00169]

Hou R B, Ma B P, Chang H, Gu X Q, Shan S G and Chen X L. 2019. VRSTC: occlusion-free video person re-identification//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 7176-7185 [DOI: 10.1109/CVPR.2019.00735]

Jiang X Z, Du S Y, Qin Z C, Sun Y J and Yu J. 2020. KBGN: knowledge-bridge graph network for adaptive vision-text reasoning in visual dialogue//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1265-1273 [DOI: 10.1145/3394171.3413826]

Jing C C, Wu Y W, Zhang X X, Jia Y D and Wu Q. 2020. Overcoming language priors in VQA via decomposed linguistic representations. Proceedings of the AAAI Conference on Artificial Intelligence, 34(7): 11181-11188 [DOI: 10.1609/aaai.v34i07.6776]

Kil J, Zhang C, Xuan D and Chao W L. 2021. Discovering the unknown knowns: turning implicit knowledge in the dataset into explicit training examples for visual question answering [EB/OL]. [2022-02-13]. https://arxiv.org/pdf/2109.06122.pdf

Lao M R, Guo Y M, Liu Y and Lew M S. 2021. A language prior based focal loss for visual question answering//Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen, China: IEEE: 1-6 [DOI: 10.1109/ICME51207.2021.9428165]

Laptev I. 2005. On space-time interest points. International Journal of Computer Vision, 64(2): 107-123 [DOI: 10.1007/s11263-005-1838-7]

Laptev I, Marszalek M, Schmid C and Rozenfeld B. 2008. Learning realistic human actions from movies//Proceedings of 2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, USA: IEEE: 1-8 [DOI: 10.1109/CVPR.2008.4587756]

Li G H, Wang X and Zhu W W. 2020. Boosting visual question answering with context-aware knowledge aggregation//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 1227-1235 [DOI: 10.1145/3394171.3413943]

Li J L, Tan H and Bansal M. 2021. Improving cross-modal alignment in vision language navigation via syntactic information [EB/OL]. [2022-02-13]. https://arxiv.org/pdf/2104.09580.pdf

Li L, Feng X Y, Boulkenafet Z, Xia Z Q, Li M M and Hadid A. 2016. An original face anti-spoofing approach using partial convolutional neural network//Proceedings of the 6th International Conference on Image Processing Theory, Tools and Applications (IPTA). Oulu, Finland: IEEE: 1-6 [DOI: 10.1109/IPTA.2016.7821013]

Li S, Bak S, Carr P and Wang X G. 2018. Diversity regularized spatiotemporal attention for video-based person re-identification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 369-378 [DOI: 10.1109/CVPR.2018.00046]

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]

Liang L H, Ai H Z and He K Z. 1999. Multi-template-matching based single face detection. Journal of Image and Graphics, 4(10): 825-830 (梁路宏, 艾海舟, 何克忠. 1999. 基于多模板匹配的单人脸检测. 中国图象图形学报, 4(10): 825-830) [DOI: 10.11834/jig.1999010197]

Lin J, Gan C and Han S. 2019. TSM: temporal shift module for efficient video understanding//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South):

IEEE：7082-7092［DOI：10.1109/ICCV.2019.00718］

Lin T Y，Dollár P，Girshick R，He K M，Hariharan B and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu，USA：IEEE：936-944［DOI：10.1109/CVPR.2017.106］

Lindeberg T. 2012. Scale invariant feature transform. Scholarpedia，7(5)：#10491［DOI：10.4249/scholarpedia.10491］

Liu M B，Yao H X and Gao W. 1998. Real-time human face tracking in color images. Chinese Journal of Computers，21(6)：527-532（刘明宝，姚鸿勋，高文. 1998. 彩色图像的实时人脸跟踪方法. 计算机学报，21(6)：527-532）［DOI：10.3321/j.issn:0254-4164.1998.06.007］

Liu P，Wei Y C，Meng Z B，Deng W H，Zhou J T and Yang Y. 2021b. Omni-supervised facial expression recognition：a simple baseline［EB/OL］.［2022-02-13］. https://arxiv.org/pdf/2005.08551.pdf

Liu Y J，Jourabloo A and Liu X M. 2018. Learning deep models for face anti-spoofing：binary or auxiliary supervision//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City，USA：IEEE：389-398［DOI：10.1109/CVPR.2018.00048］

Liu Y J，Stehouwer J，Jourabloo A and Liu X M. 2019. Deep tree learning for zero-shot face anti-spoofing//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition（CVPR）. Long Beach，USA：IEEE：4675-4684［DOI：10.1109/CVPR.2019.00481］

Liu Y J，Stehouwer J and Liu X M. 2020. On disentangling spoof trace for generic face anti-spoofing［EB/OL］.［2022-02-13］. https://arxiv.org/pdf/2007.09273.pdf

Lu C H，Wen X，Liu R L and Chen X. 2021. Multi-speaker emotional speech synthesis with fine-grained prosody modeling//Proceedings of the 2021 IEEE International Conference on Acoustics，Speech and Signal Processing（ICASSP）. Toronto，Canada：IEEE：5729-5733［DOI：10.1109/ICASSP39728.2021.9413398］

Lu C Y，Zhang C S，Wen F and Yan P F. 1999. Regional feature based fast human face detection. Journal of Tsinghua University（Science and Technology），39(1)：101-105（卢春雨，张长水，闻芳，阎平凡. 1999. 基于区域特征的快速人脸检测法. 清华大学学报（自然科学版），39(1)：101-105）［DOI：10.3321/j.issn:1000-0054.1999.01.027］

Lu J W，Peng Y X，Qi G J and Yu J. 2020. Guest editorial introduction to the special section on representation learning for visual content understanding. IEEE Transactions on Circuits and Systems for Video Technology，30(9)：2797-2800［DOI：10.1109/TCSVT.2020.3009095］

Lyu X G，Zhou J and Zhang C S. 2000. A novel algorithm for rotated human face detection//Proceedings of 2000 IEEE Conference on Computer Vision and Pattern Recognition. Hilton Head，USA：IEEE：760-765［DOI：10.1109/CVPR.2000.855897］

Ma B P，Su Y and Jurie F. 2014. Covariance descriptor based on bio-inspired features for person re-identification and face verification. Image and Vision Computing，32(6/7)：379-390［DOI：10.1016/j.imavis.2014.04.002］

Mahdi M K，Wu Q，Abbasnejad E and Shi J. 2020. Utilising Prior Knowledge for Visual Navigation：Distil and Adapt［EB/OL］.［2022-02-13］. https://arxiv.org/pdf/2004.03222v2.pdf

Maninis K K，Caelles S，Chen Y，Pont-Tuset J，Leal-Taixé L，Cremers D and Van Gool L. 2019. Video object segmentation without temporal information. IEEE Transactions on Pattern Analysis and Machine Intelligence，41(6)：1515-1530［DOI：10.1109/TPAMI.2018.2838670］

Marino K，Chen X L，Parikh D，Gupta A and Rohrbach M. 2021. KRISP：integrating implicit and symbolic knowledge for open-domain knowledge-based VQA//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville，USA：IEEE：14106-14116［DOI：10.1109/CVPR46437.2021.01389］

Miao J，Yin B C，Wang K Q，Shen L S and Chen X C. 1999. A hierarchical multiscale and multiangle system for human face detection in a complex background using gravity-center template. Pattern Recognition，32(7)：1237-1248

Murahari V，Batra D，Parikh D and Das A. 2020. Large-scale pretraining for visual dialog：a simple state-of-the-art baseline//Proceedings of the 16th European Conference on Computer Vision. Glasgow，UK：Springer：336-352［DOI：10.1007/978-3-030-58523-5_20］

Navaneet K L，Todi V，Babu R V and Chakraborty A. 2019. All for one：frame-wise rank loss for improving video-based person re-identification//Proceedings of 2019 IEEE International Conference on Acoustics，Speech and Signal Processing. Brighton，UK：IEEE：2472-2476［DOI：10.1109/ICASSP.2019.8682292］

Otberdout N，Daoudi M，Kacem A，Ballihi L and Berretti S. 2022. Dynamic facial expression generation on hilbert hypersphere with conditional wasserstein generative adversarial nets. IEEE Transactions on Pattern Analysis and Machine Intelligence，44(2)：848-863［DOI：10.1109/TPAMI.2020.3002500］

Qi J X，Niu Y L，Huang J Q and Zhang H W. 2020. Two causal principles for improving visual dialog//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle，USA：IEEE：10857-10866［DOI：10.1109/CVPR42600.2020.01087］

Qi Y K，Pan Z Z，Hong Y C，Yang M H，van den Hengel A and Wu Q. 2021. The road to know-where：an object-and-room informed sequential BERT for indoor vision-language navigation［EB/OL］.［2022-02-13］. https://arxiv.org/pdf/2104.04167.pdf

Qing L Y，Shan S G，Chen X L and Gao W. 2006. Face recognition under varying lighting based on the harmonic images. Chinese Journal of Computers，29(5)：760-768（卿来云，山世光，陈熙霖，

高文. 2006. 基于球面谐波基图像的任意光照下的人脸识别. 计算机学报, 29(5): 760-768) [doi: 10.3321/j.issn:0254-4164.2006.05.011]

Qiu Z F, Yao T and Mei T. 2017. Learning spatio-temporal representation with pseudo-3D residual networks//Proceedings of 2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE: 5534-5542 [DOI: 10.1109/ICCV.2017.590]

Ramnath K and Hasegawa-Johnson M. 2021. Seeing is knowing! Fact-based visual question answering using knowledge graph embeddings [EB/OL]. [2022-02-13]. https://arxiv.org/pdf/2012.15484.pdf

Savvides M, Kumar B V K V and Khosla P K. 2004. Eigenphases vs// Proceedings of the 17th International Conference on Pattern Recognition. Cambridge, UK: IEEE: 810-813 [DOI: 10.1109/ICPR.2004.1334652]

Shan S G, Gao W, Cao B and Zhao D B. 2003. Illumination normalization for robust face recognition against varying lighting conditions// Proceedings of 2003 IEEE International SOI Conference. Nice, France: IEEE: 157-164 [DOI: 10.1109/AMFG.2003.1240838]

Shao R, Lan X Y, Li J W and Yuen P C. 2019. Multi-adversarial discriminative deep domain generalization for face presentation attack detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE: 10015-10023 [DOI: 10.1109/CVPR.2019.01026]

Shao Z W, Liu Z L, Cai J F and Ma L Z. 2021a JAA-Net: joint facial action unit detection and face alignment via adaptive attention. International Journal of Computer Vision, 129(2): 321-340

Shao Z W, Zhu H L, Tang J S, Lu X Q and Ma L Z. 2021b Explicit Facial Expression Transfer via Fine-Grained Representations. IEEE Transactions on Image Processing. 30: 4610-4621

Shashua A and Riklin-Raviv T. 2001. The quotient image: class-based re-rendering and recognition with varying illuminations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2): 129-139 [DOI: 10.1109/34.908964]

Simonyan K and Zisserman A. 2014. Two-stream convolutional networks for action recognition in videos [EB/OL]. [2022-02-13]. https://arxiv.org/pdf/1406.2199.pdf

Su Z, Zhu C, Dong Y P, Cai D Q, Chen Y R and Li J G. 2018. Learning visual knowledge memory networks for visual question answering//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7736-7745 [DOI: 10.1109/CVPR.2018.00807]

Taigman Y, Yang M, Ranzato M A and Wolf L. 2014. DeepFace: closing the gap to human-level performance in face verification//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 1701-1708 [DOI: 10.1109/CVPR.2014.220]

Tang J S, Shao Z and Ma L Z. 2020. Fine-Grained Expression Manipulation Via Structured Latent Space// Proceedings of 2020 IEEE International Conference on Multimedia and Expo (ICME). London, UK: IEEE: 1-6 [10.1109/ICME46284.2020.9102852]

Tang J S, Shao Z and Ma L Z. 2021. EGGAN: Learning Latent Space for Fine-Grained Expression Manipulation. IEEE MultiMedia, 28(3): 42-51

Tan X, Xu K, Cao Y, Zhang Y, Ma L Z and Lau Rynson W H. 2021. Night-time scene parsing with a large real dataset. IEEE Transactions on Image Processing, 30: 9085-9098 [DOI: 10.1109/TIP.2021.3122004]

Tran D, Bourdev L, Fergus R, Torresani L and Paluri M. 2015. Learning spatiotemporal features with 3D convolutional networks//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 4489-4497 [DOI: 10.1109/ICCV.2015.510]

Tu T, Ping Q, Thattai G, Tur G and Natarajan P. 2021. Learning better visual dialog agents with pretrained visual-linguistic representation// Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 5618-5627 [DOI: 10.1109/CVPR46437.2021.00557]

Vries H D, Strub F, Chandar S, Pietquin O and Courville A. 2016. GuessWhat?! Visual object discovery through multi-modal dialogue [EB/OL]. [2022-02-13]. https://arxiv.org/pdf/1611.08481.pdf

Wang H and Schmid C. 2013. Action recognition with improved trajectories//Proceedings of 2013 IEEE International Conference on Computer Vision. Sydney, Australia: IEEE: 3551-3558 [DOI: 10.1109/ICCV.2013.441]

Wang H T, Li S Z and Wang Y S. 2004. Face recognition under varying lighting conditions using self quotient image//Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition. Seoul, Korea (South): IEEE: 819-824 [DOI: 10.1109/AFGR.2004.1301635]

Wang J G and Tan T N. 2000. A new face detection method based on shape information. Pattern Recognition Letters, 21(6/7): 463-471 [DOI: 10.1016/S0167-8655(00)00008-8]

Wang K, Peng X J, Yang J F, Lu S J and Qiao Y. 2020a. Suppressing uncertainties for large-scale facial expression recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 6896-6905 [DOI: 10.1109/CVPR42600.2020.00693]

Wang K, Peng X J, Yang J F, Meng D B and Qiao Y. 2020b. Region attention networks for pose and occlusion robust facial expression recognition. IEEE Transactions on Image Processing, 29: 4057-4069 [DOI: 10.1109/TIP.2019.2956143]

Wang L M, Xiong Y J, Wang Z, Qiao Y, Lin D H, Tang X O and van Gool L. 2016. Temporal segment networks: towards good practices for deep action recognition//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 20-36 [DOI: 10.1007/978-3-319-46484-8_2]

Wang P, Wu Q, Shen C H, Dick A and van den Hengel A. 2017.

Explicit knowledge-based reasoning for visual question answering//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press: 1290-1296

Wang P, Wu Q, Shen C H, Dick A and van den Hengel A. 2018a. FVQA: fact-based visual question answering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(10): 2413-2427 [DOI: 10.1109/TPAMI.2017.2754246]

Wang W X, Sun Q, Fu Y W, Chen T, Cao C J, Zheng Z Q, Xu G Q, Qiu H, Jiang Y G and Xue X Y. 2019. Comp-GAN: compositional generative adversarial network in synthesizing and recognizing facial expression//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 211-219 [DOI: 10.1145/3343031.3351032]

Wang Y, Joty S, Lyu M R, King I, Xiong C M and Hoi S C H. 2020c. VD-BERT: a unified vision and dialog transformer with BERT [EB/OL]. [2022-02-13]. https:arxiv.org/pdf/2004.13278.pdf

Wang Y Q. 2014. An analysis of the Viola-Jones face detection algorithm. Image Processing on Line, 4: 128-148 [DOI: 10.5201/ipol.2014.104]

Wang Z N, Zeng F W, Liu S C and Zeng B. 2021c. OAENet: oriented attention ensemble for accurate facial expression recognition. Pattern Recognition, 112: #107694 [DOI: 10.1016/j.patcog.2020.107694]

Wojciech Z, Zoran Z and Ben J A K. 2005. Keeping track of humans: have I seen this person before?//Proceedings of 2005 IEEE International Conference on Robotics and Automation. Barcelona, Spain: IEEE: 2081-2086 [DOI: 10.1109/ROBOT.2005.1570420]

Wong W K, Lai Z H, Wen J J, Fang X Z and Lu Y W. 2017. Low-rank embedding for robust image feature extraction. IEEE Transactions on Image Processing, 26(6): 2905-2917 [DOI: 10.1109/TIP.2017.2691543]

Wu J J, Jiang J G, Qi M B and Liu H. 2019. Independent metric learning with aligned multi-part features for video-based person re-identification. Multimedia Tools and Applications, 78(20): 29323-29341 [DOI: 10.1007/s11042-018-7119-6]

Wu K, Zhu H L, Hao Y Y and Ma L Z. 2017. Cascade regression based multi-pose face alignment. Journal of Image and Graphics, 22(2): 257-264 (伍凯, 朱恒亮, 郝阳阳, 马利庄. 2017. 级联回归的多姿态人脸配准. 中国图象图形学报, 22(2): 257-264) [DOI: 10.11834/jig.20170214]

Wu Q, Wang P, Shen C H, Dick A and Van Den Hengel A. 2016. Ask me anything: free-form visual question answering based on knowledge from external sources//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 4622-4630 [DOI: 10.1109/CVPR.2016.500]

Wu Q, Shen C H, Wang P, Dick A and van den Hengel A. 2018. Image captioning and visual question answering based on attributes and external knowledge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(6): 1367-1381 [DOI: 10.1109/TPAMI.2017.2708709]

Wu W S, Chang T and Li X M. 2021. Visual-and-language navigation: a survey and taxonomy [EB/OL]. [2022-02-03]. https://arxiv.org/pdf/2108.11544.pdf

Xie S Y, Hu H F and Chen Y Z. 2021. Facial expression recognition with two-branch disentangled generative adversarial network. IEEE Transactions on Circuits and Systems for Video Technology, 31(6): 2359-2371 [DOI: 10.1109/TCSVT.2020.3024201]

Xing X, Wang K Q and Shen L S. 2000. A real-time algorithm for tracking human faces based on organ tracking. Acta Electronica Sinica, 28(6): 29-31 (邢昕, 汪孔桥, 沈兰荪. 2000. 基于器官跟踪的人脸实时跟踪方法. 电子学报, 28(6): 29-31) [DOI: 10.3321/j.issn:0372-2112.2000.06.008]

Xiong X H and de la Torre F. 2013. Supervised descent method and its applications to face alignment//Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, USA: IEEE: 532-539 [DOI: 10.1109/CVPR.2013.75]

Xu B H, Ye H, Zheng Y B, Wang H, Luwang T and Jiang Y G. 2019. Dense dilated network for video action recognition. IEEE Transactions on Image Processing, 28(10): 4941-4953 [DOI: 10.1109/TIP.2019.2917283]

Yan J J, Lei Z, Wen L Y and Li S Z. 2014. The fastest deformable part model for object detection//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 2497-2504 [DOI: 10.1109/CVPR.2014.320]

Yan Y, Huang Y, Chen S, Shen C H and Wang H Z. 2020. Joint deep learning of facial expression synthesis and recognition. IEEE Transactions on Multimedia, 22(11): 2792-2807 [DOI: 10.1109/TMM.2019.2962317]

Yang J W, Lei Z and Li S Z. 2014. Learn convolutional neural network for face anti-spoofing [EB/OL]. [2022-02-03]. https://arxiv.org/pdf/1408.5601.pdf

Yang X, Luo W H, Bao L C, Gao Y, Gong D H, Zheng S B, Li Z F and Liu W. 2019. Face anti-spoofing: model matters, so does data//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3502-3511 [DOI: 10.1109/CVPR.2019.00362]

Yu J, Zhu Z H, Wang Y J, Zhang W F, Hu Y and Tan J L. 2020. Cross-modal knowledge reasoning for knowledge-based visual question answering. Pattern Recognition, 108: #107563 [DOI: 10.1016/j.patcog.2020.107563]

Yu J H, Zhang C Y, Song Y and Cai W D. 2021. ICE-GAN: identity-aware and capsule-enhanced GAN with graph-based reasoning for micro-expression recognition and synthesis//Proceedings of 2021 International Joint Conference on Neural Networks (IJCNN). Shenzhen, China: IEEE: 1-8 [DOI: 10.1109/IJCNN52387.2021.9533988]

Zafeiriou S, Zhang C and Zhang Z Y. 2015. A survey on face detection in the wild: past, present and future. Computer Vision and Image

Understanding, 138: 1-24 [DOI: 10.1016/j.cviu.2015.03.015]

Zhang F F, Zhang T Z, Mao Q R and Xu C S. 2020a. Geometry guided pose-invariant facial expression recognition. IEEE Transactions on Image Processing, 29: 4445-4460 [DOI: 10.1109/TIP.2020.2972114]

Zhang F F, Zhang T Z, Mao Q R and Xu C S. 2020b. A unified deep model for joint facial expression recognition, face synthesis, and face alignment. IEEE Transactions on Image Processing, 29: 6574-6589

Zhang K Y, Yao T P, Zhang J, Tai Y, Ding S H, Li J L, Huang F Y, Song H C and Ma L Z. 2020c. Face anti-spoofing via disentangled representation learning//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 641-657 [DOI: 10.1007/978-3-030-58529-7_38]

Zhang L Y, Liu S C, Liu D H, Zeng P P, Li X P, Song J K and Gao L L. 2021a. Rich visual knowledge-based augmentation network for visual question answering. IEEE Transactions on Neural Networks and Learning Systems, 32(10): 4362-4373 [DOI: 10.1109/TNNLS.2020.3017530]

Zhang W, Li Y M, Lu W Z, Xu X S, Liu Z W and Ji X Y. 2019. Learning intra-video difference for person re-identification. IEEE Transactions on Circuits and Systems for Video Technology, 29(10): 3028-3036 [DOI: 10.1109/TCSVT.2018.2872957]

Zhang Y F, Ming J and Qi Z. 2021b. Explicit Knowledge Incorporation for Visual Reasoning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual: IEEE/CVF

Zhao L, Gao L L, Guo Y Y, Song J K and Shen H T. 2021a. SKANet: structured knowledge-aware network for visual dialog//Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME). Shenzhen, China: IEEE: 1-6 [DOI: 10.1109/ICME51207.2021.9428279]

Zhao Y, Yang L, Pei E C, Oveneke M C, Alioscha-Perez M, Li L F, Jiang D M and Sahli H. 2021b. Action unit driven facial expression synthesis from a single image with patch attentive GAN. Computer Graphics Forum, 40(6): 47-61 [DOI: 10.1111/cgf.14202]

Zhao Z Q, Liu Q S and Wang S M. 2021c. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Transactions on Image Processing, 30: 6544-6556 [DOI: 10.1109/TIP.2021.3093397]

Zhi R C, Xu H R, Wan M and Li T T. 2019. Combining 3D convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition. IEICE Transactions on Information and Systems, E102.D(5): 1054-1064 [DOI: 10.1587/transinf.2018EDP7153]

Zhou J, Lu C Y, Zhang C S and Li Y D. 2000. A survey of automatic human face recognition. Acta Electronica Sinica, 28(4): 102-106 (周杰, 卢春雨, 张长水, 李衍达. 2000. 人脸自动识别方法综述. 电子学报, 28(4): 102-106) [DOI: 10.3321/j.issn:0372-2112.2000.04.027]

Zhou Z, Huang Y, Wang W, Wang L and Tan T N. 2017. See the forest for the trees: joint spatial and temporal recurrent neural networks for video-based person re-identification//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 6776-6785 [DOI: 10.1109/CVPR.2017.717]

Zhu F D, Zhu Y, Chang X J and Liang X D. 2020a. Vision-language navigation with self-supervised auxiliary reasoning tasks//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 10009-10019 [DOI: 10.1109/CVPR42600.2020.01003]

Zhu X K, Jing X Y, You X G, Zhang X Y and Zhang T P. 2018. Video-based person re-identification by simultaneously learning intra-video and inter-video distance metrics. IEEE Transactions on Image Processing, 27(11): 5683-5695 [DOI: 10.1109/TIP.2018.2861366]

Zhu Y, Weng Y, Zhu F D, Liang X D, Ye Q X, Lu Y T and Jiao J B. 2021. Self-motivated communication agent for real-world vision-dialog navigation//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 1574-1583 [DOI: 10.1109/ICCV48922.2021.00162]

Zhu Y K, Zhang C, Ré C and Li F F. 2015. Building a large-scale multimodal knowledge base system for answering visual queries [EB/OL]. [2022-02-13]. https://arxiv.org/pdf/1507.05670.pdf

Zhu Z H, Yu J, Wang Y J, Sun Y J, Hu Y and Wu Q. 2020b. Mucko: multi-Layer cross-modal knowledge reasoning for fact-based visual question answering [EB/OL]. [2022-02-13]. https://arxiv.org/pdf/2006.09073.pdf

作者简介

马利庄,1963 年生,男,教授,主要研究方向为计算机图形图像、计算机视觉、智能信息处理。
E-mail:ma-lz@cs.sjtu.edu.cn

吴飞,男,教授,主要研究方向为人工智能、多媒体分析与检索和统计学习理论。E-mail:wufei@cs.zju.edu.cn
毛启容,女,教授,主要研究方向为多媒体信息处理、多媒体大数据分析。E-mail:mao_qr@ujs.edu.cn
王鹏杰,男,副教授,主要研究方向为计算机图形学、数字娱乐、文化遗产数字化保护。E-mail:pengjiewang@qq.com
陈玉珑,女,博士后,主要研究方向为多媒体设计、媒体传播。E-mail:llong_c@sjtu.edu.cn