



信息通信领域产学研合作特色期刊 十佳皖刊
第三届国家期刊奖百种重点期刊 中国科技核心期刊

ISSN 1009-6868
CN 34-1228/TN

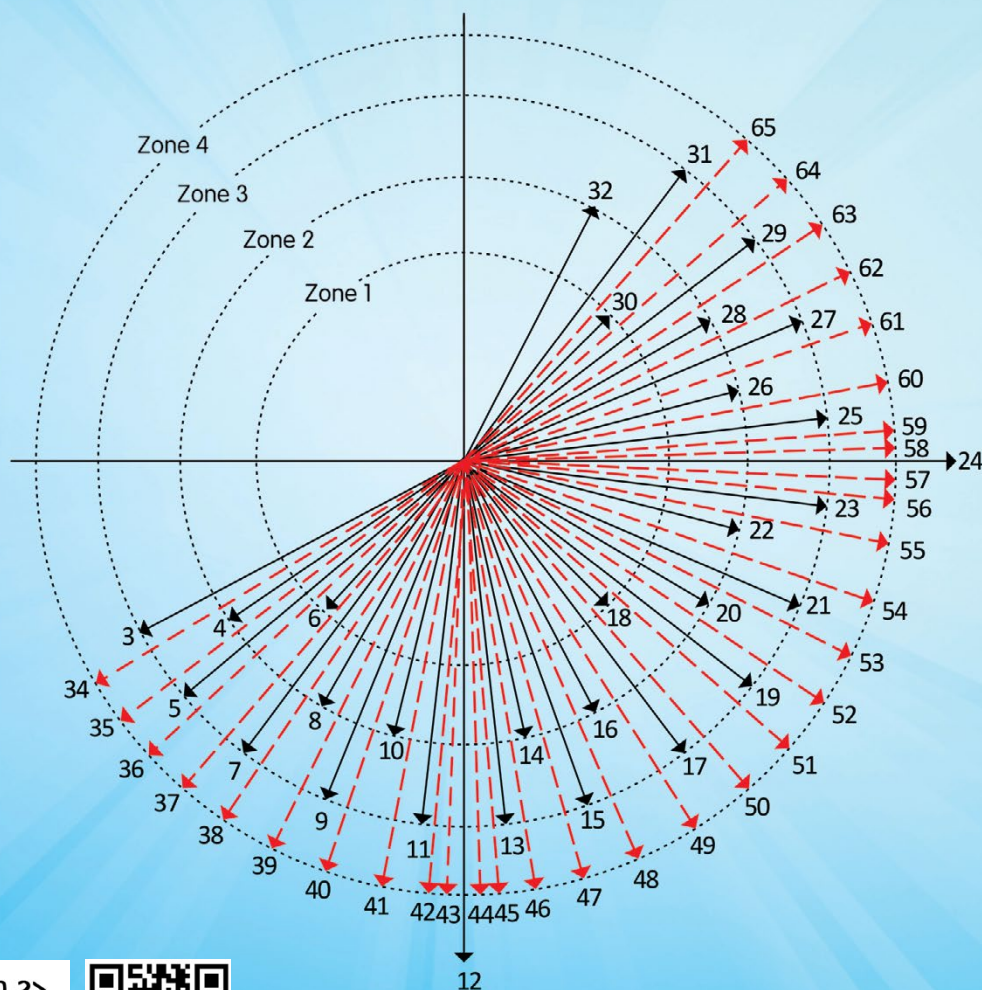
中兴通讯技术

ZTE TECHNOLOGY JOURNAL

<http://tech.zte.com.cn>

2021年2月·第1期

专题：视频技术和用户体验评测



《中兴通讯技术》第8届编辑委员会成员名单

顾问 侯为贵(中兴通讯股份有限公司创始人) 钟义信(北京邮电大学教授) 陈锡生(南京邮电大学教授)

主任 陆建华(中国科学院院士)

副主任 李自学(中兴通讯股份有限公司董事长) 糜正琨(南京邮电大学教授)

编委(按姓名拼音排序)

陈建平	上海交通大学教授	唐雄燕	中国联通网络技术研究院首席科学家
陈前斌	重庆邮电大学教授、副校长	陶小峰	北京邮电大学教授
葛建华	西安电子科技大学教授	王文博	北京邮电大学教授、副校长
管海兵	上海交通大学教授	王文东	北京邮电大学教授
郭庆	哈尔滨工业大学教授	王喜瑜	中兴通讯股份有限公司执行副总裁
洪波	中兴发展股份有限公司总裁	王翔	中兴通讯股份有限公司高级副总裁
洪伟	东南大学教授	卫国	中国科学技术大学教授
黄宇红	中国移动研究院副院长	吴春明	浙江大学教授
纪越峰	北京邮电大学教授	邬贺铨	中国工程院院士
江涛	华中科技大学教授	肖甫	南京邮电大学教授
蒋林涛	中国信息通信研究院科技委主任	解冲锋	中国电信研究院教授级高工
李尔平	浙江大学教授	徐安士	北京大学教授
李红滨	北京大学教授	徐子阳	中兴通讯股份有限公司总裁
李厚强	中国科学技术大学教授	续合元	中国信息通信研究院副总工
李建东	西安电子科技大学教授	薛向阳	复旦大学教授
李军	清华大学教授	薛一波	清华大学教授
李乐民	中国工程院院士	杨义先	北京邮电大学教授
李融林	华南理工大学教授	杨震	南京邮电大学教授、原校长
李少谦	电子科技大学教授	叶茂	电子科技大学教授
李自学	中兴通讯股份有限公司董事长	易芝玲	中国移动研究院首席科学家
林晓东	中兴通讯股份有限公司副总裁	张宏科	北京交通大学教授
刘健	中兴通讯股份有限公司高级副总裁	张平	中国工程院院士
刘建伟	北京航空航天大学教授	张卫	复旦大学教授
陆建华	中国科学院院士	张云勇	中国联通集团产品中心总经理
马建国	广东工业大学教授	赵慧玲	工业和信息化部通信科技委信息通信网络专家组组长
孟洛明	北京邮电大学教授	郑纬民	中国工程院院士
糜正琨	南京邮电大学教授	钟章队	北京交通大学教授
任品毅	西安交通大学教授	周亮	南京邮电大学教授
石光明	西安电子科技大学教授、副校长	朱近康	中国科学技术大学教授
孙知信	南京邮电大学教授	祝宁华	中国科学院半导体研究所研究员
谈振辉	北京交通大学教授、原校长		

目次

中兴通讯技术 (ZHONGXING TONGXUN JISHU)
总第 156 期 第 27 卷 第 1 期 2021 年 2 月

卷首特稿

筚路蓝缕 玉汝于成 01
王喜瑜

专题：视频技术和用户体验评测

专题导读 04
江涛，陆平

点云编码综述 05
李厚强，李礼，李竹

AVS3 视频编码关键技术及应用 10
张嘉琪，雷萌，马思伟

下一代通信助力实时分布云渲染 17
陆平，盛斌，朱方

视频质量增强模型加速算法 21
杨文哲，徐迈，白琳

基于图神经网络的视频推荐系统 27
高宸，李勇，金德鹏

脑启发视频用户体验评测关键技术 33
陶晓明，杜冰，段一平

37 超高清内容清晰度用户体验质量评价
朱文瀚，翟广涛，陶梅霞，杨小康，张文军

44 交互式视频质量评价方法研究进展
李继龙，赵雪，杨铀

48 HTTP 自适应流媒体直播系统中的用户体验
质量优化
宋新铎，张远，王博

专家论坛

54 小视频内容分析技术发展探讨
薛向阳，李斌

企业视界

60 构建智能实时网络，使能 5G 视频业务繁荣
吕达，郑清芳

技术广角

68 面向视频云微服务系统的智能运维技术
徐代刚，姜磊，梅君君

77 用于人工智能的硅基光电子芯片
白冰，裴丽，左晓燕

2021 年第 1—6 期专题计划及策划人

1. 视频技术和用户体验评测

华中科技大学教授 江涛

中兴通讯股份有限公司副总裁 陆平

2. 6G 愿景及技术挑战

中国工程院院士 张平

北京邮电大学教授 张建华

3. 边缘计算与算力网络

工信部通信科技委信息通信网络

专家组组长 赵慧玲

4. 高铁智能通信技术与应用

北京交通大学教授 艾渤

5. 低轨卫星通信技术与应用

哈尔滨工业大学教授 郭庆

6. 触觉通信技术

南京邮电大学教授 周亮

MAIN CONTENTS

ZTE TECHNOLOGY JOURNAL Vol. 27 No. 1 Feb. 2021

Guest Paper

Reflections on 5G Commercial Application **01**
WANG Xiyu

Special Topic: Video Technologies and QoE Estimation

Editorial **04**
JIANG Tao, LU Ping

A Review of Point Cloud Compression **05**
LI Houqiang, LI Li, LI Zhu

Key Technologies and Applications of AVS3 Video
Coding Standard **10**
ZHANG Jiaqi, LEI Meng, MA Siwei

Next-Generation Communications Technology Facilitates
Real-Time Distributed Cloud Rendering **17**
LU Ping, SHENG Bin, ZHU Fang

Video Quality Enhancement Model Acceleration **21**
Algorithm
YANG Wenzhe, XU Mai, BAI Lin

Video Recommender System with Graph Neural Networks **27**
GAO Chen, LI Yong, JIN Depeng

Key Techniques of Brain Inspired Video QoE Prediction **33**
TAO Xiaoming, DU Bing, DUAN Yiping

37 Quality of Experience Estimation of Ultra-High
Definition Content
ZHU Wenhao, ZHAI Guangtao, TAO Meixia,
YANG Xiaokang, ZHANG Wenjun

44 A Review of Interactive Video Quality Assessment
Methods
LI Jilong, ZHAO Xue, YANG You

48 QoE Optimization in HTTP Adaptive Live Streaming
System
SONG Jinke, ZHANG Yuan, WANG Bo

Expert Forum

54 Short Video Content Analysis Technology
XUE Xiangyang, LI Bin

Enterprise View

60 Building Smart Real-Time Networks to Enable
Prosperity of 5G Video Services
LYU Da, ZHENG Qingfang

Technology Perspective

68 Intelligent Operation and Maintenance Technology for
Video Cloud Microservice System
XU Daigang, JIANG Lei, MEI Junjun

77 Silicon Photonic Chips for Artificial Intelligence
BAI Bing, PEI Li, ZUO Xiaoyan

期刊基本参数: CN 34-1228/TN*1995*b*16*82*zh*P* ¥ 20.00*6500*15*2021-02

敬告读者

本刊享有所发表文章的版权, 包括英文版、电子版、网络版和优先数字出版版权, 所支付的稿酬已经包含上述各版本的费用。未经本刊许可, 不得以任何形式全文转载本刊内容; 如部分引用本刊内容, 请注明该内容出自本刊。



筌路蓝缕 玉汝于成

◎ 王喜瑜 / 中兴通讯股份有限公司

2020年，全球共有58个国家和地区的135个运营商宣布商用5G，但5G用户超百万的国家和地区，仅有中国、美国、韩国、日本和欧洲。5G建设规模在全球继续呈分化趋势。中国已累计建成5G基站71.8万个，拥有5G终端连接数超2亿，在全球范围内遥遥领先。短暂使用非独立组网（NSA）架构后，中国5G商用目前已加速过渡到独立组网（SA）架构。这从根本上避免了后续频繁的网络升级优化，使得运营商在垂直行业的部署更加快速灵活，给行业提供更好的网络性能保障。

规模商用加速 5G 技术成熟进程

相比于之前任何一代移动通信技术，5G有着空前的全球认知度，但这不代表它生来就成熟。实际上，正是5G的规模商用部署促进了5G技术不断进步、不断突破。

从消费者体验的角度来看，在5G的导入初期，用户面临的信号覆盖不好、语音体验差、手机能耗高等问题正在得到较好的解决。例如，大规模多输入多输出（Massive MIMO）广播波束单边带（SSB）1+X方案，可将复杂场景的垂直覆盖能力提升30%；随着终端芯片对5G语音业务（VoNR）的支持，语音体验会大幅提升；终端节能系列功能的商用，可使能耗在典型场景下的降幅达80%。公众期待的5G“杀手级”的应用，仍处于探索或生态构建的过程中，但基于视频的深度升级应用必将成为一个重要方向。

从运营商资本性支出（CAPEX）角度来看，尤其考虑到5G网络的发展定位，建设方案的选择至关重要。虽然单个5G基站的成本约为

4G基站的2.5倍，但如果把4G比作一条4车道的高速公路，5G则相当于100车道的高速公路。随着5G用户渗透率的提升和物联宽带应用的涌现，5G网络建设的边际成本也将迅速递减，因此5G网络的性价比远优于4G。目前，部分领先的运营商已经优先选择64/32通道天线的5G基站来建设网络。他们意识到，更多的天线和大带宽才会给用户带来真正的5G差异化体验，而这也将成为市场的核心竞争力。部分更加关心短期投入的客户，会优先选择相对低成本的覆盖解决方案，例如将现有频分双工（FDD）4G网络，通过软件升级至5G网络，也不得不接受性能提升有限的现实。

快速低成本地实现良好的室内覆盖是另一个需要特别关注的因素。如何保护现有分布式天线系统（DAS）系统投资，实现平滑演进，满足5G更大容量的需求，是运营商在做5G室内覆盖时需要考虑的重要问题。增强型DAS（eDAS）采用分布式多天线联合收发技术，突破传统DAS只能实现1流或者2流传输的限制，实现上/下行多流MIMO传输，提升系统容量。此外，eDAS采用5G创新算法，不再受限于传统DAS系统严苛的物理链路均衡要求，实现多流效果。eDAS技术的应用无须新增硬件，为运营商最大限度地降低了室内覆盖的成本。

从运营商运营成本（OPEX）角度来看，能耗始终是一个核心问题。运营商在2020年8月发布的数据显示，单个5G基站功耗约为4G的3~4倍。但如果把4G比作小汽车，那么5G就是大巴车。大巴车每百千米的人均能耗是小汽车的8.4%，这意味着并非单个小汽车的能耗低于大巴车，而是当运送相同数量的乘客时，大巴车与小汽车相比可以节约80%以上的能耗。即便如此，持续节能减排依然是产业界共同努力的方向，运营商与设备商正一起通过智

能算法更大幅度地降低基站能耗。以 PowerPilot 为例,该方案可以针对差异化覆盖场景、时段和基站负荷,通过引入大数据和人工智能(AI)应用,对网络话务和配置信息进行分析,“一站一策”地实现站点级、精细化、多层次节电功能的应用。通过现网验证,该方案可以降低 15% 左右的能耗。当 5G 网络负载提升后,PowerPilot 还可以做到实时识别业务和分析业务能效比、主动导航业务,并可以在不影响用户体验的基础上,合理调整用户分布,通过频间/制式间深度协同,实现整网更优能效比。

“5G 改变社会”走向现实

从 2020 年工业和信息化部组织的“绽放杯”5G 应用大赛看,“5G 改变社会”正在逐渐成为现实。从全国“两会”的全息直播到新冠肺炎疫情期间的樱花“云赏”,从云南神火铝业的天车遥控到新凤鸣集团的飞丝遥检,5G 已经开始在不同的典型行业场景中发挥价值。

- 大视频(高清/扩展现实)场景。在高清视频监控/分析、机器视觉质检、港口岸桥或挖掘机远程控制等应用场景中,端侧数据通常需要以高清图像/高清视频的方式实时采集、处理或交互,而 5G 可同时满足端侧的灵活部署和高清视频的流量诉求,让真实世界的数字化呈现成为可能。

- 时延敏感场景。行业端到端通信需要确定性的时延控制,例如电力行业的差动保护、配电网同步相量测量等场景对时延、抖动、授时精度都提出了超高要求。原有的 4G 网络已无法满足这些需求,只有 5G 网络才能应对挑战。

- 高可靠安全场景。轨道交通的列控系统、远程手术等场景与行业的核心生产系统的可靠性、生命财产的安全性密切相关。5G 网络针对行业高可靠性的特点,也在一定程度上拓展了与行业运营技术(OT)深度融合的无限可能。

在 5G 场景化推进中,建议基于成熟度并按照循序渐进的原则,对于大带宽的场景,应率先满足需求,规模实施应用;对于低时延、高可靠等相对个性化的场景,应以高价值示范应用先行,逐渐复制推进。

但碎片化和规模效益之间的平衡问题,仍

将会是 5G 行业应用推广的核心挑战。针对这一问题,中兴通讯提出了“精准云网综合解决方案”。在过去一年多的时间里,中兴通讯通过对上千个应用、近百个场景的收敛,将需求的关键技术特征和场景特征解耦,抽取成具有共性能力的组件,形成“积木式”的组件库,并通过灵活、高效的组合向上支撑各种场景的应用,再通过组件在场景上的应用试点,对组件库进行持续迭代优化。

组件可以分成两大类:一类是用来满足高性能要求的云原生、数据库、操作系统、网络资源等基础能力组件,例如与 5G 强相关的创新性组件——网络原子能力组件,它将网络资源预留、双连接、容灾备份、边缘服务质量(QoS)、小颗粒切片、极简 5G 核心网(5GC)等关键技术,封装为网络原子能力,再被灵活调配、快速组合成面向不同场景的 5G 专网定制化服务;另一类是支撑应用创新孵化的视频组件、AI 组件、安全组件、大数据组件等面向场景和应用的组件,例如视频组件可满足视频物联、视频会议、点直播和交互直播的应用需求,AI 组件满足智能化制造、视频监控分析、自动导引车(AGV)/机器人云端控制以及定位等应用的使用需求。

这种积木式的组件化平台,实现了定制化和规模化之间的平衡,并通过模块化组件服务,支持业务敏捷发放,具有低成本、快速迭代、“一点创新,多点复制”的优点,让企业真正实现“应用随心”。

5G 行业应用加速云网融合

视频业务不仅要求网络具备高带宽、强交互、确定时延的业务属性,还对由视频基础能力中台、内容分发网络(CDN)、边缘计算技术(MEC)等多种技术组件构成的云网有效协同提出了要求。运营商可利用网络优势,通过网融云(视频渲染能力上云)和网融端(终端能力上云),构建云/网/端/业一体化视频新平台,提升网络价值,增强用户粘性。

园区专网应能经济便捷地进行部署。边缘计算等 5G 技术的成熟,给园区数字化改造带来了新的机会。企业需要将本地移动网络里的

数据流在本地进行卸载,才能满足应用低时延、高可靠性的要求,同时确保企业生产的信息安全。中兴通讯 NodeEngine 方案,就是通过在 5G 网络设备上增加单板,实现边缘算力的站点级部署,既满足数据不出园区的需求,又能有效降低企业园区专网的部署成本,缩短部署工期。

专属云与公有云相结合的混合云可能更加符合中国企业的需求。中国的公有云已经发展到比较成熟的阶段,头部企业正在形成。但是,中国企业仍有其特别诉求:一方面,大部分大型企业和组织都有自己的科研团队,自身具备开发和维护能力;另一方面,数据不出园区的需求,也需要企业生产域和信息域的系统能够快速持续迭代。中兴通讯 TCF 云底座方案,向下可以兼容各种基础设施即服务(IaaS);向上可以提供服务化接口,为上层应用屏蔽跨平台细节。同时,可以根据公有云、边缘云,甚至单机服务器等不同场景进行协议裁减,帮助企业客户在享有专属云安全与成本优势的同时,兼享公有云的快速部署与随时可及的便利。

中兴通讯甘当数字经济的“筑路者”

2020 年初,中国正式提出了“新基建”的发展思路,5G 基础网络建设将作为产业转型升级的重要抓手。利用“5G+ 大数据 +AI+ 云化算力”,构建以科技为核心的新型工业制造体系,将推动智能制造与产业数字化转型,成为生产力发展的新引擎。

三年新基建,十年新动能。中国三大电信运营商将与领头的公有云运营商一起,在这场轰轰烈烈的产业数字化进程中,充当核心数字运营体的角色。在此基础之上,也将有无数的创新型企业涌现并壮大。

在数字经济的大潮中,中兴通讯坚持不做公有云服务,坚持不碰客户数据,坚持不伤害客户与产业生态链,坚持自己在国家产业链中的定位,坚持将最难的事做到最好,甘当数字经济的“筑路者”,以持续的技术突破创新和可靠供应链,助力数字运营体的发展壮大,促进企业的数字化转型。在核心技术创新方面,除了端到端数据信息通信技术(DICT)解决方案之外,中兴通讯将持续向下扎根,如自主设计软基带芯片、可编程交换芯片及全系列连接芯片,为中国大型工业基础设施提供应用级的工业操作系统,为银行提供交易系统数字化转型的分布式数据库等;在行业赋能方面,中兴通讯将与生态伙伴合作,提供云底座、云化 AGV 以及 DevOps 工具链及集成工具;在应用实践方面,中兴通讯在南京打造智能制造基地,实现“用 5G 制造 5G”。另外,中兴通讯也在持续深化自身的数字化和智能化转型,坚定地向着“极致的云公司”的愿景迈进。

不驰于空想,不骛于虚声,初心因为坚持而伟大。筌路蓝缕,愿玉汝于成。

作者简介



王喜瑜,中兴通讯股份有限公司执行副总裁、CTO、移动网络和移动通讯多媒体技术国家重点实验室学术委员会主任,教授级高工;1998 年入职中兴通讯,先后担任无线研究院院长、技术规划部部长,现全面负责中兴通讯系统产品规划及研发;曾获国家科学技术进步二等奖、广东省科学技术进步一等奖、中国通信学会科技进步一等奖等多项荣誉。

视频技术和用户体验评测专题导读



专题策划人 江涛



华中科技大学教授、IEEE Fellow、中国第6代移动通信技术研发总体专家组成员、教育部“长江学者”特聘教授、国家杰出青年科学基金获得者，享受国务院政府特殊津贴，入选中共中央组织部第二批“万人计划”科技创新领军人才；长期从事宽带移动、多媒体通信、天地一体化信息网络等研究，先后承担国家重点研发计划、“973”计划、“863”计划和国家科技重大专项等项目和课题；以第一完成人身份先后获国家技术发明二等奖、湖北省自然科学奖一等奖、中国电子学会自然科学奖一等奖等，所提出的校验级联极化码（PCC Polar Code）被正式采纳为5G标准。

专题策划人 陆平



中兴通讯股份有限公司副总裁、移动网络和移动通讯多媒体技术国家重点实验室副主任；研究方向包括云计算、大数据、增强现实、基于多媒体服务的技术；支持和参与了国家科技重大专项、国家科技支撑项目等；发表多篇论文，撰写了《物联网能力开发与应用》《云计算中的大数据技术与应用》等多部著作。

现代信息化社会对视频业务的需求呈爆发式增长，人们希望在任何地点、任何时间都能享受到超高质量的视频业务服务。高速率、低时延的新一代通信技术与人工智能技术相结合，可助力视频技术向高清晰度、高交互性和虚实结合等新方向发展，并与工业制造、医疗、教育等垂直行业有机融合。

视频质量评估为衡量视频技术的效能提供了标尺，其重要性不言而喻。传统的视频质量评价指标主要是客观的，包括画面清晰度、时延和抖动等，并不一定适用于增强现实（AR）、虚拟现实（VR）和交互式视频等新兴业务。当前的视频技术发展趋势是从“人”的角度去寻找视频质量评价指标。“人”这个新的维度，会催生一批崭新的视频质量测评框架，并进一步推动视频编解码、传输和分发技术的发展。

点云编码可压缩用于三维物体、空间数字化建模的点云数据，是新一代视频编解码技术标准的重要组成部分。《点云编码综述》与《AVS3 视频编码关键技术及应用》分别从点云编码和全局角度介绍了视频编解码技术领域的最新研究成果。从视频的实时渲染到视频特征的智能分析，新一代视频技术对算力和人工智能技术的需求越来越高。《下一代通

信助力实时分布云渲染》从通信的角度阐述了视频技术发展的新可能，即借助5G通信和云计算技术来克服终端视频实时渲染所面临的算力瓶颈问题。《视频质量增强模型加速算法》与《基于图神经网络的视频推荐系统》分别运用深度神经网络技术与图神经网络技术，提升视频增强与视频内容推荐算法的效能。在用户体验测评方面，《脑启发视频用户体验评测关键技术》从全新的脑电响应数据入手，实现小样本下对用户体验质量更为精准的评估。《超高清内容清晰度用户体验质量评价》聚焦超高分辨率的视频质量评估，并可辨别真假超高清内容。《交互式视频质量评价方法研究进展》从主观质量评估与客观质量评估两个方面，对新一代视频业务中的交互式视频进行了综述。《HTTP自适应流媒体直播系统中的用户体验质量优化》分析了影响直播系统中用户体验质量的关键因素，并总结了用户体验优化策略。

本期论文能够反映出在视频技术与用户体验评测这一领域里中国研究者的主要成果与学术观点。希望这些文章能为多媒体新技术领域的研究提供多种可能，并起到积极的推动作用。

江涛 陆平

2021年1月16日

DOI: 10.12142/ZTETJ.202101002
收稿日期: 2021-01-17



点云编码综述

A Review of Point Cloud Compression

李厚强 / LI Houqiang¹, 李礼 / LI Li¹, 李竹 / LI Zhu²

(1. 中国科学技术大学, 中国 合肥 230027;
2. 美国密苏里大学堪萨斯分校, 美国 堪萨斯城 64110)
(1. University of Science and Technology of China, Hefei 230027, China;
2. University of Missouri-Kansas City, Kansas City 64110, USA)

摘要: 点云编码是支撑点云广泛应用的关键技术之一, 是近期技术研究和标准化领域的热点。对点云几何信息和属性信息编码技术演进进行了回顾, 并针对稠密点云和稀疏点云的几种典型编码方法的编码效率进行了比较。未来点云编码研究将集中于利用帧间预测去除动态点云的不同帧之间的相关性, 以及端到端点云编码、任务驱动的点云编码等方面。

关键词: 3D 点云编码; 几何信息编码; 属性信息编码

Abstract: 3D point cloud compression is one of the key technologies supporting the wide-spread use of point clouds. Recently, it is one of the focuses for both research and standardization groups. The latest advance of the compression technologies for both the 3D point cloud geometry and attribute information is reviewed. Compression efficiencies of several typical compression technologies for both the 3D dense and sparse point clouds are compared. In the future, more studies will focus on inter-frame prediction to exploit the correlations between different frames in 3D dynamic point clouds, end-to-end point cloud compression, and task-driven point cloud compression.

Keywords: 3D point cloud compression; geometry information coding; attribute information coding

DOI: 10.12142/ZTETJ.202101003

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210122.1752.002.html>

网络出版日期: 2021-01-25

收稿日期: 2020-12-08

点云是一系列高维空间点(例如三维空间点)的集合。每一个点包含几何信息(x, y, z)以及颜色和反射率等属性信息。根据点云中点的密度, 点云可以粗略地分为稠密点云和稀疏点云。稠密点云可以用来精细重建3D物体例如人物等, 可被广泛应用于虚拟现实和增强现实。稠密点云重建的

3D物体支持6自由度, 相比360°全景视频仅能支持3自由度, 可以给用户带来更好的视觉体验。稀疏点云可以高精度重建3D场景, 结合2D摄像头采集的高清图像视频, 可被用于自动驾驶和机器人视觉等应用中。由于点云数据量巨大, 点云编码成为了上述应用中不可或缺的一环。相比成熟的图像视频编码技术, 点云编码由于其独有的特点成为近期的研究热点。图像视频中的像素在2D空间中均匀分布, 而点云中的点在3D空间的分布是

稀疏且无规律的。点云的稀疏性是指3D空间仅有很小一部分3D位置被点占用。从压缩的角度来看, 相比于编码整个3D空间, 仅仅编码被占用的部分信息会更加高效。同时, 点云的无规律性使得点云的不同点之间的相关性难以被有效去除。点云编码可以根据其包含的信息分为两个部分: 几何信息编码指明空间中哪些位置存在3D点, 属性信息编码指明空间中3D点的颜色和反射率等属性信息。在大部分点云编码算法中, 都是先编码几何信

基金项目: 国家自然科学基金(62021001); 中国科学技术大学引进人才科研启动经费(KY21000000108)

息,然后基于重建的几何信息和原始点云对点云进行重着色,最后编码重着色之后的属性信息。

1 几何信息编码

几何信息编码主要分为3类:基于树结构的方法、基于表面近似的方法、基于映射的方法。下面我们将分别对这些方法进行详细介绍。

1.1 基于树结构的方法

基于树结构的方法是最直接的几何信息编码方法。其基本思想是对包含点云的最小立方体以树的形式进行迭代划分,如果划分完的子立方体包含点,则编码“1”,且会被进一步划分;不包含点,则编码“0”,且不会被进一步划分。在基于树结构的方法中,使用的树结构通常为八叉树和二叉树。

早在21世纪初,基于八叉树的方法就被用于编码点云^[1]。基于八叉树的方法首先迭代地把包含点云的最小立方体划分为8个子立方体,然后用一个字节编码8个子立方体是否包含点这一信息。由于父节点和子节点,以及相邻节点的字节之间存在很强的相关性,通常使用基于上下文的算术编码进一步去除该相关性。由于该方法简单有效,它在国际动态图像专家组征集的所有稀疏点云编码方法中取得了优胜,最终发展成为基于几何信息的点云编码标准之一^[2]。为了进一步提升编码效率,我们提出了使用该字节中1的个数和组合来代表该字节,1的个数和组合也可以使用父节点和邻近节点近似成的面来估计^[1]。八叉树的主要缺点是表征八叉树需要的比特数会随着树深度的增加而急剧增加,所以使用二叉树来编码几何信息的方法被提出^[3]。点云编码使用基于数据的二叉树可以一定程度上缓解因深度增加所需要的比特数,但是基于数据

而非空间的二叉树需要编码分割节点信息,尤其在树的深度较浅时会消耗大量比特。

1.2 基于表面近似的方法

因为完整点云很难被近似成一个参数化的表面,所以基于表面近似的方法通常与基于树的方法结合使用。首先使用八叉树或二叉树把点云分割成互不包含的小立方体,然后小立方体被近似成表面以进一步编码。表面近似的方法的本质是降维,编码一个小立方体相当于编码三维信息,而把小立方体近似成一个表面则仅需要编码二维信息。

在所有基于表面近似的方法中,最常用的表面是平面。我们首先对点云进行八叉树划分,划分到一定的深度后,再使用平面对立方体中的点进行近似,编码平面与立方体的交点来代表平面,最后对平面进行采样恢复最终的点。该方法在国际运动图像专家组征集的所有静态稠密点云编码方法中取得了优胜,最终发展成为基于几何信息的点云编码标准之一^[2]。除了使用八叉树作为树分割的方式,二叉树也可以作为一种树分割的方式使用。除了使用采样来恢复最终的点,也可以使用四叉树对近似形成的平面进行基于树结构的编码^[3]。基于平面的编码方法相比于基于树的编码方法,在低码率上可以带来明显的性能提升,但是由于平面近似始终存在误差,基于平面的编码方法无法实现无损编码。除此之外,为了进一步提升表面近似精度,二阶曲面也被用于表面近似^[4],但是二阶曲面相比平面需要传输更多的头信息,这会导致编码性能提升有限。

1.3 基于映射的方法

基于映射的方法最初是针对网格

(mesh)编码设计的。近些年来,基于映射的方法逐渐开始被用于点云编码。基于映射的方法的基本思想是把点云从3D空间映射到2D空间,然后使用成熟的2D图像视频编码方法进行编码。此方法的核心在于找到一种合适的映射,既能在投影的过程中减少点的损失,又能使投影之后的图像视频具有较高的时空相关性以更好地利用2D图像视频编码方法中高效的预测技术。

为了尽可能在投影过程中减少点的丢失,我们以一定顺序扫描点云八叉树,把3D点云转化为2D图像或视频^[5]。这种投影方式不会造成任何点的丢失,但形成的2D图像视频时空相关性弱,编码效率低。为了提高2D图像视频的时空相关性,我们提出把点云完整地投影到包围着该点云的圆柱体或立方体上^[6]。此方法的2D图像视频编码效率高,但会造成部分被遮挡的连续点丢失,从而导致3D点云质量较差。为了兼顾投影点的数量和2D图像视频编码效率,我们提出把具有相似法向量的点按片投影到包围该点云的立方体上,不同的点云会形成几十到数百个片^[2]。此基于片的投影不会导致被遮挡的连续大量点丢失,因为它们会形成一个新的片投影到2D空间。此外,基于片的投影方法把具有相似法向量的点投影成一个片,使得属于同一个片的点的深度方差较小,有利于提升编码效率。该方法在国际运动图像专家组征集的所有动态点云编码方法中取得了优胜,最终发展成为基于视频的点云编码标准^[2]。

2 属性信息编码

属性信息编码主要可以分为3类:基于变换的方法、基于预测的方法、基于映射的方法。下面我们将分别对这些方法进行详细介绍。

2.1 基于变换的方法

变换是编码中一种常用的去相关方法。基于变换的属性信息编码方法的基本思想是利用重建的几何信息来设计一个内容自适应的属性信息变换,以去除属性信息之间的相关性。去除相关性之后的属性信息经过量化和熵编码后形成属性信息码流。

为了充分利用已经编码的几何信息,我们提出使用图变换的方法对属性信息进行变换编码^[7]。首先根据点与点之间的距离构建图,然后对图拉普拉斯矩阵进行特征值分解,最后使用特征向量构建的变换对属性信息进行变换。除此之外,我们还提出使用高斯过程来近似点与点之间的关系,推导出高斯过程对应的K-L变换来编码属性信息^[8]。上述方法能达到较好的编码性能,但是需要进行复杂的特征值分解。这会导致很高的复杂度,不利于实际使用。为了更好地取得编码性能和复杂度之间平衡,我们提出使用基于区域的自适应分层变换对属性信息进行编码^[9]。基于区域的自适应分层变换本质上是加权Haar小波变换。根据八叉树的每一个子节点包含的点的数量,对属性信息进行加权小波变换,以利用几何信息。基于区域的自适应分层变换被基于几何信息的点云编码标准采纳,成为被推荐的静态点云属性编码方法^[2]。除了以上常规的基于变换的编码方法,基于几何信息的稀疏表达变换也被用于压缩属性信息^[10],但是稀疏位置信息的编码限制了其效率。

2.2 基于预测的方法

除了变换以外,预测也是一种常用于编码的去相关方法。不同于变换对信号进行旋转使得其更适合编码,预测本质上是已编码的信息作为条件,使用条件熵代替原信号的熵,从

而提升编码效率。和变换一样,预测之后的信号经过量化和熵编码后形成码流。

在图像视频编码中,基于邻近已重建图像块对当前图像块进行预测,在各代图像视频编码标准中一直沿用。在基于八叉树的几何信息编码中,点云被分割成多个等大的小立方体,基于邻近已经重建的小立方体的属性信息对当前小立方体进行预测,是2D预测编码到3D预测编码的一个简单扩展^[11]。但是一方面,点云的稀疏性导致邻近可用预测块较少,3D预测不如2D预测有效;另一方面,如果想要使3D预测和2D预测一样精细,在3D空间进行预测编码需要使用比2D空间多得多的预测方向。因此,针对3D点云进行类似图像视频的预测并不高效。3D点云属性信息预测通常使用分层预测^[2]。我们把点云属性信息分成不同的层进行逐层编码,并使用已经编码的层对待编码的层进行加权预测。在此种方法的发展过程中,涌现出了多种点云分层方式:基于点与点之间距离的分层方法,以及基于二叉树的分层方法等。

除此之外,我们还提出了基于提升的方式使用编码残差来进一步修正层间预测,以更好地提升性能^[2]。由于分层预测的方法在编码性能和复杂度之间取得了很好的均衡,该方法被基于几何信息的点云编码标准采纳,成为了推荐的属性压缩方法之一。

2.3 基于映射的方法

大部分属性信息编码方法利用任何信息去除属性信息之间的相关性,以提升编码效率;但是基于映射的属性编码方法则有所不同,它采用与基于映射的几何编码方法相同的投影方式,然后使用成熟的视频编码技术对重着色之后的属性视频进行编码。从基本的流程上来说,基于映射的属性编码方法和基于映射的几何编码方法并没有太大不同。在不使用几何信息的情况下,基于高效成熟的2D图像视频编码技术已经能够取得非常好的性能。基于已经编码的几何信息,我们对属性2D图像视频编码器进行运动矢量预测率,并对失真方面进行优化^[12],这使得基于映射的属性视频编码方法取得了进一步的性能提升。基于映射的属性信息压缩方法和基于映射的几何信息压缩方法组成了基于视频的点云编码标准^[2]。

3 点云传统编码方法的比较

3.1 几何信息压缩性能

基于映射的方法和基于表面近似的方法都不适合稀疏点云,所以稀疏点云几何信息几乎只能使用基于树的方法进行压缩。不同于稀疏点云,稠密点云几何信息可以使用上述3种方法进行压缩。表1给出了相对于表面近似的方法,基于映射的方法、基于树结构的方法压缩稠密点云几何信息

▼表1 使用基于映射的方法和基于树结构的方法压缩稠密点云几何信息所得的率失真性能结果

测试序列	基于映射的方法		基于树的方法	
	D1/%	D2/%	D1/%	D2/%
Loot	-80.8	-72.0	117.4	67.9
Redandblack	-77.1	-60.6	103.0	117.9
Soldier	-74.7	-62.8	91.5	61.4
Queen	-87.3	-76.8	94.3	74.9
Longdress	-83.0	-75.0	51.2	54.5

D1: 点到点距离

D2: 点到平面的距离

的率失真性能结果。表1中, $D1$ 表示点到点的距离, $D2$ 表示点到平面的距离; 数字表示相同点云几何信息质量下的码率变化。从表1可以看出, 针对稠密点云, 基于映射的方法会比基于表面近似的方法带来显著的性能提升, 在相同的点到点和点到平面的距离下, 基于映射的方法分别会带来近80%和70%的码率节省。此外, 基于树的方法相比于基于表面近似的方法, 需要额外90%和70%的比特数。综上所述, 基于映射的方法可以带来最好的稠密点云几何信息压缩效果。

3.2 属性信息压缩性能

针对稠密点云属性信息, 基于变换、预测和映射的方法均可以使用。但是基于映射的属性信息压缩方法通常和基于映射的几何信息压缩方法结合起来使用, 而基于变换和预测的方法通常和基于树和表面的几何信息压缩方法结合起来使用。不同的几何信息压缩方法会带来不同的重着色之后的点云, 所以很难单独对基于映射的方法和基于预测和变换的方法进行直接对比。表2给出了使用基于变换的方法和基于预测的方法压缩稠密点云属性信息的率失真性能比较。从表2可以看出, 相比于基于预测的方法, 基于变换的方法对于亮度分量会带来大约3.6%的码率增加, 对于色度分量的性能损失则更加明显。因此, 针对稠密点云, 基于预测的方法可以带来比基于变换的方法更好的属性信息压缩性能。

针对稀疏点云属性信息, 基于映射的方法难以使用, 所以我们主要对基于变换的方法和基于预测的方法进行了对比, 率失真性能如表3所示。从表3可以看出, 针对稀疏点云属性信息, 相比基于预测的方法, 基于变换的方法能带来大约3%的码率节省。

综上所述, 基于变换的方法是业界效果比较好的稀疏点云属性信息压缩方法。

4 点云编码最新进展和发展方向

近几年来, 点云几何和属性信息编码技术的发展取得了长足进步, 但是和传统的图像视频编码标准所能达到的编码效率相比, 仍有较大的距离。如何进一步提升编码性能是点云编码未来发展的目标之一。

帧间预测是传统视频编码中提升压缩效率最显著的部分, 但是对于动态点云而言, 帧间预测效率目前还远远不够。对于稠密动态点云帧间预测, 基于片的映射方法取得了目前最优的性能, 但基于片的映射方法仍存在两个问题: 首先, 点云按片映射到2D视频的过程复杂度很高, 不同于视频编码存在成熟的市场优化方案, 此映射过程目前还不适合实时应用; 另外, 按片映射过程不可避免地破坏了视频的时空相关性。尽管一些人们尝试在视频编码过程中通过寻找3D空间对应块来解决此问题^[12], 但如何从更源头产生时空更连续的视频仍然是稠密动

态点云编码中一个非常关键的问题。对于稀疏动态点云帧间预测, 需要直接在3D空间进行运动估计和运动补偿。但由于相邻点云帧点数不完全相同, 且不同点之间不存在和视频像素一样的——对应关系, 所以3D运动估计和运动补偿是业界一个非常困难的问题, 目前还没有一个成熟的解决方案。

基于深度学习的端到端图像视频编码近期取得了长足的进步, 几乎达到或超越了传统图像视频编码的性能, 这就促进了以端到端的方式对点云进行压缩编码的方法的使用。端到端点云几何属性信息压缩是目前的研究热点^[13]。几何信息编码使用3D普通或稀疏卷积神经网络来编码每个空间位置是否存在3D点这一信息; 属性信息编码使用神经网络结合坐标信息编码对应的颜色和反射率等。目前端到端点云编码仅在稠密静态点云方面取得了较好的效果, 而针对稀疏点云和动态点云, 目前都没有较好的解决方案。另外, 稀疏点云主要针对机器视觉, 易于被端端点云压缩利用, 也是未来非常值得尝试的方向。

▼表2 使用基于变换的方法压缩稠密点云属性信息所得的率失真性能结果

测试序列	Y/%	U/%	V/%
Loot	-4.2	21.3	19.2
Redandblack	7.7	12.2	42.3
Soldier	2.7	21.5	21.3
Queen	5.0	22.1	2.9
Longdress	6.6	22.4	53.4

U、V: 统称为色度分量 Y: 亮度分量

▼表3 使用基于变换的方法压缩稀疏点云属性信息所得的率失真性能结果

测试序列	反射率 /%
Ford_01	-2.5
Ford_02	-1.5
Ford_03	-1.5
Qnxadas-junction-approach	-5.4
Qnxadas-junction-exit	-4.6
Qnxadas-motorway-join	-5.9
Qnxadas-navigating-bends	-1.3

5 结束语

点云几何和属性信息编码是支撑点云广泛应用的关键技术之一。点云几何和属性信息编码近年来取得了长足的进步,但在帧间预测、编码应用等方面仍有许多悬而未决的问题。未来人们需要进一步研究帧间预测、基于深度学习的端到端点云编码等技术,以更高层的应用为目标设计更高效点云几何和属性信息压缩技术。

参考文献

- [1] SCHNABEL R, KLEIN R. Octree-based point-cloud compression [C]//IEEE VGTC conference on Point-Based Graphics. Goslar, Germany: IEEE VGTC, 2006, (6): 111-120
- [2] SCHWARZ S, PREDA M, BARONCINI V, et al. Emerging MPEG standards for point cloud compression [J]. IEEE journal on emerging and selected topics in circuits and systems, 2019, 9(1): 133-148. DOI:10.1109/jet-cas.2018.2885981
- [3] KATHARIYA B, LI L, LI Z, et al. Scalable point cloud geometry coding with binary tree embedded quadtree [C]//2018 IEEE International Conference on Multimedia and Expo (ICME). San Diego, CA, USA: IEEE, 2018: 1-6. DOI:10.1109/icme.2018.8486481
- [4] XU Y Z, ZHU W J, XU Y L, et al. Dynamic point cloud geometry compression via patch-wise polynomial fitting [C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton, United Kingdom: IEEE, 2019: 2287-2291. DOI:10.1109/icassp.2019.8682413
- [5] BUDAGAVI M, FARAMARZI E, HO T, et al. Sam-sungs response to Cfp for point cloud compression (Category 2) [R]. Macau, China, 2017
- [6] HE L Y, ZHU W J, XU Y L. Best-effort projection based attribute compression for 3D point cloud [C]//2017 23rd Asia-Pacific Conference on Communications (APCC). Perth, Australia: IEEE, 2017: 1-6. DOI:10.23919/apcc.2017.8304078
- [7] ZHANG C, FLORENCIO D, LOOP C. Point cloud attribute compression with graph transform [C]//2014 IEEE International Conference on Image Processing (ICIP). Paris, France: IEEE, 2014: 2066-2070. DOI:10.1109/icip.2014.7025414
- [8] DE QUEIROZ R L, CHOU P A. Transform coding for point clouds using a Gaussian process model [J]. IEEE transactions on image processing, 2017, 26(7): 3507-3517. DOI:10.1109/tip.2017.2699922
- [9] DE QUEIROZ R L, CHOU P A. Compression of 3D point clouds using a region-adaptive hierarchical transform [J]. IEEE transactions on image processing, 2016, 25(8): 3947-3956. DOI:10.1109/tip.2016.2575005
- [10] GU S, HOU J H, ZENG H Q, et al. 3D point cloud attribute compression using geometry-guided sparse representation [J]. IEEE transactions on image processing, 2020, 29: 796-808. DOI:10.1109/tip.2019.2936738
- [11] COHEN R A, TIAN D, VETRO A. Point cloud attribute compression using 3-D intra prediction and shape-adaptive transforms [C]//2016 Data Compression Conference (DCC). Snowbird, UT, USA: IEEE, 2016: 141-150. DOI:10.1109/dcc.2016.67
- [12] LI L, LI Z, ZAKHARCHENKO V, et al. Advanced 3D motion prediction for video-based dynamic point cloud compression [J]. IEEE transactions on image processing, 2020, 29: 289-302. DOI:10.1109/tip.2019.2931621
- [13] QUACH M, VALENZISE G, DUFAUX F. Learning convolutional transforms for lossy point cloud geometry compression [C]//2019 IEEE International Conference on Image Processing (ICIP). Taipei, Taiwan, China: IEEE, 2019: 4320-4324. DOI:10.1109/icip.2019.8803413

作者简介



李厚强, 中国科学技术大学教授; 主要研究领域为视频编码与通信、图像处理与计算机视觉、多媒体信息检索等; 主持国家自然科学基金委重点项目、“973”项目、“863”项目等国家级科研项目 10 余项; 获 2019 年国家技术发明二等奖 (排名第 2)、2015 年国家自然科学二等奖 (排名第 2)、2012 年安徽省科学技术一等奖 (排名第 1); 发表论文 200 余篇, 获授权发明专利 60 余项, 被视频编码国际标准采纳提案 45 项。



李礼, 中国科学技术大学特任研究员; 主要研究领域为图像视频编码、3D 点云编码等; 获 2019 年国家技术发明二等奖 (排名第 5); 发表论文 50 余篇, 获授权发明专利 9 项, 被视频编码国际标准采纳提案 8 项。



李竹, 美国密苏里大学堪萨斯分校副教授; 主要研究领域为图像视频编码、图像视频处理、图像视频通信等; 获国际会议 ICIP 2006 最佳论文奖; 发表论文 100 余篇, 获授权美国发明专利 40 余项。

AVS3 视频编码 关键技术及应用

Key Technologies and Applications of AVS3 Video Coding Standard

张嘉琪 /ZHANG Jiaqi¹, 雷萌 /LEI Meng², 马思伟 /MA Siwei^{2,3}

(1. 中国科学院计算技术研究所, 中国 北京 100086;

2. 北京大学, 中国 北京 100871;

3. 北京大学信息技术科创中心, 中国 绍兴 312300)

(1. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100086, China;

2. Peking University, Beijing 100871, China;

3. Information Technology R&D Innovation Center of Peking University, Shaoxing 312300, China)



摘要: 超高清 (UHD) 视频能为用户带来质量更高、沉浸感更强的视觉体验, 但高带宽成本限制了其推广和应用。为解决超高清视频传输和存储的难题, 中国数字音视频编解码技术标准 (AVS) 工作组制定了新一代视频编码标准——AVS3, 并在超高清产业化应用方面取得重要进展。介绍了 AVS3 视频编码关键技术, 以及其与 AVS2、多功能视频编码 (VVC)、开放媒体联盟视频 (AV1) 等标准的性能对比情况。

关键词: 视频编码; AVS3; 超高清

Abstract: Ultra-high definition (UHD) videos can provide users a higher quality and more immersive visual experiences. However, the application of UHD is limited by high bandwidth cost. To solve the transmission and storage problem of UHD, China Audio and Video Coding Standard (AVS) workgroup established a new generation of video coding standard—AVS3. Currently, AVS3 has made a great contribution to the development of UHD industries in China. Key technologies of AVS3 are described, and a comprehensive comparison with AVS2, versatile video coding (VVC) and alliance for open media video 1 (AV1) is conducted.

Keywords: video coding; AVS3; UHD

视觉是人类获取信息的重要来源, 视频承载了海量非结构化视觉信息, 是应用最广泛的多媒体数据格式, 它与人们的生活息息相关, 是人类获取信息的重要途径之一。目前, 互联网 70% 以上的流量来自于图片和视频, 并且这个比例仍在持续攀升^[1], 视频已成为网络上体量最大的数据格式。据统计^[1], 2017 年标清和高清视频内容约各占视频流量的一半; 2019 年标清内容的占比约下降到 1/3, 高清内容

成为主流, 而超高清内容的占比正在逐步攀升; 预计到 2022 年, 超高清内容的占比约提升到 1/4。

超高清视频具有更高的空间和时间分辨率、更广的色域和更宽的动态范围, 是继视频数字化、高清化之后的新一轮重大技术革新。视频技术从高清向超高清的演进, 不仅引发了内容制播、芯片制造、网络传输等产业链各环节的升级换代, 而且驱动了广播电视、安防监控、智能交通等以视频为核心的行业服务转型。自 2018 年起, 中国超高清视频产业已达万亿元级别。预计到 2022 年, 中国超高清视

频产业总体规模将超过 4 万亿元^[2]。

成倍增长的数据量给超高清视频的高效传输和存储带来了巨大的挑战。以 8K、10 bit、120 帧/秒的 YUV (一种颜色编码方法) 420 格式的超高清视频为例, 其原始数据的码率会达到约 88.99 Gbit/s。若采用第 2 代数字音视频编解码技术标准 (AVS2) / 高效视频编码 (HEVC)^[3-4] 标准对原始数据进行压缩, 压缩码率约 310 Mbit/s, 带宽传输压力极大。因此, 超高清视频应用迫切需要更加高效的压缩技术。

为解决超高清视频带宽需求大、存储难等问题, 中国 AVS 工作组率先

基金项目: 国家自然科学基金 (61961130392); 北京大学百度基金 (2019BD003)

展开了具有自主知识产权的、针对超高清视频的视频编码标准的制定工作。在 2017 年 12 月举行的会议中，AVS 工作组决定开展面向超高清视频应用的新一代数字视频编码标准（以下简称 AVS3）的制定工作。AVS3 的制定工作分为两个阶段：第 1 阶段（基准档次）是从 2018 年 3 月到 2019 年 6 月，制定面向复杂度优先的应用，其性能相较于 AVS2 提升 30%；第 2 阶段（增强档次）是从 2019 年 6 月到 2021 年 12 月，目标是编码效率比 AVS2 提升 1 倍以上，同时编码性能超越同时代的其他国际标准。2020 年 5 月 13 日，AVS3 基准档次标准正式获批并被颁布为团体标准。

一直致力于制定高压缩率和友好专利政策的视频编码标准。经历了 19 年的发展, AVS 工作组已经制定从 AVS 到 AVS3 这 3 代视频编码标准。面向超高清视频应用, AVS3 沿用了基于块的预测变换混合编码框架, 具体如图 1 所示。AVS3 包括块划分、帧内预测、帧间预测、变换量化、熵编码、环路滤波等模块。相较于 AVS2, AVS3 在保留部分编码工具的同时, 针对不同模块引入了一些新的编码工具^[5], 并采用了更灵活的块划分结构、更精细的预测模式、更具适应性的变换核, 实现了约 30% 的码率节省, 显著提升了编码效率。

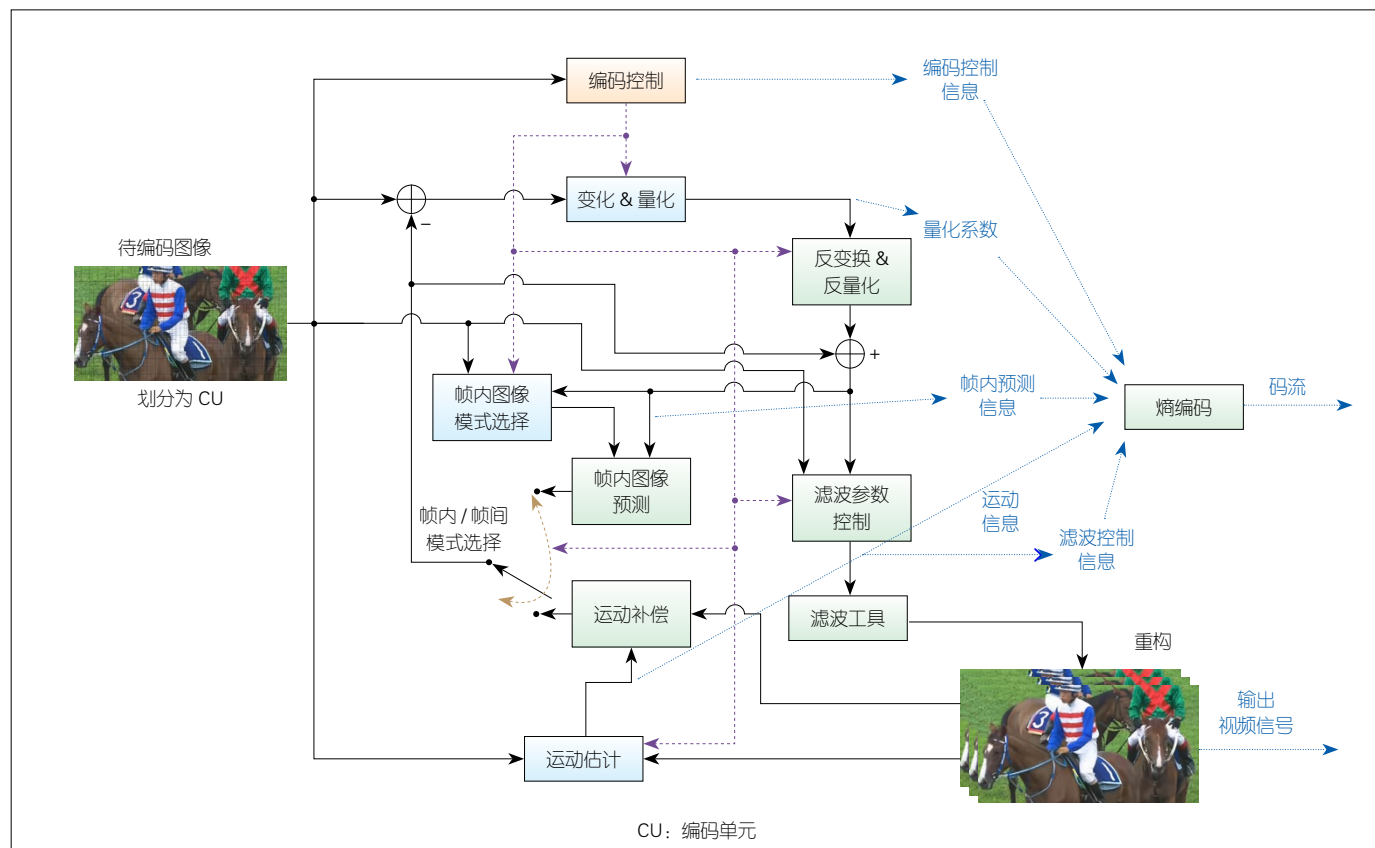
架，每个编码单元（CU）的尺寸都是方形且允许被进一步划分为不同形状的预测单元（PU）。为提升划分的灵活性，AVS3 引入了基于四叉、二叉（QTBT）和扩展四叉树（EQT）的划分方式，如图 2（b）。QTBT 加 EQT 的划分方式允许出现非方形编码单元，编码单元是后续预测、变换和量化的基础，非方形划分更加符合纹理精细和为了便于硬件实现，AVS3 采用了局部分离树（LST）。LST 技术为了避免色度出现边长等于 2 像素的变换块，在亮度块划分时，如果亮度块出现边长等于 4 像素的边，则仅对亮度块划分，无须对色度块划分。为提高硬件流水处理效率，AVS3 对一些小块添加了模式限制。当块大小满足限制后，该节点及其划分得到的编码块的编码模式只能全部选择同一种预测模式，

1 AVS3 视频编码关键技术

AVS 工作组自 2002 年成立以来,

1.1 块划分

如图 2 (a) 所示, AVS2 采用了基于四叉树 (QT) 的递归划分编码框



▲图1 基于块的混合编码框架

如帧间预测或帧内预测。

1.2 帧间预测

帧间预测工具可以分为3类：一类是针对跳过模式和直接模式候选项的扩充，一类是差分运动矢量（MVD）编码，最后一类则是基于子块的运动补偿。

跳过模式和直接模式是一项使用相邻编码块的运动矢量（MV）进行预测的高效编码技术。AVS2中的跳过模式和直接模式候选项只有4个相邻模式和1个时域模式，对图像非相邻结构性和纹理多变性的区域编码效率不高。AVS3引入了基于历史运动矢量的预测（HMVP）和高级运动矢量表达（UMVE）等技术。HMVP利用非局部

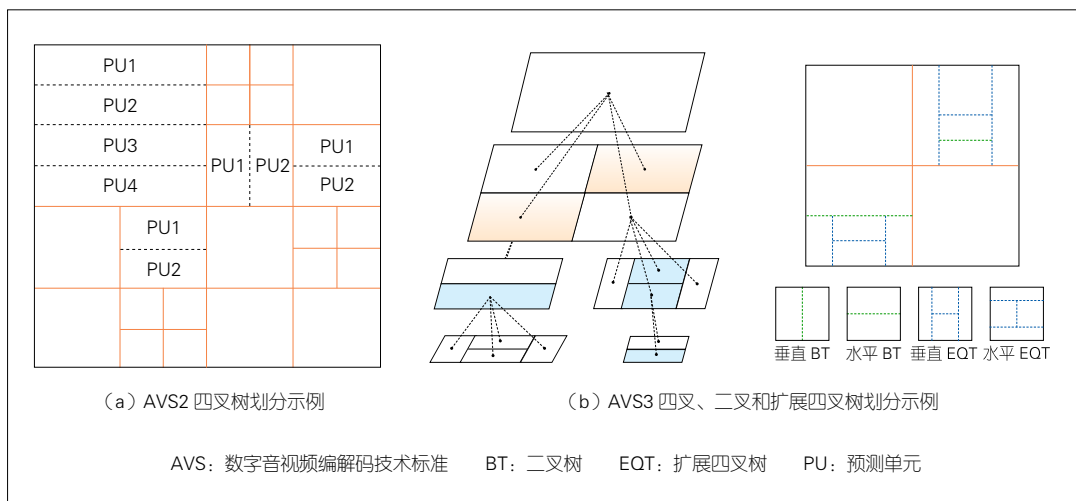
相似性的原理获取更多非相邻的运动矢量候选，如图3（a）所示。HMVP通过动态更新运动候选矢量列表，保留了与当前块运动相关性最高的候选项，提高了跳过模式和直接模式、处理非局部相似性运动的能力。UMVE通过对跳过模式和直接模式候选项加入运动矢量偏移，对运动矢量进行更精细的表达，可以更好地消除视频场景中因剧烈运动而带来的匹配误差。

自适应运动矢量精度（AMVR）和扩展运动矢量精度（EMVR）的引入提升了MVD的编码效率。在AVS2中，运动矢量精度只有1/4像素和1/2像素，且无法灵活选择。AVS3中的AMVR使用了1/4、1/2、1、2、4像素精度的运动矢量，根据视频内容自适应地选

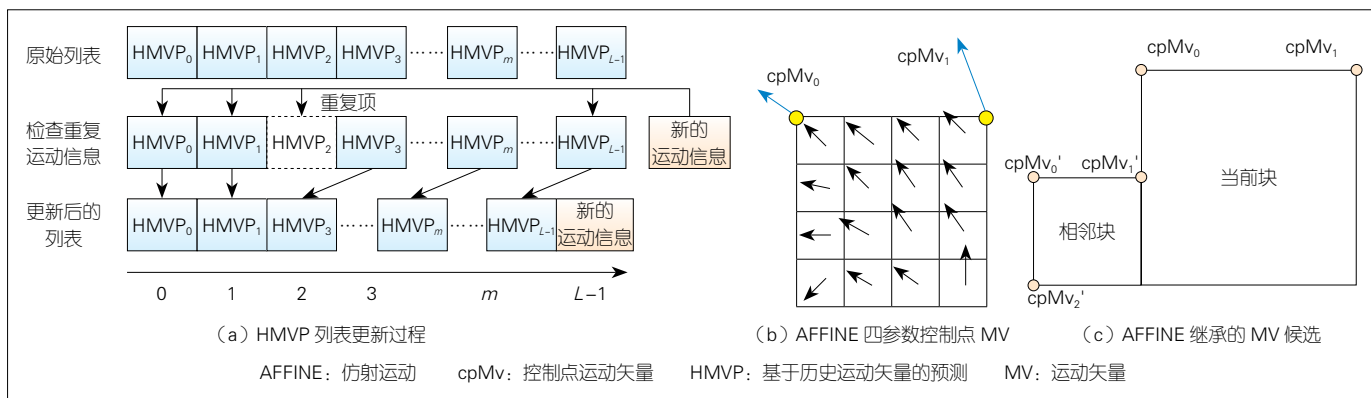
择预测精度，提高了帧间预测在不同区域的适应性。EMVR提供了不同的运动搜索起始点，扩大了运动矢量的搜索空间，有效提升了运动估计的准确性。

双向光流（BIO）^[6]、仿射运动（AFFINE）和解码端运动矢量修正（DMVR）^[7]等技术采用基于子块的运动补偿，提高了帧间预测准确度。基于物体运动轨迹是平滑的这一假设，BIO通过最小化每个子块的前向和后向预测样本之间的差异来计算运动细化差，然后使用运动细化差来调整每个子块的预测样本值。如图3（b）所示，AFFINE根据仿射变换模型，利用2个（四参数）或3个（六参数）控制点的运动矢量导出当前编码块的运动矢

量场。AFFINE运动模型相对于AVS2中的平移运动模型，可以有效提升具有缩放、旋转、透视和其他不规则运动等性能的视频序列编码。DMVR将编码区域划分为若干个不重叠的子块，以初始MV为起始位置，使用最小化均方误差的模板匹配方法对当前MV进行偏移，进一步修正双向



▲图2 AVS2、AVS3 块划分示例



▲图3 HMVP 和 AFFINE 示意图

预测样本值。

1.3 帧内预测

帧内预测方面的新技术包括帧内预测模式扩展(EIPM)、预测像素滤波、跨分量预测等。

EIPM^[8]扩展了帧内预测的角度,如图4(a)所示。帧内预测模式从33种扩展到66种,包括62种角度模式和4种特殊模式,提高了对方向性纹理的预测能力,可以适应纹理丰富的超高清视频内容。

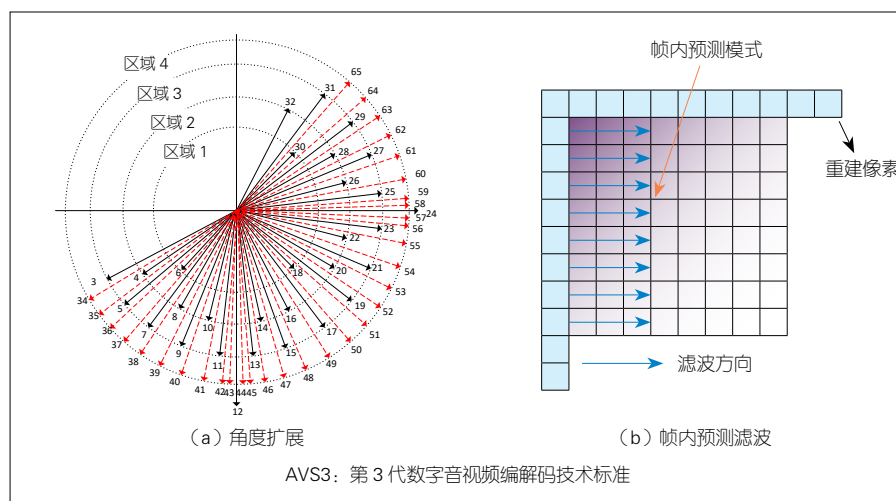
帧内预测滤波包含分像素插值滤波和预测像素值滤波。多组滤波(MIPF)根据块内像素点的个数和所在位置^[9],使用4组不同的插值滤波器生成预测像素。多组滤波适用于不同的颜色分量和像素平滑程度,在复杂度极低的情况下,取得了可观的性能增益。MIPF得到预测像素后,还可以对预测像素进行帧内预测滤波(IPF)。IPF使用高斯平滑滤波器,根据参考像素、预测模式和与参考像素的距离对预测像素做进一步的修正,如图4(b)所示。跨分量预测是指在色度预测编码过程中,通过两步预测模式(TSCPM)对色度进行预测编码。其原理是假定亮

度和色度分量之间线性相关,通过最小二乘法求解对应线性回归的参数,在求得参数后,使用亮度重构像素以精细重建对应位置的色度像素,在色度上取得了显著的增益。

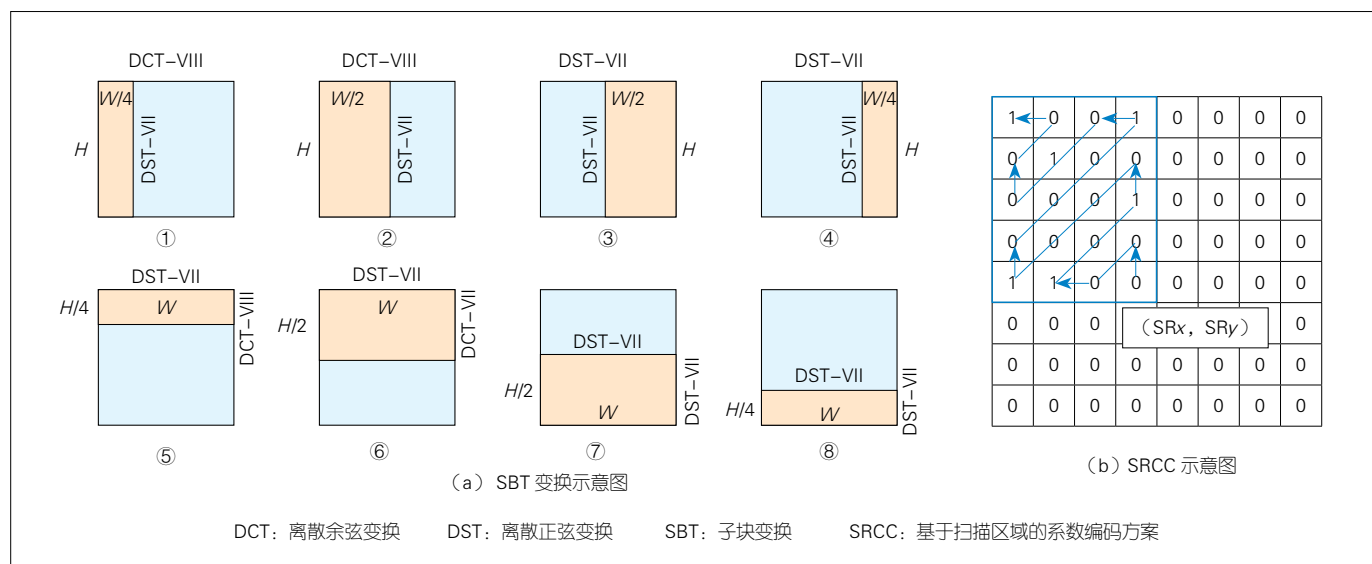
1.4 变换与量化

变换可以集中能量,利于熵编码进行系数压缩。离散余弦变换(DCT)具有很好的去相关能力^[10],且由于其对称性有利于软硬件实现,因此能够在视频压缩领域得到广泛的应用。在上二代视频编码标准中,DCT-II作为

主要应用的变换核,适用于均匀分布的残差变换,但缺乏处理不均匀残差分布的能力。在AVS3中,隐则变换(IST)和子块变换(SBT)引入了新的变换核DST-VII和DCT-VIII,能够聚集不均匀分布残差的能量。IST^[11]通过量化块中偶数系数个数的奇偶性隐式地导出变换核的类型,在提高变换灵活性的同时,没有引入额外的比特消耗。基于帧间预测残差分布的局部性,SBT把预测残差分布的位置限制在残差块的1/2或者1/4区域,如图5(a)所示,从而降低变换系数的局部分量,



▲图4 AVS3 帧内编码工具



▲图5 变换与系数编码

并减少了全零块的编码代价,提高了压缩性能。

在系数编码中,AVS3采用了一种基于扫描区域的系数编码方案(SRCC)^[12]。SRCC使用参数(SR_x , SR_y)控制量化系数非零的区域。为了达到码率和失真之间的平衡以及提高系数编码的灵活性,SRCC使用率失真优化选择最优扫描区域。在扫描编码区域内的非零系数时,SRCC采取了从右下到左上的反Z形扫描方式,如图5(b)所示;非零系数采用了分层编码,不同层级使用多套上下文,根据系数在扫描区域的位置和扫描区域的面积确定上下文模型。精确的上下文建模显著提升了压缩效率。

1.5 基于卷积神经网络的环路滤波

为了探索神经网络在编码标准中的可实现性,AVS3工作组设立了智能编码专题小组,对基于卷积神经网络的环路滤波(CNNLF)^[13]进行了深入研究。CNNLF能够代替传统的去块(Deblock)滤波和样本自适应偏移(SAO)滤波,并取得了6%左右的性能增益。

CNNLF使用神经网络探索视频信号之间的非线性关系和变化规律,对视频信号的全局信息和局部关系进行了联合建模。得益于海量的训练数据和算力的提升,CNNLF的网络泛化能力要远高于传统滤波方式。CNNLF训练时以残差块为单位,加速了网络收敛过程,并且设置不同量化参数(QP)段为训练单元,

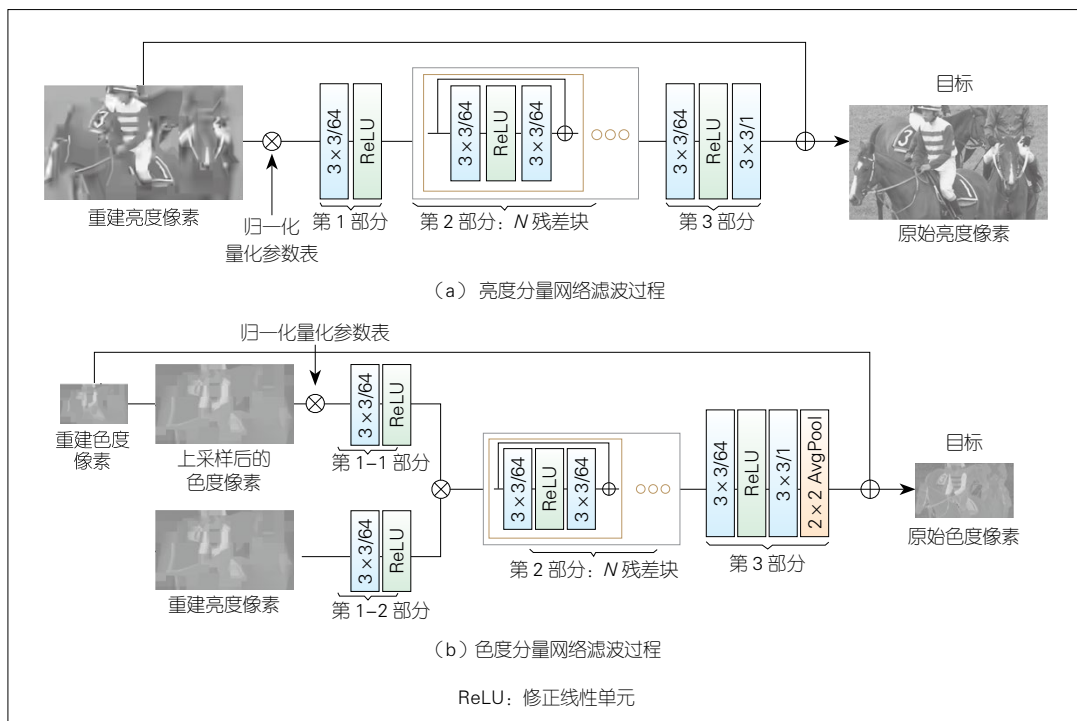
增强了网络对QP的泛化能力。如图6(a)所示,CNNLF的网络由全局残差、残差块、卷积层和激活层组成,采用亮度、色度分量分离训练的方式,且亮度分量指导色度分量滤波,进一步提升色度分量的重建质量,如图6(b)所示。

1.6 性能对比

本文中,我们首先对AVS3与AVS2进行了性能对比,测试时使用的参考软件版本分别为参考设计模型19.5(RD 19.5)、高性能平台4.0(HPM 4.0)和高性能平台9.0(HPM 9.0),其中HPM 4.0和HPM 9.0分别用于测试AVS3第1阶段和第2阶段。测试配置为随机访问(RA)配置,测试结果见表1。可以看出AVS3第1阶段相比AVS2平均可以获得24%的性能提升,且对4K分辨率序列的提升更为明显,达到了平均25%的性能提升。AVS3第2阶段是基于第1阶段的进一步推进。相比第1阶段,AVS3第

2阶段实现了约8%的性能提升;相比AVS2,实现了平均31%的性能提升,同时各分辨率序列的性能提升较为均衡,在部分4K序列上可以达到超40%的性能提升。此外,我们还测试了AVS3采用了CNNLF后的性能,如表1最右侧所示可以再获得近3%的性能提升。

我们还将AVS3和VVC^[14]、开放媒体联盟视频标准(AV1)^[15]进行了对比。我们选取了5个2K序列及6个4K序列进行测试,测试平台分别为通用编码测试平台(VTM 10.0)、开放媒体联盟视频标准测试平台(AOM-Oct)、HPM 4.0和HPM 9.0。如表2所示,VVC、AVS3和AV1相较于HEVC,在客观性能上都有较大的提升,尤其是VVC的性能提升最为显著,平均达到了40%;其他各标准中,AV1平均提升了25.5%的性能,对于AVS3,第1阶段和第2阶段分别达到了平均23%和30%的性能提升。综合来看,VVC、AV1和AVS3在超高清序列方



▲图6 基于卷积神经网络的环路滤波过程

面都表现出了优异的性能,达到了平均 25% 及以上的性能提升。VVC 和 AVS3 更是达到了超 30% 的性能提升,个别序列能达到 40% 的性能提升。

2 AVS3 超高清产业应用

随着超高清、全景视频等应用的高速发展,8K 超高清,乃至 16K、32K 等更高分辨率的视频内容将进一步流行。2019 年,中国发布的《超高清视频产业行动计划(2019—2022)》明确指出超高清视频将成为未来视频产业的重要发展方向。

AVS3 标准的颁布显著加速了超高清产业链的升级革新。为了缩短标准制定和成果落地的时间,AVS3 工作组在标准制定过程中,采用了分档制定与芯片集成技术协同研发的推进方式,同步推进全产业链应用开源合作。2019 年 6 月,AVS3 第 1 阶段基准档完成;2019 年 9 月,在阿姆斯特丹举办的第五十届荷兰广播电视设备展览会上,海思发布了首个基于 AVS3 标准的 8K 端到端解决方案,同时推出了全球首颗基于 AVS3 标准的支持 8K 分辨率、120 帧的超高清解码芯片 Hi3796CV300,如图 7(a)所示;随后,北京大学、北京博雅睿视科技有限公司和英特尔合作推出了 SVT-AVS3 8K 实时编码器,并搭建了 8K 端到端实时编解码系统,如图 7(b)所示。北京大学深圳研究生院开发了支持 AVS3 标准,8K 分辨率、60 帧实时解码器 uAVS3d。2020 年 5 月,当虹科技 AVS3 8K 超高清编码器和上海海思 AVS3 8K 超高清解码板完成了 AVS3+5G+8K 全国直播首测,主要测试在 5G 链路下的 8K 超高清节目直播传输应用。近期,中央广播电视总台启动“5G+4K/8K 超高清制播示范平台”项目,其中包括搭建 AVS2/AVS3 标准超高清电影院直播系统以及 5G 和超

▼表 1 AVS3 与 AVS2 在通用测试条件下的性能对比

序列	HPM 4.0			HPM 9.0			HPM 9.0-Mod AI 4.0		
	Y/%	U/%	V/%	Y/%	U/%	V/%	Y/%	U/%	V/%
通甲 4K	-23.9	-27.5	-29.6	-32.5	-36.8	-37.8	-36.0	-35.3	-38.8
通乙 1 080P	-23.7	-28.0	-29.6	-32.6	-35.5	-36.4	-35.3	-37.1	-34.9
通丙 720P	-23.0	-27.2	-26.5	-28.4	-30.2	-28.1	-32.7	-38.9	-40.5

AVS: 数字音视频编解码技术标准 HPM: 高性能平台 U、V: 统称为色度分量 Y: 亮度分量

注: 测试配置为随机访问, 均与设计模型 RD19.5 对比。

▼表 2 VVC、AV1、AVS3 和 HEVC 的性能对比

测试软件	随机访问编码配置		
	Y/%	U/%	V/%
VTM10.0	-38.17	-39.81	-39.71
AOM-Oct	-23.09	-33.69	-31.84
HPM 4.0	-23.20	-19.19	-18.66
HPM 9.0	-31.16	-27.87	-26.73

AOM-Oct: 开放媒体联盟视频标准测试平台

AV1: 开放媒体联盟视频标准

AVS3: 第 3 代数字音视频编解码技术标准

HEVC: 高效视频编码

HPM: 高性能平台

U、V: 统称为色度分量

VTM: 通用编码测试平台

VVC: 多功能视频编码

Y: 亮度分量



▲图 7 基于 AVS3 的 8K 应用实例

高清相关的测试体系。中央广播电视总台会将 AVS3 8K 超高清现场直播运用在 2022 年北京冬季奥运会中。

3 结束语

本文简要介绍了新一代视频编码标准 AVS3 的关键技术和 AVS3 超高清应用情况。与 AVS2 视频编码标准相比, AVS3 编码效率显著提升。AVS3 标准在技术创新、专利政策与生态建设方面已有全面的布局,为中国 8K 超高清

视频产业的发展奠定了坚实的基础。可以预见的是,随着 5G 的快速发展和超高清时代的来临, AVS3 标准前景广阔,将获得更广泛的应用。

参考文献

- [1] FORECAST G. Cisco visual networking index: global mobile data traffic forecast update, 2017–2022 [R]. 2017
- [2] 工信部等三部门联合印发《超高清视频产业发展行动计划(2019–2022 年)》[EB/OL]. [2020–12–22]. http://www.gov.cn/gongbao/content/2019/content_5419224.htm

- [3] MA S W, HUANG T J, READER C, et al. AVS2 ? making video coding smarter [standards in a nutshell] [J]. IEEE signal processing magazine, 2015, 32(2): 172–183. DOI:10.1109/msp.2014.2371951
- [4] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard [J]. IEEE transactions on circuits and systems for video technology, 2012, 22(12): 1649–1668. DOI:10.1109/tcsvt.2012.2221191
- [5] ZHANG J Q, JIA C M, LEI M, et al. Recent development of AVS video coding standard: AVS3 [C]//2019 Picture Coding Symposium (PCS). Ningbo, China: IEEE, 2019: 1–5. DOI:10.1109/pcs48520.2019.8954503
- [6] 王凡, 欧阳晓, 吕卓逸, 等. CE: BIO 双向光流 [C]// 数字视频编解码技术标准化工作组第六十九次会议. 成都, 中国: AVS 工作组, 2019
- [7] 徐巍伟, 赵寅, 杨海涛. CE3.1: 简化 DMVR 方案 [C]// 数字视频编解码技术标准化工作组第七十次会议. 海口, 中国: AVS 工作组, 2019
- [8] 雷萌, 罗法蕾, 王苕社, 等. CE2-related: 帧内角度模式扩展 [C]// 数字视频编解码技术标准化工作组第七十次会议. 海口, 中国: AVS 工作组, 2019
- [9] 王英彬, 许晓中, 李一鸣, 等. CE1-related: 帧内预测参考像素滤波设计方法 [C]// 数字视频编解码技术标准化工作组第七十次会议. 海口, 中国: AVS 工作组, 2019
- [10] NUSSBAUMER H J. The fast Fourier transform [M]//Fast fourier transform and convolution algorithms. Berlin, Heidelberg: Springer Berlin Heidelberg, 1981: 80–111
- [11] 张玉槐, 张凯, 张莉, 等. 帧内自适应变换 [C]// 数字视频编解码技术标准化工作组第六十九次会议. 成都, 中国: AVS 工作组, 2019
- [12] 王凡, 欧阳晓, 吕卓逸, 等. SRCC 基于扫描区域的系数编码 [C]// 数字视频编解码技术标准化工作组第七十次会议. 海口, 中国: AVS 工作组, 2019
- [13] 林凯, 贾川民, 赵政辉, 等. CE: 基于残差网络的神经网络滤波 [C]// 数字视频编解码技术标准化工作组视频组 2020 年 1 月加会. 北京, 中国: AVS 工作组, 2020
- [14] BROSS B, CHEN J, LIU S, et al. Versatile video coding (Draft 10): ITU-T and ISO/IEC JVET-S2001 [S]. 2020
- [15] CHEN Y, MURHERJEE D, HAN J N, et al. An overview of core coding tools in the AV1 video codec [C]//2018 Picture Coding Symposium (PCS). San Francisco, CA, USA: IEEE, 2018: 41–45. DOI:10.1109/pcs.2018.8456249

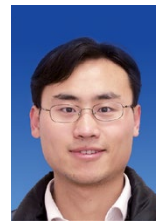
作者简介



张嘉琪, 中国科学院计算技术研究所读博士研究生; 主要研究方向为视频编码及处理; 所提多项标准提案已被 AVS3、VVC 等标准采纳。



雷萌, 北京大学信息科学技术学院在读博士研究生; 主要研究方向为视频编码及处理; 发表论文 2 篇。



马思伟, 北京大学信息科学技术学院教授、博士生导师, 国家杰出青年科学基金获得者, 现担任 AVS 视频组组长; 主要研究方向为视频编码及处理; 曾主持“863”计划、科技支撑计划、国家自然科学基金重点项目等多项国家级课题; 曾获国家技术发明奖二等奖、国家科学技术进步奖二等奖、中国电子学会特等奖等奖励; 已发表 SCI 论文 70 余篇, 已获得授权发明专利 50 余项。



下一代通信 助力实时分布云渲染

Next-Generation Communications Technology Facilitates Real-Time Distributed Cloud Rendering

陆平 /LU Ping¹, 盛斌 /SHENG Bin², 朱方 /ZHU Fang¹

(1. 中兴通讯股份有限公司, 中国 深圳 518057;
2. 上海交通大学, 中国 上海 200240)

(1. ZTE Corporation, Shenzhen 518057, China;
2. Shanghai Jiao Tong University, Shanghai 200240, China)

摘要: 实时分布式云渲染技术和 5G 通信技术的应用, 可以解决工业模型中实时渲染的问题。实时分布式云渲染技术将渲染工作放在服务器端, 再将渲染结果利用 5G 通信技术快速传输到客户端以实现实时的浏览和交互。介绍了实时分布式云渲染技术的基本框架、关键技术、发展状况, 以及当前的一些应用, 总结了其面临的技术难题。认为新基建中人工智能的建设, 可以进一步提升实时分布式云渲染的性能。

关键词: 工业数字化; 实时分布式云渲染; 服务器端渲染; 5G 传输

Abstract: Combining real-time distributed cloud rendering technology with 5G technologies can solve the problem of real-time rendering in industrial models. The rendering work is placed on the server side by distributed cloud rendering technology, and then the rendering results are transmitted to the client side quickly by 5G technologies to realize real-time browsing and interaction. Basic framework, key technologies, development status and some current applications of real-time distributed cloud rendering technology are introduced, and technical problems still faced by real-time distributed cloud rendering technology are summarized. It is believed that the construction of artificial intelligence in the new infrastructure can further improve the performance of real-time distributed cloud rendering.

Keywords: industrial digitization; distributed cloud rendering; server rendering; 5G transmission

DOI: 10.12142/ZTETJ.202101005

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210122.1756.004.html>

网络出版日期: 2021-01-25

收稿日期: 2020-12-22

在工业 4.0 时代, 为了更加方便地进行工业设计和生产工作, 人们已经逐步开始将大规模的工业模型制作成数字化模型后再渲染出来, 供技术人员和设计人员使用。渲染是将三维的场景模型, 通过添加材质、灯光、特殊效果等方式, 最终将其转化成可以在显示器上呈现的二维图像的过程^[1]。渲染分为实时渲染和离线渲染: 实时渲染是指计算机在完成每帧的渲染工作之后, 就直接将渲染结果输出显示; 离线渲染是指当完成全部

的渲染工作后, 再将渲染结果逐帧输出。对于工业设计和生产来说, 最理想的方式是多人交互的实时渲染。进行三维场景的实时渲染需要消耗大量的计算资源, 还需要考虑模型的复杂度、纹理贴图的精细度、光线计算的复杂度等。为了达到高保真度的写实渲染效果, 需要配备大型和专业的图形显卡等硬件设备, 这是一笔不小的硬件费用支出。同时这些工业模型往往占用很大存储空间并且具有一定的保密性, 无法在单机渲染和存储。为

了解决这些问题, 研究人员利用云计算技术和 5G 传输技术, 发明了实时分布式云渲染技术。

实时分布式云渲染是指在云服务器中进行资源的分布式存储、动态分配和模型场景的实时渲染, 它保证了渲染资源的共享性和安全性。在服务器的总控制节点上, 根据各个渲染分节点的渲染性能进行渲染资源分配, 之后利用多台服务器的计算资源进行渲染任务的渲染, 最后各个分节点将渲染结果推流到总节点并进行合并,

再传递到本地客户端。利用实时分布式云渲染技术可以完成单机无法完成的大规模复杂场景的渲染工作，渲染出让人无法分辨真假的图像，同时大大降低客户端处硬件设备的需求。网络传输的速度和带宽是分布式云渲染实时性的瓶颈。借助 5G 传输的高速度、大带宽，可以突破这些瓶颈。多个用户可以使用不同的硬件设备并利用 5G 网络进行场景接入。多用户在不同的视角下，同时对同一个大规模的复杂场景模型进行实时访问浏览和交互式操作的功能，完全可以满足工业设计和生产工作的需求。

1 实时分布式云渲染的技术特点

(1) 分布式渲染技术

在早期对并行渲染技术进行研究时，人们将这种形式描述为一种“分类整理问题”（Sorting Problem），并根据“分类整理”（Sort）操作在渲染管线中的发生位置分为 Sort-first、Sort-middle、Sort-last 3 种主要类型。其中，发展与应用得最好的方式是 Sort-middle^[2]，它是指“整理”行为发生在几何阶段与光栅化阶段之间，此时场景实体已经被转换到显示坐标系中。由于几何阶段和光栅化阶段在实现原理上是分配给不同处理器运行的，因此 Sort-middle 是一种很自然的处理手段，而它也是集成电路独立显卡的理论基础。

为了追求更逼真的视觉效果，渲染管线流程复杂程度显著提升，集成显卡的电路复杂程度也在随之提升。此外，神经网络渲染等新兴技术的出现也对原有的技术框架不断提出挑战。人们之前就考虑过多渲染机并行的分布式解决方案，不过限于当时分机之间通信方式的吞吐量、时延、可靠性等，以并行集成电路为主要承载形式的图形加速卡硬件占据了主流。而 5G 网络

的出现为这一领域开辟了新的可能，它能在保持超低时延的同时保证超大带宽，因而弥补了多机分布式方案的短板。这种组织结构有望更加充分地发掘现有硬件设备的潜力，特别是神经网络这一对存储空间与算力要求非常高而对数据全局性依赖较低的计算过程。

(2) 可交互云渲染技术

可交互云渲染技术的前身——“远程实时协助”于 1968 年被实现，最初这项技术的主要内容是通过网络连接实现多台终端设备共享操作界面以及交叉控制。21 世纪 10 年代以来，在智能手机、移动网络和云技术逐渐发展和普及过程中，直播行业兴起，人们开始关注小型设备通过网络与云端服务器实时交互与传输媒体流的可能性，由此引申出了“云渲染”与“云游戏”等一系列概念。与传统“远程桌面”技术不同的是，可交互云渲染技术对各终端之间的网络性能要求非常高，实时交互要求整个系统中每一部件之间的信息交换均具有极低的通信时延，而高分辨率图像高速传输同时要求网络具有超大吞吐量，且能快速处理拥堵与异常。除此之外，云渲染技术包括渲染环境部署、图像渲染、图像编码、网络传输、图像解码、显示等一系列细分技术^[3]。

2 实时分布式云渲染的发展

早在 1994 年，学术界就对分布式渲染（又称并行渲染）有了较为早期的理论定义，而世界上第一款真正意义上的大众消费级 3D 渲染加速卡——Voodoo 于 1995 年才正式问世。在此之前，人们就逐渐意识到，面向通用型计算的传统结构中央处理器（CPU）的图像绘制运算效率并不高，CPU 不同部件的工作负载十分不均衡，这限制了运算效率。为了满足图像绘制对

计算能力日益增加的需求，人们开始研究并行计算的加速。此后，行业的研究方向主要集中于针对大规模场景的并行绘制和基于众核硬件架构的并行光照计算。其中，多独立节点式的并行绘制由于对节点之间的通信网络反应时间、吞吐量等指标要求很高，在很长的一段时间内网络通信技术无法满足该需求，这导致该技术一直进展缓慢。近年来，随着 5G 技术的逐渐成熟与普及，独立渲染节点之间的超高性能交换网络的组建成为可能，这让该技术重新回到人们的视野之中。

多节点并行绘制主要的技术痛点是大规模场景中全局光照计算时的内存瓶颈问题，其中一种实现方式是内外存交换调度，另一种是将数据分布式存储到不同节点内存。目前大多研究基于同构计算资源，通过对空间数据进行负载均衡的合理分割，实现高利用率的计算单元并行。目前的研究主要有基于图像空间分解策略和基于场景空间分解策略两个方向。

实时分布式云渲染的其中一个创新点是云技术。3G/4G 网络与直播的普及向人们证明了使用性能比较受限的小型设备，诸如手机平板等，是可以实现移动、流畅地播放高质量流媒体的。早在这一时期就已经有商家开始布局云游戏领域，不过受制于移动网络稳定性、网速、延迟等，云渲染的发展也长期处于停滞状态。5G 技术的出现正好弥补了上一代移动网络的短板，让用户通过小型便携设备实时与云端游戏客户端进行交互，并让流媒体传输成为可能。云技术有望在 5G 通信的加持下解放终端设备，任一具有网络功能的显示设备均能进行高性能图像软件的操作使用。

实时光线追踪在计算机图像领域异军突起，带来了前所未有的真实绘制效果，引发一场新的绘制技术革命。

而这种手段同时要求巨大计算量与时间复杂度,以完成对复杂光路的高数据密度计算,因此必须考虑用计算并行化方式提高速度。由于光线追踪绘制具有全局化等特点,多节点式存储的实现途径将依赖不同节点的大量信息交换,这增添了新的运算负载,故在目前情况下光线追踪领域分布式渲染并行方式具有局限性。

3 实时分布式云渲染的应用

当前,实时分布式云渲染以其高质量、高速度、对客户端无压力的特点成为了渲染技术研究的热点。该技术目前的研究关键在于如何实现分布式节点在工作任务分配过程中的负载均衡,以及如何解决网络通信开销巨大、传输速度快等问题^[3],这使得该项技术还没有被大规模地产业化应用。但这项技术一旦成熟就可以带来技术红利^[4],因此值得我们去大力探索。

具体来看,实时分布式云渲染技术在实践中将展现出以下优势:

(1) 实时分布式云渲染技术极大地改善了用户体验。首先,云渲染技术的存在,使得用户体验不再囿于本地主机配置,用户可以享受一些高画质、高分辨率的动画电影。其次,对于实时性强的应用,例如3D游戏,实时分布式云渲染能够快速渲染出高帧率、高分辨率的画面并进行传输。同时,实时分布式云渲染技术加强了对于闲置计算资源的利用,降低了资源的消耗。分布式云渲染概念是由分布式并行集群渲染技术衍生而来的,它将原先由本地客户端处理的图形渲染(2D、3D)转移至云端,以使得网络中闲置的计算资源得到利用,从而降低本地渲染时的硬件消耗成本。云渲染平台的计算资源可以即使用即申请,不用时更不需要浪费本地各类资源进行维护^[5]。

(2) 5G技术的逐渐成熟使得云计算、云渲染技术在由云端向客户端传输数据的瓶颈不复存在。这进一步弥补了实时分布式云渲染技术的应用过程中薄弱的一环^[6]。同时,相比于光纤传输和WiFi传输的方式,5G传输技术的移动性强,部署便捷性高。光纤传输方式需要铺设大量的光纤线路,光纤入户需要的成本高。而5G基站的建设成本低,单个基站就可以覆盖周围的区域。相比于WiFi传输方式如WiFi6,其主要用于室内无线终端上网,并不适合高速移动通信。而5G通信技术,既可以用于广域高速移动通信,又可以用于室内无线上网。

对于实时分布式云渲染技术的应用,鉴于这一技术高效高质量渲染的特点,其潜在的服务对象为影视动画、游戏、虚拟现实(VR)等行业的建模、灯光、渲染等领域。

影视动画的渲染在集群式的渲染方式下即可完成;但在3D游戏、VR游戏等行业,因为实时性以及对渲染帧速率和分辨率的需求,实时分布式云渲染将大有可为。根据统计,近年来全球游戏市场规模逐渐扩大,2019年已达到1457亿美元,2020年预计将达到1593亿美元,据此可以推测云游戏将是基于实时分布式云渲染架构的一个新的极具发展前景的应用^[4]。云游戏服务这一概念首先由OnLive提出并证明了可行性,在云游戏的运行模式下,所有游戏都在服务器端运行,并将渲染完毕后的游戏画面压缩后通过网络传送给用户。在客户端,用户的游戏设备不需要任何高端处理器和显卡,只需要基本的视频解压能力即可流畅运行游戏。之后,HUANG C. Y.等又提出了第一个开放式云游戏系统GamingAnywhere,其具有高扩展性、可移植性(Windows、Linux、OS X)和可重新配置性,相较于OnLive和

StreamMyGame两个系统具有更低的反应延迟、更高的画面质量^[7]。

另外,实时可视化渲染同样是该技术的一个重要应用场景。可视化内容包含数据可视化、过程模拟可视化、虚实内容的融合等。渲染技术作为生成可视化内容的重要手段和流程,在当前新基建的提出和数据运用多样化的大背景下,将会是非常重要的环节。在工业4.0时代,“智能”二字将更加深入人心,数据和信息将会更多地用于实际的科学研究和工业生产中。如果能在实际的生产中实现可视化,让研究和生产过程更加直观清晰,将对整个社会的发展有着巨大的促进作用。若高质量高速度的渲染、数据可视化、音响技术、传感器技术等能进一步融合成为成熟的基于高分辨率立体投影的虚拟现实显示技术,将可以用于任何具有沉浸感需求的虚拟仿真应用领域,如虚拟设计与制造、虚拟装配、模拟训练、虚拟演示演示、虚拟生物医学工程、地质、矿产、石油、航空航天、科学可视化、军事模拟地形地貌、地理信息系统(GIS)等。

随着5G和人工智能(AI)时代的到来,实时分布式云渲染在未来必将发挥更大的优势,视频、游戏等行业也将会发生巨大改变;但未来还有很长的路需要探索。

4 问题与展望

实时的分布式云渲染技术已经得到一定的发展,达到了较好的效果,但仍面临着一些技术难题:

(1) 渲染任务的分配问题。使用实时分布式云渲染技术进行渲染的一般是大规模的复杂场景模型,因此需要通过总控制节点将渲染任务分配到不同的渲染子节点上。在该过程中需要创建一个高效的大规模复杂模型的管理框架,并考虑资源划分的速度、

各个渲染节点上的分配的工作负载均衡等问题。这是提高分布式渲染性能的必经之路。

(2) 渲染结果的合成问题。各个节点渲染得到的效果需要汇总在总节点上,并合成为同一张图像,再传递至客户端处。如何将不同渲染节点上得到的渲染结果正确、快速地合成为一帧的图像,也是一个影响实时分布式云渲染性能的重要问题。

(3) 渲染结果的压缩问题。需要将渲染得到的让人无法分辨真假的图像,高效无损地进行压缩,最大程度满足人们的视觉效果要求。

(4) 网络传输的带宽和速度问题。为了将高质量的渲染结果在不同节点之间、客户端与服务器端传输,需要较大的传输带宽。为了达到实时渲染和实时交互的目的,需要较快的传输速度。

(5) 安全问题。在计算机领域,安全性是指保证存储在计算机上的数据不被没有权限的人盗取和访问,而云渲染的过程需要将数据从本地上传至网络^[8],这中间就存在着数据泄露的可能,从而威胁信息安全。人们因为信息安全的隐患而产生对技术本身的信任问题。一旦信任问题被解决,云渲染技术的发展速度就会得到极大提高。

新基建的提出将会进一步促进实时分布式云渲染的发展。新基建是指新型基础设施建设,它通过吸收新科技革命成果,实现国家生态化、数字化、高速化、新旧动能转换与经济结构对称态。2020年3月,中共中央政治局常务委员会提出要加快5G网络、数据中心等新基建的建设进度。

5G网络因其大带宽和高传输速度的特点,为实时分布式云渲染技术提供了硬件上的支持。随着5G网络的进一步普及,利用5G进行实时分布式云

渲染各个节点之间、服务器与客户端之间的通信,将会大大提升实时分布式云渲染的性能。可以想象在不远的将来,用户可以直接使用手机、平板电脑等便携式设备,通过连入5G网络,直接进行大规模工业模型的多人交互设计工作。

新基建中人工智能的建设,也可以进一步提升实时分布式云渲染的性能。例如,可以利用深度学习技术,先进行用户行动的智能预测,再利用预测得到的用户接下来的动作提前进行下一步的渲染工作。这样一来可以降低渲染延迟,也可以利用神经网络进行注视点预测,以减少渲染计算量,提升计算资源的利用率。神经渲染已经成为计算机图形学领域发展最为迅猛的发展方向之一,利用神经网络可以达到独特的渲染效果,如场景合成、视角变化等,这些都可以为实时分布式云渲染技术助力。这些人工智能方法的加入可以优化实时分布式云渲染技术的时延、渲染效果、渲染流程等。随着新基建的发展,实时分布式云渲染将会迎来更大的发展空间和更好的发展前景。

致谢

感谢上海交通大学计算机系秦义明同学、华东理工大学计算机系黎宇航同学的调研工作,以及北京大学马雷老师的指导工作。

参考文献

- [1] 渲染 [EB/OL]. [2020-12-18]. <https://baike.baidu.com/item/渲染/464729>
- [2] MOLNAR S, COX M, ELLSWORTH D, et al. A Sorting classification of parallel rendering [J]. IEEE computer and graphics and applications, 1994, 14(4): 23-32. DOI: <https://doi.org/10.1109/38.291528>
- [3] LIU Z, ZOU H. AzureRender: a cloud-based parallel and distributed rendering system [C]// 2015 IEEE 17th International Conference on

High Performance Computing and Communications. USA: IEEE, 2015: 1881-1886. DOI: 10.1109/HPCC-CSS-ICSS.2015.328

- [4] 徐婵娟. 基于服务器端的三维渲染技术综述 [J]. 中国传媒大学学报 (自然科学版), 2019, 26(1): 20-26
- [5] HONG H J, CHUANG J C, HUS C H. Animation rendering on multimedia fog computing platforms [C]//IEEE International Conference on Cloud Computing Technology & Science. USA: IEEE, 2016
- [6] 郑海林, 朱峰. 基于5G网络的移动云计算优化措施研究 [J]. 信息与电脑 (理论版), 2019, (9): 156-157+160
- [7] HUANG C Y, HUS C H, CHANG Y C, et al. GamingAnywhere: an open cloud gaming system [C]//ACM Multimedia Systems Conference. USA: ACM, 2013
- [8] ZHOU F. Analysis of computer network security issues in cloud computing environment [J]. Lifelong education, 2020, 9(6): 57-59

作者简介



陆平, 中兴通讯股份有限公司副总裁、移动网络和移动通讯多媒体技术国家重点实验室副主任; 研究方向包括云计算、大数据、增强现实、基于多媒体服务的技术; 主持和参与了国家科技重大专项、国家科技支撑项目等; 发表多篇文章, 撰写了《物联网能力开发与应用》《云计算中的大数据技术与应用》等多部著作。



盛斌, 上海交通大学计算机科学与工程系副教授; 研究方向包括虚拟现实和计算机图形学。主持国家自然科学基金面上项目2项、国家自然科学基金青年项目1项, 参与“863”计划1项、国家自然科学基金重点项目1项; 发表论文121篇。



朱方, 中兴通讯股份有限公司数字视频与视觉技术委员会主任、移动网络和移动通讯多媒体技术国家重点实验室多媒体方向学术带头人, IEEE高级会员; 研究方向包括云架构和基于移动计算的XR&Smart Vision特定应用目的的加速芯片组等; 发表文章10余篇, 已授权发明专利3项。

视频质量增强模型加速算法

Video Quality Enhancement Model Acceleration Algorithm



杨文哲/YANG Wenzhe, 徐迈/XU Mai, 白琳/BAI Lin
(北京航空航天大学, 中国 北京 100191)
(Beihang University, Beijing 100191, China)

摘要: 提出了一种应用于视频质量增强算法的动态结构性剪裁算法 Maskcut, 它可以有效提高基于深度学习的视频质量增强算法的运行速度。Maskcut 是一种通用的剪裁思路, 支持绝大多数的基于卷积神经网络(CNN)深度学习网络模型的剪裁加速。基于原模型中已经训练好的参数数据, Maskcut 使用一种针对剪裁加速的二次训练策略来进一步微调参数, 从而在保证模型有效性损失不大的同时, 缩短模型运行时间。以一种先进的视频质量增强算法——多帧质量增强 2.0(MFQE 2.0)为目标, Maskcut 剪裁后可以快速达到峰值信噪比(PSNR)指标损失低于 1%、时间缩短 10% 以上的加速指标。

关键词: 模型加速; 图像质量增强; 结构性剪裁

Abstract: Maskcut, a dynamic structural clipping algorithm for video quality enhancement is proposed, which can effectively improve the speed of video quality enhancement algorithm based on deep learning. Maskcut is a general tailoring idea that supports most of the tailoring acceleration based on the deep learning network models for convolutional neural networks (CNN). Based on the trained parameter data in the original model, the secondary training for tailoring acceleration is carried out to further fine-tune the parameters. With an advanced video quality enhancement algorithm, the multi-frame quality enhancement 2.0 (MFQE 2.0) as the goal, the peak signal-to-noise ratio (PSNR) index is less than 1% and the time is shortened by more than 10% after Maskcut clipping.

Keywords: model acceleration; image quality enhancement; structural tailoring

DOI: 10.12142/ZTETJ.202101006

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210125.1047.004.html>

网络出版日期: 2021-01-25

收稿日期: 2020-12-25

随着多媒体及 5G 时代的到来, 视频传输的速率和带宽得到了有效提高, 人们对于高清视频的需求也变得越来越高。但是由于许多拍摄软硬件条件不高或多层多级中转压缩过程复杂等原因, 使得视频质量不够清晰, 因此高清视频仍有着很大的提升空间。目前, 针对图像视频质量增强的算法, 大多是基于深度学习的

庞大计算量。这些算法在应用时往往存在参数冗余、计算量大、时间耗费多等问题, 这也是近两年来深度学习领域需要着重解决的问题。目前大量的优化算法大多是基于浮点运算次数(FLOPs)的仿真工作。面对实际问题和模型, 如果没有底层的硬件支持和推理加速, 很多算法使用时并不能获得理想的加速效果。本文中,

我们主要聚焦于实际硬件加速效果, 而非理论计算量的改变。

1 视频质量增强模型加速算法的相关工作

1.1 视频质量增强

在视频质量增强方面, 由 GUAN Z. Y. 等提出的压缩视频质量提升 2.0

(MFQE 2.0)模型^[1],是目前实用性相对较好的一种深度学习算法。该算法以卓越的速度和性能效果优于同时期的其他视频质量增强算法。本文中,我们以此作为剪裁算法的实践模型对象,提出了一种具有通用性的剪裁算法。在MFQE 2.0模型中,增强算法分为两个子网络,分别对应数据特征提取运动补偿(MC)子网络和数据恢复质量增强(QE)子网络。其中,数据恢复QE子网络为卷积神经网络(CNN)模型,对于剪裁算法的操作性和兼容性较高。针对不同质量(QP)的视频数据集,MFQE 2.0能够训练对应的模型,并能根据视频前后好帧的同一物体的像素信息,对当前低帧图像进行质量增强。由于MFQE 2.0训练较慢,且所采用的从训练后的大模型剪裁小模型的方法也存在合理性^[2],故在MFQE 2.0训练好的模型基础上进行剪裁,可以达到更快、更好的效果。

1.2 模型剪裁加速

在剪裁加速方面,近两年来相关论文的成果众多,如HAN S.等^[3]提出的剪裁-量化-编码的三部曲结构,是压缩模型比较经典的方法。本课题也是从这种压缩技巧入手,但目的是缩短模型的运行时间。本文中的剪裁方式主要为随机剪裁,即根据网络中的参数大小进行剪裁。这种方法能够减少计算量和参数数量,但剪裁结果为稀疏性矩阵,而运算时许多框架对于稀疏性矩阵的卷积(底层会转为乘法运算)并无有效加速;因此,时间上的加速效果不明显。一般做法是,底层运算采用特定的计算库,但是这种做法的通用性较差,与框架结合效果不佳。而李浩等^[4]提出的基于滤波器剪裁的加速方法,以通道为单位,并以滤波器的标准差作为衡量重

要性标准来进行剪裁。虽然这种思路能够在不依赖其他框架的情况下,有效缩短模型的运行时间,但依然存在优化空间,如缺乏在重训过程中的更自由的调整策略和自动调整剪裁阈值的机制。本文所提的加速算法也在此基础上进行了改进。

面对工业界亟待解决的问题,一种基于算法层面的有实际加速效果的剪裁就显得非常重要。因此,本文主要针对深度学习算法的时间耗散问题,并以一种视频质量增强算法为例,提出一种可以通过自动调整剪裁标准进行通道性剪裁的方法。同时,实验结果表明,该方法切实缩短了模型的运行时间,加速模型在工业界的落地应用。

2 Maskcut动态剪裁方法

2.1 动态通道剪裁

基于李浩提出的通道性剪裁的基础概念,本文的剪裁算法得以提出。在以往的通道性剪裁中(如图1所示),每一层的滤波器维度为四维,图1的每一个卷积核被视作一个质点;核矩阵为滤波器在输入通道、输出通道维度的二维表示。无论是逐层剪裁还是整体剪裁,都要先通过L1-norm或核矩阵的标准差等指标排

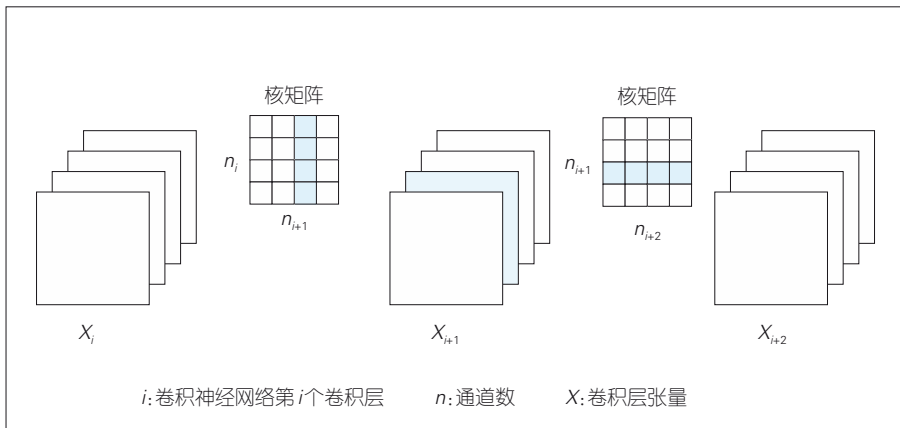
序来确定要剪裁掉的通道,然后重训练,提高模型准确率,并同时对应剪掉下一层的核矩阵对应通道的输出。

在剪裁后重训的过程中,滤波器的数值会经损失函数反向梯度传播而不断改变。如果我们依然按照原先的标准,那么此时可能有一些通道变得不符合规则。模型在剪裁后都是需要训练的,所以我们可以将训练过程中的剪裁做形式上的改变:不需要立刻导出小通道数模型,而是如图2所示,将对应通道的前向传递函数和反向梯度传递函数进行可控阻断,在保留被剪裁参数的同时,取得和剪裁滤波器通道一样的效果;在重训过程中,也可以随时更改开启关闭的通道。

2.2 半自动剪裁标准调整机制

通道性剪裁有着简单直接的优势,但一个比较突出的问题是每层剪裁的数量需要预先设定。无论剪裁标准是滤波器的标准差、L1-norm值,还是其他,都只是标准的差异而已,而选定的剪裁比例或阈值却没有对应的优化机制。本文提出了一种半自动剪裁标准调整机制,对于动态剪裁可以考虑选择。

以输出通道为单位,进行权重排



▲图1 传统通道性剪裁

序,记作 $list$,则各卷积层中每个输出通道对应的L1-norm正则化值的百分比为:

$$y = \frac{list(x) - \min[list]}{\max[list] - \min[list]}$$

根据 y 的分布情况可首先获得分布曲线,再根据趋势来合理设定阈值或百分比,以决定哪些通道应该被剪掉^[4]。通过分析此类曲线斜率的物理意义,我们可以得出这样的结论:如果某处斜率过高,那么对应一阶导数的极值点,就可得到一个由参数数据大小决定的简单阈值;我们通过拟合样条曲线后进行求导(或差分),并根据导函数极大值点或最大值点来确定对应的合适分割位置,并将其作为一种近似的剪裁阈值。

通过以上的导函数寻找极大值点,可以辅助寻找适合被剪裁的通道。如预设剪裁 $[a,b]$ 范围的通道数(百分比),然后使用寻找极大值点的方法,从 $[a,b]$ 区间内寻找导函数最大值点,以此作为当前剪裁的真正比例。通过动态剪裁的方法,可以在每轮重新训练时,不断地重新确认剪裁的比例值。

2.3 Maskcut剪裁算法

本文使用L1-norm作为衡量通道重要性的指标。通过对L1-norm进行排序,算法将通道整体的重要度进行区分,剪裁那些不太重要的通道,并采用动态剪裁的方式随时调整选择的通道位置,或利用半自动剪裁标准调整机制,随时调整选择的通道数量。

剪裁算法的步骤具体如下:

(1)以输入通道为主,以卷积层为单位,计算各个通道的L1-norm并排序;

(2)设定各层的初始剪裁比例或范围;

(3)根据比例范围和半自动剪裁

标准调整机制,找出此轮真正的剪裁比例;

(4)确认被剪裁的通道对应的开关关闭,对模型进行微调重训;

(5)重复步骤(3)~(4),直至满足要求,最后将开通道对应参数导出至新模型,完成剪裁。

其中,步骤(3)非必需,可根据实际情况选择是否进行。

2.4 动态剪裁实践流程

采用通道开关后,算法由原来的筛选-重训的流程变成了如图3所示的流程。在每轮训练中,新的流程可以更自由地重新修改训练通道。最重要的是,通过不断调整训练的通道数量,并搭配自动选择阈值分割线的机制,就可以进一步实现动态剪裁和重训,从而将两者有机结合起来。待训练效果可以接受时,再通过导出开通道的参数,原模型就可以转变为一个简单且低维的小模型,从而完成剪裁。

2.5 理论分析

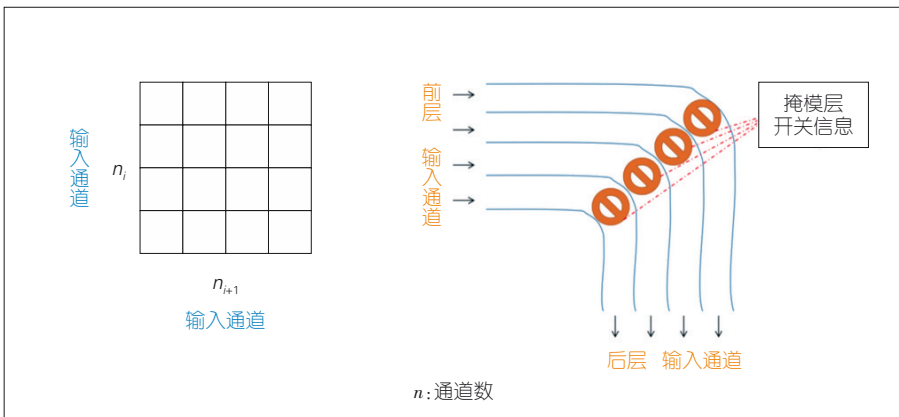
本文所提的剪裁算法的最终目的是实现时间加速,但为了保持内容完整,我们同样进行了FLOPs的理论分析。该算法减少的计算量是相同的^[4],以图1的通道为例,假设相邻两个特征图之间卷积核维度为 $k \times k \times n_i \times n_{i+1}$,输出特征图维度为 $w_{i+1} \times h_{i+1} \times n_{i+1}$,故此时FLOPs为:

$$FLOPs = n_{i+1} \times n_i \times k^2 \times h_{i+1} \times w_{i+1} \quad (1)$$

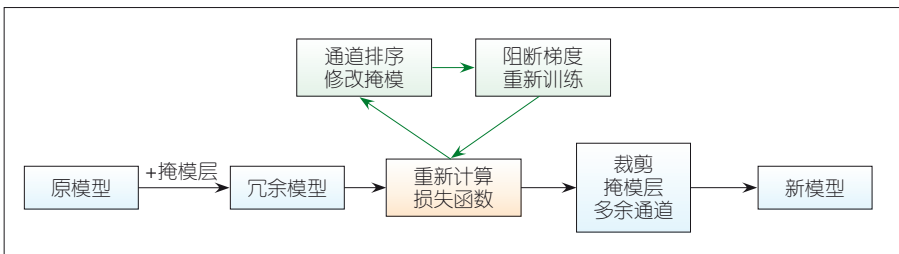
假设裁剪一个通道(如图2中的蓝色部分),那么减少的浮点计算量包括相邻两个卷积层的影响,总共减少了 $(n_i h_{i+1} w_{i+1} + n_{i+2} h_{i+2} w_{i+2}) k^2$ 的计算量。

3 实验结果

本文所提的基础模型是基于Tensorflow 1.0版本的MFQE 2.0模型,故我们以Tensorflow 1.14为运行框架环境,来测试MFQE 2.0的QP32视频



▲图2 通道开关实现剪裁效果



▲图3 加入通道开关后剪裁流程

数据集及模型。

3.1 MFQE 2.0 模型分析

3.1.1 概述

在剪裁前,应对模型的运行时间、计算时间开支进行分析和了解。

我们利用 Tensorflow 内置的时间分析工具 timeline 进行测量。由于 timeline 是官方内嵌的时间测量工具,虽有一定的波动性,但相对来说可靠得多。通过 timeline 分析,不仅可以计算出整个模型的单次运算时间,还可以获得每个卷积层的运算时间;因此对于卷积层的通道剪裁来说,有着更明显的对比效果,故我们以此工具作为时间测量的手段。

在 MFQE 2.0 网络中,前面紧凑计算对应的是特征提取网络,而后面相对耗时长的一部分是重建网络,因此优先剪裁的对象应是 QE 重建网络中的 CNN 那一部分。这种做法对于其他算法模型来说也有着很强的通用价值。

3.1.2 参数分析

为了进一步了解模型和参数,应

对模型的参数进行分析。首先我们以层为单位,对 MFQE 2.0 中比较容易剪裁的 QE 重建网络部分中的各个卷积层的参数进行分析,并且对每个权重值的绝对值进行排序和计数,对权重值和索引比进行归一化,从而能够统计出权重的分布,具体的情况如图 4 所示。

这里,我们以 QP=32 数据集下的 MFQE 2.0 模型为例,提取 QE 子网中各卷积层的参数,并以卷积层为单位,分析权重绝对值的分布情况。图 4 中 x 为参数排序后的数组 $list$ 索引比例, y 为排序后 x 对应索引号处的比例值, $y = \frac{list(x) - \min[list]}{\max[list] - \min[list]}$ 。我们从曲线斜率的变化可以看出:绝大多数的参数都较小, CNN 网络中参数较大的相对较少。由于参数绝对值的大小一般表征着重要性,绝对值大的权重对于计算结果影响更大。所以相对来说,应该优先剪裁那些权重较小的值。从图中也可以看出,在模型参数的分布上体现了模型目前仍然具有冗余性,并且存在可以剪裁的空间。

3.2 剪裁实验结果

我们以 QP=32 的视频数据为例,对 MFQE 2.0 的 QE 部分网络进行剪裁加速实验。

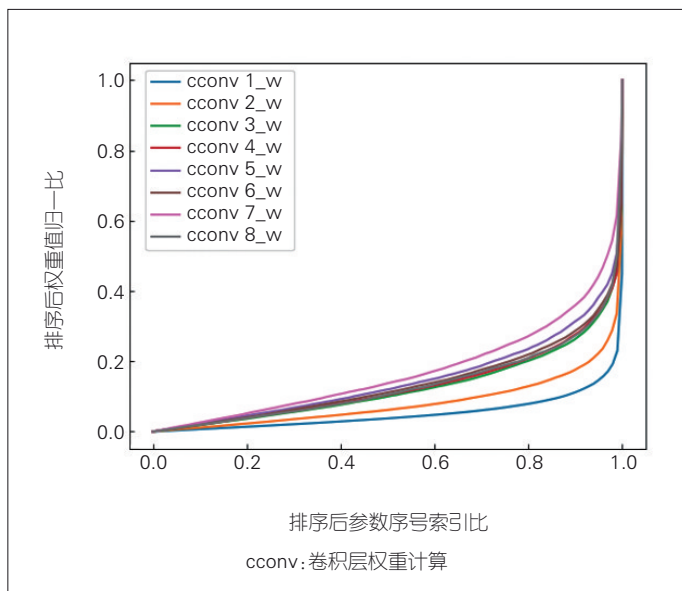
QE 部分网络的原始维度主要有 7 层,通道数均为 32,我们在此基础上进行整体剪裁加速实验。为了方便统计和测试,采取各层剪裁的通道数时刻保持一致,通道数不断调整的方法进行实验。

3.2.1 剪裁时间结果

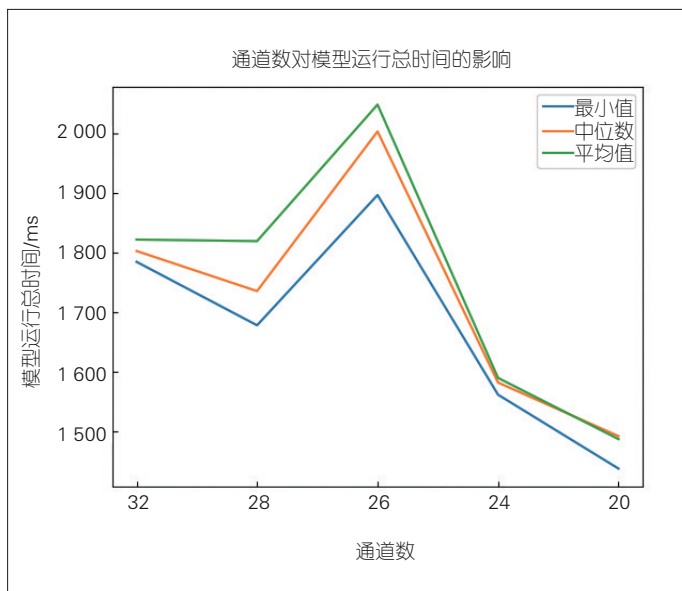
Tensorflow 静态图的特点导致在模型剪裁时,只能采取稀疏+参数数据迁移的方案。而在稀疏之前,需要先对不同剪裁的具体时间加速效果进行实验,这也能给后面的稀疏实验提供优化对比。

因此,我们将原方案 32 通道的模型做更改,并用 timeline 来测试时间,经平均处理消除不确定性,其结果如图 5 所示。

可以看出,随着通道数的减少,模型运算的总时间也相应减少。虽然只是裁剪部分层的通道,但由于 QE 网络所占运行时间很长,所以相



▲图4 各层参数绝对值统计曲线(QP=32)



▲图5 模型总时间与通道数折线图

对来说只检测 QE 网络部分的时间加速效果依然不错。

当方案中的通道数为 26 时,数据平均处理后的时间仍比原模型时间更久一些。经查阅资料后得知,在计算机底层计算时,由于通道数是卷积层滤波器四维矩阵中的两个维度,因此在计算时都是按照感受野为单位进行的,可能会因维度数目和计算机底层的一些内存单元大小的匹配,存在时间长短的区分。原模型通道数仅为 32,因此在通道数目减小的过程中,如果不是减少量显著,则存在时间消耗不减反增的可能,这与计算机底层内存块的空间大小等都有关系。

总的来说,通过剪裁通道数能够很方便地实现时间加速,但至于性能如何,则需要通过稀疏实验进行验证。

3.2.2 剪裁结果分析

通过动态的通道剪裁方案,算法能够在每次训练时不直接将某些参数置为 0,而是通过对掩模层的“开

关”进行学习,这样就能实现损失函数和反向梯度传播的对应更新,以更快速准确地实现视频质量增强的网络剪裁效果。我们对之前提到的算法和方案进行了落实,在 MFQE 2.0 网络中,以 QP=32 的数据和训练好的模型对低质量帧(non-PQF)进行实验。

如果不进行相应的参数恢复,而直接进行剪裁,那么峰值信噪比(PSNR)和结构相似性比(SSIM)的运行结果如表 1 所示。

可以看出,如果直接进行剪裁,会导致模型的性能变得非常差。也就是说,虽然模型存在着大量权重小的参数,但仍不能简单忽略。如果直接删除一些小权重的通道,而不对其他的通道进行适当的修改和调整,那么准确率和质量增强效果依然会大打折扣。

使用掩模层进行不断地重训后,通过调整通道开关闭合,就可以计算损失函数,阻断部分梯度反向传播的更新。

经上述训练后,我们分别测试了

所得模型的时间和效果,结果如表 2 所示。分析 PSNR 对应列与直接剪裁不重训,我们可以看到模型重训对模型的性能恢复效果。

我们采用 PSNR 作为质量增强效果的客观标准。基于此种动态通道的剪裁方法有着较好的卷积神经网络的通用性和一般规律性,并在本文的视频质量增强任务上实现了有效的效果。

总的来看,如果主要分析 Δ PSNR 和总时间的平衡,分析结果如表 3 所示。在目前的实验结果中,在通道数为 24 时,能够实现以 0.7% 的性能损失换取 12.5% 的总时间缩短,达到预期指标。

4 结束语

本文提出了一种动态的通道剪裁方法,以 L1-norm 作为衡量滤波器通道重要程度的标准,通过动态剪裁、统一设定剪裁比例,对已经训练好的 MFQE 2.0 模型进行剪裁加速。在 MFQE 2.0 的 QP 32 数据集中,通过微调模型后,我们发现:当将 QE 网络的 32 通道数剪裁至 24 时,可以达到以 0.7% 的性能损失换取 12.5% 的总时间缩短的指标。在剪裁后网络最终导出之前,随着迭代次数的推进,模型剪裁效果会更好,而且具有一定外扩性,即最后的模型再适当增大通道时,由于动态剪裁的选通机制,可以相对更轻松微调至其他数目通道的模型。通过动态通道剪裁,完成了 MFQE 2.0 的视频质量增强的模型有效加速。

本文还提出一种半自动剪裁标准调整策略,通过拟合函数的一阶导数极大值或最大值寻找,辅助决定剪裁最佳比例,后续应尝试从结果反馈信息或二分类辅助自动决定剪裁比例。

▼表 1 直接剪裁峰值信噪比结果对比

模型	原模型 32 通道	28 通道	24 通道
PSNR	0.305	0.108	0.007
损失比/%	—	64.6	97.7

PSNR: 峰值信噪比

▼表 2 重训后峰值信噪比结果对比

通道数	Δ PSNR	损失比/%
32	0.305	—
28	0.291	4.6
24	0.303	0.7
20	0.230	24.6

PSNR: 峰值信噪比

▼表 3 重训后 SSIM 与增强每帧总时间指标情况

通道数	Δ SSIM	损失比	总时间	加速比/%
32	0.011	—	1 783.9	—
28	0.009	18.2	1 677.4	6.0
24	0.009	18.2	1 560.9	12.5
20	0.008	27.3	1 436.8	19.5

SSIM: 结构相似性比

参考文献

- [1] GUAN Z Y, XING Q L, XU M, et al. MFQE 2.0: a new approach for multi-frame quality enhancement on compressed video [J]. IEEE transactions on pattern analysis and machine intelligence, 2019: 1. DOI: 10.1109/tpami.2019.2944806
- [2] ZHU M, GUPTA S. To prune, or not to prune: exploring the efficacy of pruning for model compression [EB/OL]. (2018-06-23) [2020-12-22]. <http://arxiv.org/abs/1710.01878>
- [3] HAN S, MAO H Z, DALLY W J. Deep compression: compressing deep neural networks with pruning, trained quantization and Huffman coding [EB/OL]. [2020-12-22]. <https://arxiv.org/abs/1510.00149>
- [4] 李浩, 赵文杰, 韩波. 基于滤波器裁剪的卷积神经网络加速算法 [J]. 浙江大学学报(工学版), 2019, 53(10): 1994-2002
- [5] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient ConvNets [EB/OL]. [2020-12-22]. <https://arxiv.org/abs/1608.08710>
- [6] LEBEDEV V, LEMPITSKY V. Fast ConvNets using group-wise brain damage [C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 2554-2564. DOI: 10.1109/cvpr.2016.280
- [7] WEN W, WU C P, WANG Y D, et al. Learning structured sparsity in deep neural networks [EB/OL]. [2020-12-22]. <https://arxiv.org/abs/1608.03665>
- [8] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications [EB/OL]. [2020-12-22]. <https://arxiv.org/abs/1704.04861>
- [9] LI H, KADAV A, DURDANOVIC I, et al. Pruning filters for efficient ConvNets [EB/OL]. [2020-12-24]. <https://arxiv.org/abs/1608.08710>
- [10] POLYAK A, WOLF L. Channel-level acceleration of deep face representations [J]. IEEE access, 2015, 3: 2163-2175. DOI: 10.1109/access.2015.2494536

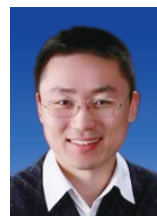
作者简介



杨文哲, 北京航空航天大学电子信息工程学院在读硕士研究生; 研究方向包括深度学习、图像压缩等。



徐迈, 北京航空航天大学电子信息工程学院教授、教育部“青年长江学者”、中国图象图形学会青工委副主任; 研究方向包括图像处理、视频压缩、视频通信、计算机视觉与人工智能等; 2016年获教育部霍英东青年基金资助, 2017年获人工智能学会技术发明一等奖(第二完成人), 2018年获教育部科技进步一等奖、中国电子学会优秀科技工作者, 2019年获国家优秀青年基金资助, 2020年获北京市杰出青年基金资助; 发表论文100余篇。



白琳, 北京航空航天大学网络空间安全学院教授; 主要研究方向包括通信网络安全、无线通信、物联网、无人机通信等领域等; 主持国家自然科学基金项目3项、国家重点研发计划项目子课题1项, 获国家自然科学基金优秀青年科学基金资助, 获第四届中国出版政府奖、中国电子学会自然科学二等奖、国家科技进步二等奖; 发表SCI期刊文章66篇, 著有英文专著2部、中文专著3部。



基于图神经网络的 视频推荐系统

Video Recommender System with Graph Neural Networks

高宸/GAO Chen, 李勇/LI Yong, 金德鹏/JIN Depeng

(清华大学, 中国 北京 100084)
(Tsinghua University, Beijing 100084, China)

摘要: 提出了一种基于图神经网络的视频推荐模型,将用户的视频观看序列型行为建模为图结构,用结点代表用户与视频,用边代表行为,引入两种类型的向量传播方法分别对用户的长期兴趣与短时兴趣进行建模。其中,通过用户结点与视频结点的双向传播刻画长期兴趣,借助视频结点切换关系的单向传播刻画短时兴趣,并通过多层向量传播实现对图上高阶邻接信息的捕捉。在一个真实世界的视频网站观看数据集上的实验表明,提出的方法与现有最佳方法相比,其推荐精准度得到了有效提升。进一步的实验表明,该方法能够有效缓解数据稀疏性的问题。

关键词: 视频推荐系统;用户兴趣建模;图神经网络;深度学习

Abstract: A novel recommendation model with graph neural networks is proposed. Users' sequential video-watching behaviors are first constructed as a graph, which represents users and videos as nodes, and behaviors as edges. Then two kinds of embedding propagation methods are introduced for capturing users' long-term and short-term preferences, respectively. Specifically, a user-item bi-directional embedding propagation layer is used for capturing long-term preferences while an item-item embedding propagation layer for capturing short-term preferences. Moreover, the multi-layer propagation is proposed to extract high-order connectivity. Experiments on a real-world video-watching dataset verify that the proposed method can outperform the state-of-the-art methods. Further experiments demonstrate that the proposed method can effectively alleviate the data sparsity issue.

Keywords: video recommender system; user preference modeling; graph neural network; deep learning

DOI: 10.12142/ZTETJ.202101007

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210120.1010.002.html>

网络出版日期: 2021-01-20

收稿日期: 2020-12-10

在信息超载时代,个性化推荐系统^[1-2]成为用户获取信息的主要方式。推荐系统通过收集用户的

历史行为来推断用户兴趣,进而生成推荐列表。与常见的电商网站推荐系统^[3]不同,视频网站上的用户行为具有两个重要特性。首先,用户的视频观看行为呈现出高度的序列性。一段时间内浏览的视频表现出极高的相关性,且浏览的前后顺序十分重要,因此需要对用户行为进行序列化

建模。其次,用户可能存在短期观看某一类/系列多个视频的“短时”兴趣,呈现出突发、多样的特点。因此,我们需要从长期兴趣与短时兴趣两方面对用户的兴趣进行细粒度化的建模。

针对序列化行为的推荐问题,现有的方法^[4-6]仍然存在两部分缺陷。

基金项目: 国家重点研发计划(2018YFB1800804);国家自然科学基金(U1936217、61971267、61972223、61941117、61861136003);北京自然科学基金(L182038);北京国家信息科学与技术研究中心基金(20031887521);清华大学-腾讯联合实验室项目

首先,仅仅使用权重或者卷积/循环神经网络对不同历史行为进行隐式建模的方法,缺乏对序列化行为中视频切换关系的显式建模;其次,目前的推荐方法没有考虑针对用户长期与短时兴趣的细粒度建模。本文中,我们设计了一种基于图神经网络的推荐模型,通过两种向量传播方式来分别对用户的长期兴趣和短时兴趣进行建模。此外,我们还引入了多层向量传播以捕捉图上高阶邻接信息。

1 问题定义

视频推荐系统的目标是尽可能地满足用户的需求,即为用户推荐最符合其兴趣的视频。在视频推荐系统中,相关输入数据为用户历史视频观看的行为序列,其中,序列中的前后关系代表用户观看视频的先后顺序关系。输出数据则为可计算给定用户下一次观看给定视频的概率模型。在得到该模型后,我们可对所有

候选视频进行概率计算,并按照概率预估值从大到小排序,得到推荐列表。

2 方法设计

这里我们提出一种基于图卷积网络的视频(VGCN)推荐模型,具体如图1所示。该推荐模型主要由4个流程部分构成:构建包含用户视频结点与行为边的异构图、构建嵌入层以得到用户与视频的表征向量、设计向量以刻画用户的长期兴趣与短时兴趣、引入预测层得到用户观看视频的概率。

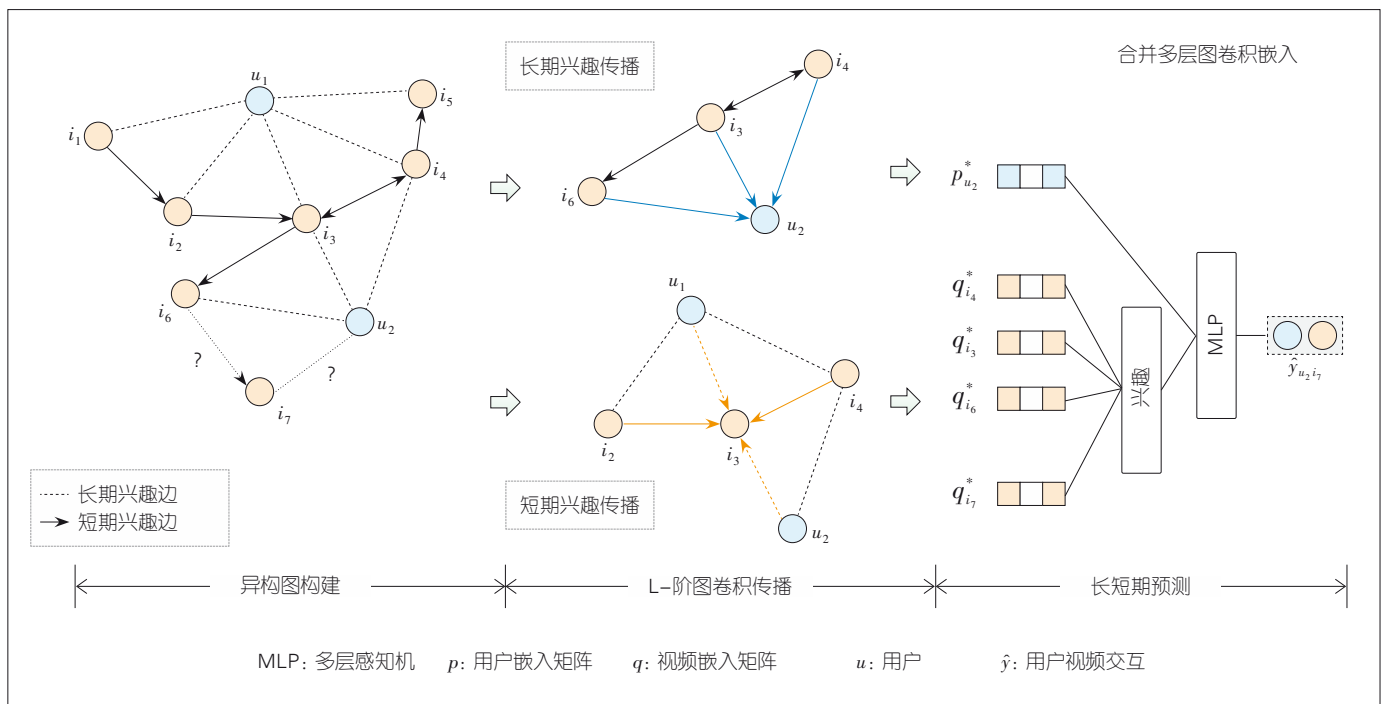
2.1 异构图构建

图是一种具备强大数据表示能力的结构。在视频推荐系统中,一种直观并且有效的做法是,把用户与视频分别表示为图中两种类型的结点,并将用户的观看行为建模为图上的边。具体而言,我们构建异构图 $G=\langle V, E \rangle$, 其中, V 表示所有结

点的集合, E 表示所有边的集合。同时集合 V 中有两类结点:用户结点 $u \in U$ 与视频结点 $i \in I$ 。集合 E 中有两类边:一类是用户与视频的交互边 $r \in R$, 其中 r_{ui} 代表用户 u 与视频 i 存在交互行为(观看);另一类是视频切换边 $r \in T$, 其中 t_{ij} 代表视频 i 到视频 j 的切换行为,并且该边的权重由数据集中所有用户的切换次数决定。如果该权重为0,即边不存在,则代表没有用户产生从观看 i 切换到观看 j 的行为。总的来说,我们得到了包含两类结点与两类边的异构图。

2.2 嵌入层

针对图的表征学习可定义为:通过机器学习的方法,为点、边或图学习其在低维空间的表征。该方法可以将高维的图数据转换为低维特征向量,实现预测、分类等任务^[7]。在通用视频推荐任务中,由于用户画像、视频属性等数据较难收集,用户与视频一般而言仅有身份标识(ID)



▲图1 视频图卷积网络推荐模型示意图

特征,即仅有用户与视频的原始编号。因此,我们针对ID特征设计适用于独热编码的嵌入层,为用户与视频分别建立嵌入矩阵 \mathbf{P} 与 \mathbf{Q} 。 \mathbf{P} 的维度为 N 乘以 D , \mathbf{Q} 的维度为 M 乘以 D 。其中, N 为用户的数目, M 为视频的数目。 D 为低维空间的维度,是一个可以调整的超参数,其过大的维度会带来过拟合问题,而过小的维度则存在欠拟合问题。

我们通过用户与视频的嵌入矩阵与独热编码可得到每个用户与视频的低维表征向量,如公式(1)所示:

$$p_u = \mathbf{P}^T v_u^U, q_i = \mathbf{Q}^T v_i^I, \quad (1)$$

其中, v_u^U 与 v_i^I 分别为用户与视频的独热编码。

独热编码是一种仅有一个位置为1且其余位置为0的高维向量(值为1的位置即为原始编号)。用户独热编码的长度为 N ,视频独热编码的长度为 M 。在嵌入矩阵完成随机初始化后,模型的后续部分将从嵌入矩阵得到最终的预测结果。当基于预测损失的随机梯度下降时,嵌入矩阵即可从初始化的随机向量逐渐调整至可刻画用户与视频特征的高质量表征向量。

2.3 向量传播层

我们首先建立了上述用户与视频的嵌入矩阵。该嵌入矩阵可以被视为第0层用户/视频向量。接着,我们设计向量传播层以利用图上的高阶邻接关系,以捕捉用户的长期与短时兴趣。图卷积网络是一类最典型的图神经网络^[8]。向量传播是图卷积网络的核心模块,其核心思想是将向量传播给图上的邻居结点,以实现图结构邻接性到向量相似性的转化,并可通过多层向量传播实现对高阶邻接关系的建模。借助向量传播方

法,图卷积网络在诸多任务上取得了当前最佳性能^[8-9]。

在视频推荐系统中,需要对用户进行两方面的兴趣建模:长期兴趣与短时兴趣。其中长期兴趣侧重于用户较为固定的、不随时间变化的兴趣,短时兴趣则与之相反。具体而言,我们通过用户结点的表征向量对其长期兴趣进行建模,通过用户上一时刻交互的结点的表征向量对其短期兴趣进行建模。这种做法与用户长期兴趣与短时兴趣的物理意义相契合。

2.3.1 长期兴趣向量传播层

用户观看视频的行为被表示为图上用户结点与视频结点之间的边。如果将用户观看过的所有视频结点的向量表征往其传播,聚合之后自然就代表了与具体时间无关的长期兴趣。因此,我们可以通过用户结点-视频结点向量传播对长期兴趣进行建模。具体而言,某个用户观看过的所有视频往其传播,向量传播公式如下:

$$p_u^{(l+1)} = \sigma(W_1^{(l+1)}(p_u^{(l)} + \text{aggregate}(q_i^{(l)} | i \in R_u)) + b_1^{(l+1)}), \quad (2)$$

$$q_{i,1}^{(l+1)} = \sigma(W_1^{(l+1)}(q_{i,1}^{(l)} + \text{aggregate}(p_u^{(l)} | u \in R_i)) + b_1^{(l+1)}), \quad (3)$$

其中, σ 是非线性激活函数, $W_1^{(l+1)}$ 与 $b_1^{(l+1)}$ 分别是第1层往 $l+1$ 层传播时的网络参数与偏置参数(此处我们添加下标1以与后文的短时兴趣相关参数加以区分), R_u 是用户 u 交互过的所有视频, R_i 是观看过视频 i 的所有用户, $\text{aggregate}(\cdot)$ 是向量聚合操作(常见做法有求平均等), $p_u^{(l)}$ 为用户 u 第 l 层的表征向量, $p_u^{(l+1)}$ 为用户 u 第 $l+1$ 层的表征向量, $q_i^{(l)}$ 为视频 i 第 l 层的表征向量, $q_i^{(l+1)}$ 为视频 i 第 $l+1$ 层的表征

向量。

总的来说,前文所述的向量操作实现了长时兴趣侧从低层向量到高层向量的计算方式。随着层数的逐渐提升,更高阶的邻接关系将会被提取至表征向量中。但值得一提的是,层数不能过高,这是因为向量传播可以被理解成一种局部图的近邻平滑作用,如果层数过深,则相当于实现了全局平滑,反而会使学习到的表征向量无效。

2.3.2 短时兴趣向量传播层

上述长期兴趣向量传播层通过忽略序列关系的历史行为边传播,来刻画用户的长期兴趣。接着,我们进一步设计用于对用户的短时兴趣进行建模的向量传播方法。考虑到用户的短时兴趣与视频观看的切换行为需要相契合,我们采用基于视频切换行为的有向边来设计向量传播方法。换言之,向量传播的路径就是上一个视频到下一个视频的有向边。由于此处不涉及用户结点的表征向量,因此,我们可以实现长期兴趣与短时兴趣的解耦建模。

具体而言,传播公式如公式(4)所示:

$$q_{i,2}^{(l+1)} = \sigma(W_2^{(l+1)}(q_{i,2}^{(l)} + \text{aggregate}(q_j^{(l)} | j \in T_i)) + b_2^{(l+1)}), \quad (4)$$

其中, σ 是非线性激活函数, $W_2^{(l+1)}$ 与 $b_2^{(l+1)}$ 是从第 l 层往 $l+1$ 层传播时的网络参数与偏置参数(此处我们添加下标2,以与前文所述的长期兴趣相关参数加以区分), T_i 是存在往 i 有向边的所有视频的集合。此处的 $\text{aggregate}(\cdot)$ 是带权重的聚合操作,该权重为往 i 的有向边的权重,即存在该切换行为的用户的个数,同时该权重可从一定程度上对切换关系的强弱进行建模。

2.4 预测层

在以上向量传播层的基础上,我们进一步设计得到最终预测结果的预测层。通过长期兴趣向量传播与短时兴趣向量传播,我们得到了两部分视频向量。此处我们使用求和操作,将两部分合并为一部分,即 $q_i^{(l+1)} = q_{i,1}^{(l+1)} + q_{i,2}^{(l+1)}$ 。

考虑到不同层数的用户或视频向量包含了不同阶数的图上邻接关系,我们将不同层数的向量使用拼接操作进行聚合,如公式(5)和公式(6)所示:

$$p_u^* = p_u^{(0)} \parallel p_u^{(1)} \parallel \dots \parallel p_u^{(L)}, \quad (5)$$

$$q_i^* = q_i^{(0)} \parallel q_i^{(1)} \parallel \dots \parallel q_i^{(L)}, \quad (6)$$

其中, L 为一个可以调整的超参数, \parallel 代表向量的拼接操作。

随后,我们通过一个基于注意力网络的预测函数,对给定的用户与视频预测观看概率,具体如公式(7)所示:

$$\hat{y}_{ui} = MLP(p_u^* \parallel ATN(q_i^*, R_u)), \quad (7)$$

其中, MLP 代表一个多层感知机, ATN 代表一个注意力网络。最终输出的 \hat{y}_{ui} 为一个从0到1之间的概率值,该值越大,用户 u 越有可能观看视频 i 。

2.5 训练方法

在获得对于任意给定用户与视频的观看概率预估后,我们基于对数损失函数进行优化。由于数据中仅记录了用户观看过的视频,即正样本,我们需要从未观看的视频中随机采集一些样本作为负样本。对于正样本而言,模型的预测结果要尽可能接近1;对于负样本,模型的预测结果要尽可能接近0。损失函数具体计算方式如公式(8)所示:

$$Loss = - \sum_{(u,i) \in Y^+ \cup Y^-} y_{ui} \log \hat{y}_{ui} + (1 - y_{ui}) \log (1 - \hat{y}_{ui}), \quad (8)$$

其中, Y^+ 与 Y^- 分别为正样本集合与负样本集合, y_{ui} 与 \hat{y}_{ui} 分别为数据真实标签与模型预测输出。

基于以上损失函数,我们可通过随机梯度下降进行更新。在我们提出的VGCN模型中,需要学习的参数主要为用户与视频的嵌入矩阵 P 与 Q 。向量传播操作中的 W 与 b 等参数则占据着较少的参数量,而每加深一次图卷积网络,只会引入额外的 W 与 b 等参数。因此,不同深度的VGCN模型的参数量基本相当,均约等于嵌入矩阵的参数量大小。而嵌入矩阵参数量大小不仅与用户数目和视频数目的和呈线性关系,也与嵌入矩阵向量维度 D 呈线性关系。

3 实验验证

为了验证提出的VGCN方法的有效性,我们对真实视频观看数据集进行了推荐性能的验证。

3.1 实验设置

3.1.1 数据集

我们一个视频网站上收集了2020年10月的用户视频观看行为数据。由于完整数据规模过大,我们随机选取了一部分用户。经过预处理后的数据,包括了60 813个用户与292 286个视频产生的14 952 659条

观看记录。对于每一个用户而言,其观看记录为一条包含了若干个视频的序列。

3.1.2 性能指标

视频推荐乃至通用推荐系统最常使用的指标为排序指标,其中最具有代表意义的指标为特征曲线下方的面积(AUC)、平均倒数排名值(MRR)与归一化折损累计增益(NDCG)^[10]。AUC可衡量模型对于所有正样本与负样本相对关系的区分能力,MRR衡量模型将正样本排在列表靠前位置的能力,NDCG则衡量模型排序结果与理想排序结果的距离。

3.1.3 基线模型

我们选取两个极具竞争力的模型作为基线模型:卷积序列嵌入推荐模型(CASER)^[5]与深度兴趣网络(DIN)^[6]。其中,CASER通过卷积网络建模用户的行为序列,DIN通过注意力网络建模用户行为序列。

3.2 推荐性能比较

我们首先对整体的推荐结果进行比较,如表1所示。

由表1可知,与现有模型相比,我们提出的VGCN模型在AUC、MRR、NDCG@1、NDCG@2等指标上,可以有效且稳定地提升推荐性能,且平均相对提升值约为1.7%。对于推荐系统模型而言,该提升值是显著的。

图2则展示了不同方法训练时的模型损失曲线。由图2可以看出,我

▼表1 视频推荐精准度性能比较

模型	AUC	MRR	NDCG@1	NDCG@2
CASER	0.7471	0.8788	0.7576	0.9105
DIN	0.7561	0.8856	0.7712	0.9156
VGCN	0.7781	0.8943	0.7886	0.9220

AUC:特征曲线下方的面积

CASER:卷积序列嵌入推荐模型

DIN:深度兴趣网络

MRR:平均倒数排名值

NDCG:归一化折损累计增益

VGCN:基于图卷积网络的视频推荐算法

们的VGCN方法可以取得更小的训练损失。

3.3 稀疏度影响研究

在推荐系统尤其是视频推荐系统中,数据稀疏十分重要。具体而言,对于不同稀疏性的用户,能否均取得较好的效果,是衡量一个推荐模型好坏的重要指标。因此,我们将用户的历史交互行为数目分3组进行研究:0~50、50~200、200以上。每组均有足够的用户数目,以消除随机性。对于每组的用户,为计算其平均推荐精准度,我们选取了AUC与NDCG@2两个排序指标,具体结果如图3所示。

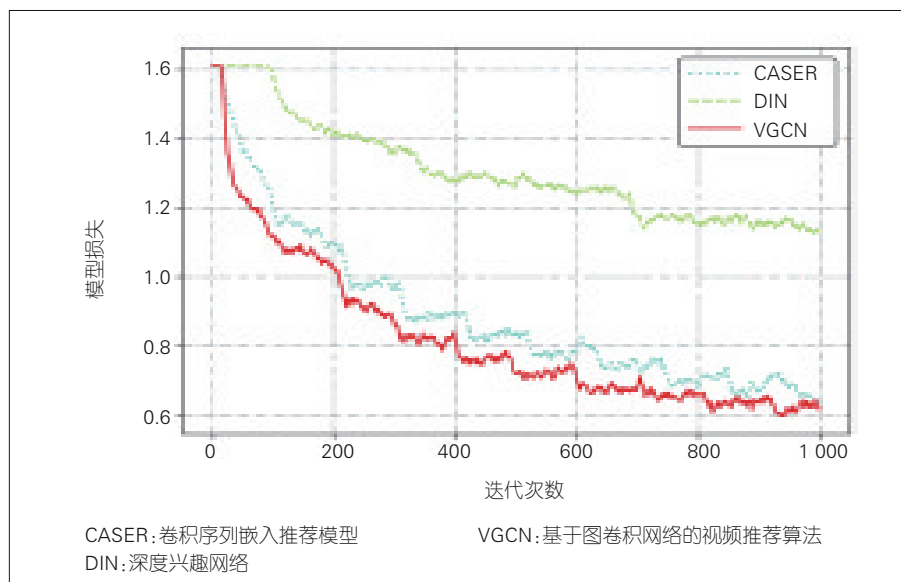
由图3可知,我们提出的VGCN方法在不同稀疏度的用户组里,均可取得有效且稳定的性能提升。这一结果进一步验证了VGCN方法的有效性。

3.4 超参数影响研究

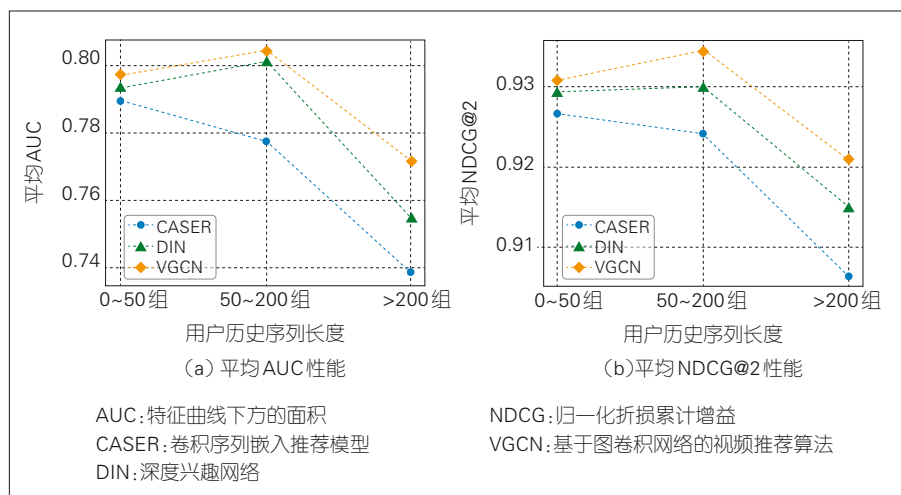
在模型训练的过程中,L2正则系数是一项重要的超参数,图4展示了不同L2正则系数对视频推荐性能的影响。根据图4可以看出,不论选择何种L2正则系数,我们提出的VGCN方法均可以取得最佳推的荐性能。此外,L2正则系数对于模型视频推荐精准度性能的影响较小,即模型对于该超参数的敏感度较低,这意味着模型不需要花费太多的调参时间与算力。

4 结束语

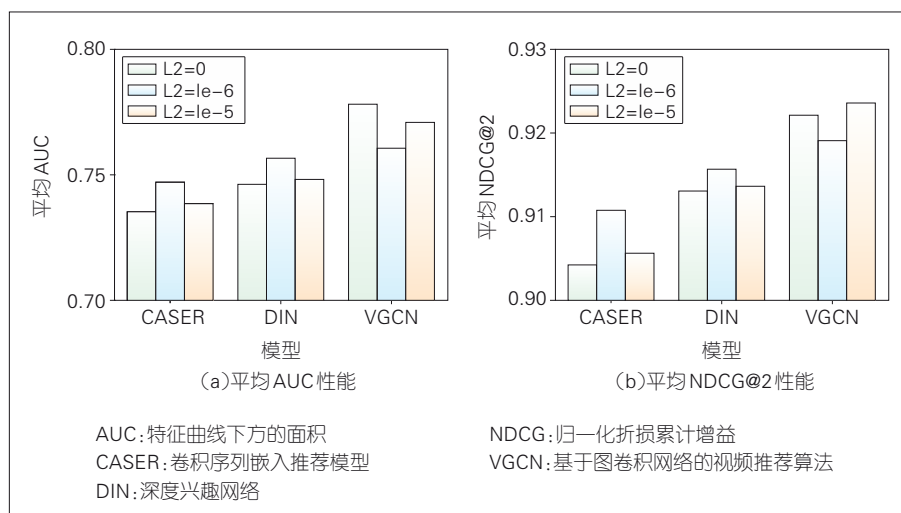
视频推荐系统是提升用户视频观看体验的重要技术。本文设计了一种基于图神经网络的推荐模型,并通过两种向量传播方式对用户长期兴趣与短时兴趣建模。基于真实数据集的实验有效验证了整体推荐精



▲图2 训练时模型损失曲线比较



▲图3 不同观看次数用户的视频推荐精准度性能比较



▲图4 不同超参数设置(L2正则系数)的视频推荐精准度性能比较

准度与不同稀疏度用户推荐精准确度的性能提升。同时,超参数影响的实验进一步验证了推荐精准确度性能提升的稳定性。

致谢

本研究得到清华大学常健新同学的帮助,谨致谢意!

参考文献

- [1] 许海玲, 吴潇, 李晓东, 等. 互联网推荐系统比较研究 [J]. 软件学报, 2009, 20(2): 350-362. DOI: 10.3724/SP.J.1001.2009.03388
- [2] LU J, WU D S, MAO M S, et al. Recommender system application developments: a survey [J]. Decision support systems, 2015, 74: 12-32. DOI: 10.1016/j.dss.2015.03.008
- [3] 朱岩, 林泽楠. 电子商务中的个性化推荐方法评述 [J]. 中国软科学, 2009(2): 183-192. DOI: 10.3969/j.issn.1002-9753.2009.02.022
- [4] RENDLE S, FREUDENTHALER C, SCHMIDT-

THIEME L. Factorizing personalized Markov chains for next-basket recommendation [C]// Proceedings of the 19th International Conference on World Wide Web. North CA, USA: ACM Press, 2010. DOI: 10.1145/1772690.1772773

- [5] TANG J X, WANG K. Personalized top-N sequential recommendation via convolutional sequence embedding [C]// Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. Los Angeles, CA, USA: ACM, 2018. DOI: 10.1145/3159652.3159656
- [6] ZHOU G R, ZHU X Q, SONG C R, et al. Deep interest network for click-through rate prediction [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. NY, USA: ACM, 2018. DOI: 10.1145/3219819.3219823
- [7] GOYAL P, FERRARA E. Graph embedding techniques, applications, and performance: a survey [J]. Knowledge-based systems, 2018, 151: 78-94. DOI: 10.1016/j.knsys.2018.03.022
- [8] 徐冰冰, 岑科廷, 黄俊杰, 等. 图卷积神经网络综述 [J]. 计算机学报, 2020, 43(5): 755-780. DOI: 10.11897/SP.J.1016.2020.00755
- [9] KIPF T N, WELING M. Semi-supervised classification with graph convolutional networks [C]// Proceedings of International Conference on Learning Representations. Toulon, France: University of Montreal, 2017
- [10] MANNING D C, SCHÜTZE H, RAGHAVAN P. Introduction to information retrieval [M]. Cambridge: Cambridge University Press,

2008

作者简介



高宸, 清华大学电子工程系在读博士研究生; 主要研究领域为用户行为建模与挖掘。



李勇, 清华大学电子工程系副教授; 主要研究领域为网络科学、城市计算、用户行为建模与挖掘。



金德鹏, 清华大学电子工程系教授; 主要研究领域为网络科学、城市计算、用户行为建模与挖掘。



脑启发视频用户体验 评测关键技术

Key Techniques of Brain Inspired Video QoE Prediction

陶晓明 /TAO Xiaoming¹, 杜冰 /DU Bing², 段一平 /DUAN Yiping¹

(1. 清华大学, 中国 北京 100084;

2. 北京科技大学, 中国 北京 100083)

(1. Tsinghua University, Beijing 100084, China;

2. University of Science and Technology Beijing, Beijing 100083, China)

摘要: 基于脑电图 (EEG) 响应趋同性的生理学机理, 研究了基于脑电图响应特征的体验质量 (QoE) 度量方法, 实现小样本稳定度量; 建立网络关键性能指标 (KPI)、服务关键质量指标 (KQI) 与 QoE 之间的映射关系模型。该模型可有效地提升多媒体服务和网络资源协同的优化空间, 为显著提升多媒体业务支持能力提供新途径。

关键词: 用户体验; 深度学习; 脑电图; 评测

Abstract: Based on the physiological mechanism of electroencephalogram (EEG) response convergence, the quality of experience (QoE) measurement method based on EEG response characteristics is studied to realize small sample stability measurement. Furthermore, the key performance indicator (KPI) and key quality indicator (KQI) are established based on this model. By this way, the optimization space of multimedia services and network resources collaboration can be effectively improved, and a new way to significantly enhance multimedia business support capabilities is provided.

Keywords: quality of experience; deep learning; electroencephalogram; evaluation

DOI: 10.12142/ZTETJ.202101008

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210204.0943.002.html>

网络出版日期: 2021-02-04

收稿日期: 2020-12-22

1 视频体验质量评测概述

无线多媒体业务的爆炸式增长已成为无线通信行业快速、持续发展的主要推动力。多媒体信息在为用户带来多类型、全方位的视听体验的同时, 也决定了用户对多媒体业务的主观评价必将受多维因素的综合影响; 因此用户的体验质量 (QoE)^[1] 才是衡量服务质量的最优指标。然而, 传统

通信技术的发展是基于通信系统中各层面客观技术指标来提升用户的服务质量 (QoS)。该发展思路只能单纯地提升一个或多个客观通信指标, 同时可能会偏离用户满意度提升这一根本目标, 从而造成通信和计算资源的浪费。可以看到, 与传统通信业务相比, 无线视频通信的功能定位已从准确、快速、及时地完成信息传输, 扩展为满足用户多样化的通信业务体验需求。未来, 无线视频通信质量的评价方法应当具备更强的主观属性和整体属性。也就是说, 无线视频通信质量评价的

结果应当能全面反映一项技术或者业务是否能够满足其用户在特定应用场景下的各方面体验需求。

面向 QoE 的视频通信成为全球学术界的研究热点。现有视频通信系统大部分采用基于 QoS 的度量方法, 即以微观精确性为前提的客观评价指标, 具体包括峰值信噪比 (PSNR)、结构相似性 (SSIM) 等^[2-3]。这个度量方法使得网络服务能力的提高单纯依赖于传输带宽的增加。A. V. MOORSEL 等学者于 2001 年率先提出 QoE 的概念, 用于度量用户在使用网络基础设施、

基金项目: 国家重点研发计划 (2019YFB1803404);
国家自然科学基金 (61925105、61801260)

享受网络服务过程中的主观感受^[4]。研究表明,人们在获取信息时,往往只关注主观上的视听体验,并不关注客观上的 QoS 通信指标,例如带宽、排队时延、时延抖动、吞吐量、丢包率、峰值信噪比等^[5]。也就是说,用户体验与网络传输能力并不构成比例关系,单纯通过增加传输带宽来提高用户体验具有局限性;而面向 QoE 的视频通信有可能从根本上减少业务对无线网络带宽需求的压力。如果以体验质量 QoE 为准则,比如关注度、清晰度、流畅度等,那么可放松对精确性的要求,获得新的优化空间,提高视频通信的效率。

目前,通信领域对于 QoE 的研究,主要集中在度量和建模^[6]。主观评价的方法是从用户的感知出发,让用户直接对所使用的业务做出评价,因此它是能够最直接、准确地反映用户体验的方法,具体包括平均意见评分(MOS)、差分平均意见评分(DMOS)、恰可识别失真(JND)等。视频视觉主观质量评价的研究者们提出了最小可觉失真(JND)^[7]的重要概念,并将其作为评判视频视觉质量是否达到最优的标准值,以对 QoE 进行度量。文献[8]将场景因素与人类因素作为模型的调节因素,提出了真实环境下基于加权函数的 QoE 估计框架。华为 mlab^[9]提出了移动视频平均意见评分(MV-MOS),对视频的体验质量进行度量。上述方法主要在于建立关键指标和用户 QoE 之间的映射关系,从而反映用户体验质量。这类度量方法主要依赖于用户行为,在一定程度上反映用户的主观感受,为视频通信提供了新的思路。通信过程中的编解码、传输等都会影响视频质量。因此,面向 QoE 的视频质量评价、面向 QoE 的视频编码以及面向 QoE 的视频质量增强是本文研究的主要内容。文章中,我们将

从 QoE 评价、QoE 编码、QoE 提升 3 个方面提升视频通信系统的效率。

面向 QoE 的视频通信面临着新的机遇和挑战。当前的视频通信 QoS 缺少面向人类感知的评价准则。现有视频通信系统的服务质量多使用 PSNR、SSIM 等 QoS 评价指标,它们以微观精确性为前提,不能反映用户感知质量。视频通信系统的最终目标是终端用户,衡量通信服务品质的根本标准是用户的 QoE^[10-11],即用户的宏观满意度。当前,基于用户的评价方法依赖于用户的行为反馈,这易受人们认知偏差的影响;基于业务的评价方法尚难以实现在实际业务场景下对业务的综合体验质量的准确评价,且难以推广应用。视频感知质量是视频 QoE 的一个客观评价指标,常用的评价方法有特征相似性、相对熵^[12]、用 Fréchet Inception 距离^[13]、Inception Score^[14]等,但它们难以直接准确地反应主观感受;因此,如何建立一套视频感知质量评价准则,为面向 QoE 的视频通信提供通用的评价标准,是必须要解决的问题。

QoE 的模型本质上是可测参数到视频 QoE 的映射,包括 QoE 的度量与评测。QoE 的度量是指 QoE 的标尺,目前不同场景、不同业务存在 QoE 度量标准不统一的问题,难以真实反映用户的 QoE。传统方法往往采用主观评分,因人而异、差异大,难以找出度量关系。这样一来,传统方法会需要大量样本,如利用 10 万左右的用户打分,才能归纳出视频质量退化与 QoE 相对稳定的度量关系。近年来的研究表明,通过测量人类视听刺激的早中期神经信号,脑电图(EEG)可以反映主观感受,并可有效排除个体的影响。QoE 评测本质上是从可测参数到用户体验的映射,包括数学模型和机器学习模型两大类。数学模型主要包括对数模型、线性回归、信息增

益、相关性分析、E 函数模型等,也是目前使用较多的方法。数学模型的主要步骤是通过采集大量的数据,包括可测参数和用户打分,建立可测参数和 QoE 的数学关系,从而建立图表、分段函数等数学映射公式,其本质是数据拟合的思想。

影响 QoE 的因素众多,因此通过数据拟合的方法,是很难获得显式的数学表达的。近几年,机器学习模型取得了突破性进展,被大量应用于计算机视觉任务中。机器学习模型主要是采用数据驱动的方法,从大量数据中挖掘有意义的规律或模式,这在一定程度上克服了线性回归和相关性分析方法的缺点。因此,将机器学习的方法用于 QoE 模型被开始大量地研究。该方法的核心思想是从大量数据中学习可测参数和视频 QoE 间的复杂映射关系。一方面,视频 QoE 可以建模为一个预测问题,即使用神经网络建立影响 QoE 的因素与 QoE 分数之间的非线性复杂映射关系,用于视频 QoE 预测,例如,基于逻辑斯蒂回归的预测模型。另一方面,视频 QoE 可以建模为一个分类问题(通常为五分类问题),可以支持向量机(SVM)和决策树等分类算法。该类算法通常难以有效处理高维数据,无法为大规模业务提供稳定评测。

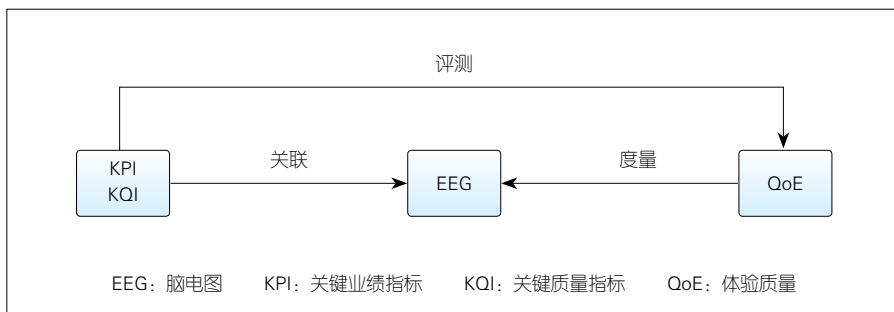
2 脑启发 QoE 评测关键技术

借助 EEG 实验手段,我们测量专业被试者观看不同播放质量视频时相对稳定的脑电响应,剖析无线网络参数、多媒体业务特性对用户体验质量的影响作用。另外,还利用多参数特征选择方法提取关键特征,量化其与用户的感知体验的相关性,从而确定影响用户体验质量的关键特征参数。利用机器学习方法,挖掘这些关键特征指标与用户主观体验之间的映射关

系,建立从可测参数到用户主观体验的预测模型。脑启发 QoE 评测主要包括 3 个方面的关键技术,如图 1 所示。

2.1 基于脑电响应特征的 QoE 度量方法

对体验质量评价方法的研究除了需要准确理解体验质量的定义外,还需要全面考虑体验质量的各项影响因素,清晰梳理体验质量的形成过程并科学地提出体验质量的预测方法。同时,体验质量评价方法的建立本质上是一项多学科交叉的研究内容。在传统信息与通信科学的基础上,生理学、心理学和社会科学等有关学科也是本领域研究的基础。EGG 系统是一种能够连续测量并记录人类头皮不同部位电位信号的设备,它通过在头皮不同部位放置弱贴合的电极,实现非侵入式的测量。对头皮电位的有效记录与分析为人类探索自身思维活动提供了一种科学、客观、可行的研究手段。脑电信号的空间分布、波形等特征也根据具体感官刺激事件的类型有着显著的模式。借助脑电图实验手段,对用户观看不同播放质量视频过程中的脑电信号进行实时测量,并通过基于事件相关电位(ERPs)和时频特征的分类算法,对用户关于视频关键性能指标的感知及认知行为进行量化表征,进而测定其感知极限。此外,从大规模用户日常移动视频业务实测数入手,将智能计算引入用户质量体验预测模型,并利用机器学习的计算工具,挖掘无线视频业务参数与用户体验中间分值之间复杂的映射关系,建立符合用户体验的评价标准。利用时域 P300 成分检测与小波变换时频分析,确定了用户对于卡顿、清晰度、启动时延等关键性能指标的感知极限;根据实测数据,计算视频业务参数和用户体验之间的斯皮尔曼相关系数,确定对用户体验有显著影响的关键网络及业



▲图 1 脑启发 QoE 评测关键技术

务指标;为建立无线视频业务参数与用户体验分值之间模型,引入深度学习方法,对关键的视频业务参数进行归一化计算,并利用深度感知器模型,训练无线视频业务参数与用户体验分值的非线性变换关系,从而获得基于客观业务参数的用户体验映射模型,实现对广大区域内多用户体验质量的在线实时预测。

2.2 QoE 关键特征选择

影响视频 QoE 的参数种类各异,数量繁多,例如,比特率、误码率(BER)、信号强度、缓冲速率、缓冲时延、重缓冲比率、重缓冲次数、视频播放时长等,可综合反映视频的清晰度、流畅度、关注度等。从大规模实测数据中,可以采集可测参数和相应的用户打分。高维可测参数对用户 QoE 的影响是有差别的。因此,通过计算可测参数和用户打分的相关系数,进行特征选择。我们把每 100 条用户对视频的打分为一组,每一条用户行为有多个可测参数和用户关于清晰度、启动延时、卡顿和视频总印象的主观打分。我们计算这 100 条用户行为的每一个特征和 4 个打分的斯皮尔曼相关性系数,找到影响 QoE 的关键因素。通过计算相关系数,得到可测参数对 QoE 的一个“加权”,之后就可以产生和用户主观打分相关性较强的特征子集。我们可以采用阈值法、双向搜索方法等提取关键特征。

2.3 基于深度学习的用户体验质量预测模型

在进行 QoE 度量之后,如何建立可测参数 KPI 和 KQI 到 QoE 的映射,仍然是需要解决的问题。综合考虑客观与主观因素,定量分析各参数与 QoE 的关系后,可以采用深度学习方法,定义适用于无线环境的视频质量评价模型。通过采集大量实测数据,获得网络参数、业务参数和用户打分,构建视频 QoE 评测数据集,并根据采集到的实际数据,采用 One-hot 进行编码。数据集可以分为训练数据集和测试数据集。首先,通过分析,确定对用户体验有显著影响的关键网络及业务指标,在数据输入深度学习模型之前,需要先对关键指标进行归一化。该深度学习模型由两部分组成:全连接层和分类预测层(使用了分类器)。根据深度学习的原理,整个模型分为两个过程:训练和预测。在训练过程中,采集的大量真实数据经过 One-hot 编码后,输入深度神经网络,学习网络业务客观指标和视频平均意见分数(V MOS)之间的非线性映射关系,并实现其在线模式。预测过程是将实时采集的数据输入到训练好的网络中,从而得到用户体验质量的预测值。

2.4 视频 QoE 评测系统

随着 4G 网络的快速发展及 5G 业务层面多样化,传统的投诉问卷调查无法满足用户需求,因此构建视频

QoE 评测系统以实时预测视频 QoE 是必须要解决的问题。该评测系统主要包括：数据采集系统、脑电感知系统、智能模型。该系统主要的输入为：初始缓冲时延、卡顿数组、卡顿量、卡顿总时长、播放时长、视频总时长、分辨率、帧率等参数；输出为：用户感知分值。通过采集用户观看视频过程中采集的参数，调用视频 EEG 标准模型接口，可反馈当前观看视频的用户感知量化分值。

数据采集是指通过数据接口采集网络数据，并在云端建立采集数据库，基于不同业务类型，提取相应字段输出给脑电感知系统和智能模型两个单元。

脑电感知系统是系统的第一核心单元。脑电感知系统具体是指基于生物特征识别技术建立脑电标准数据库，再使用输入的网络采集数据，提取相应业务的脑电数据库进行模型映射，并将评分准则校正结果输出给智能模型进行在线质量评测。

智能模型是系统的第二核心单元。智能模型具体是指基于已有算法库，根据用户应用选择相应的人工智能模型，并使用输入的网络采集数据，依托脑电感知系统输入的评分准则校正结果，进行质量评测以及其他的可扩展应用结果输出。智能模型是基于 Keras + Tensorflow 框架的，采用模块化和参数化设计，主要包括数据载入和参数输入、数据清洗、模型训练单元、模型精度评估单元，以及评测结果应用程序编程接口（API）等主要模块。

3 结束语

本文中，借助 EEG 实验手段，我们对用户观看不同播放质量的视频过程中的脑电信号进行实时测量。通过基于事件相关电位（ERPs）和时频特

征的分类算法，对用户关于视频关键性能指标的感知及认知行为进行量化表征，进而测定其感知极限。此外，我们从大规模用户日常移动视频业务实测数入手，将智能计算引入用户质量体验预测模型，并利用机器学习的计算工具，挖掘无线视频业务参数与用户体验中间分值之间复杂的映射关系，建立符合用户体验的评价标准，从而实现用户体验质量的实时评测。

参考文献

- [1] BALACHANDRAN A, SEKAR V, AKELLA A, et al. Developing a predictive model of quality of experience for Internet video [C]//Proceedings of the ACM SIGCOMM 2013 conference on SIGCOMM. New York, NY, USA: ACM, 2013: 339–350. DOI:10.1145/2486001.2486025
- [2] MINNEN D, Toderici G, COVELL M, et al. Spatially adaptive image compression using a tiled deep network [C]//2017 IEEE International Conference on Image Processing (ICIP). Beijing, China: IEEE, 2017: 2796–2800. DOI:10.1109/icip.2017.8296792
- [3] KLOPP J, WANG F Y, CHIEN S, et al. Learning a code-space predictor by exploiting intra-image dependencies [EB/OL]. [2020–12–20]. <http://www.bmva.org/bmvc/2018/contents/papers/0491.pdf>
- [4] MOORSEL A. V. Metrics for the Internet age: quality of experience and quality of business [EB/OL]. [2020–12–20]. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.24.3810>
- [5] TAO X M, DUAN Y P, XU M, et al. Learning QoE of mobile video transmission with deep neural network: a data-driven approach [J]. IEEE journal on selected areas in communications, 2019, 37(6): 1337–1348. DOI:10.1109/jsac.2019.2904359
- [6] PENG X, DUAN Y P, GENG B R, et al. A QoE-based alarm model for terminal video quality [C]//2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP). Ottawa, ON, Canada: IEEE, 2019: 1–5. DOI:10.1109/globalsip45357.2019.8969366
- [7] JAYANT N, JOHNSTON J, SAFRANEK R. Signal compression based on models of human perception [J]. Proceedings of the IEEE, 1993, 81(10): 1385–1422. DOI:10.1109/5.241504
- [8] MOLLER S, RAAKE A. Quality of experience: advanced concepts, applications and methods [M]. Germany: Springer, 2014
- [9] HUAWUEI mLab [EB/OL]. [2020–12–20]. <http://mlab.huawei.com>
- [10] CHEN Y, WU K, ZHANG Q. From QoS to QoE: a survey and tutorial on state of art, evolution and future directions of video quality analysis [J]. IEEE communications surveys and tutori-

als, 2014, 99(1): 1

- [11] GREGPR K, BESSE F, REZENDE D J, et al. Towards conceptual compression [C]//NIPS' 16: Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2016: 3556–3564
- [12] MINNEN D, Toderici G, SINGH S, et al. Image-dependent local entropy models for learned image compression [C]//2018 25th IEEE International Conference on Image Processing (ICIP). Athens, Greece: IEEE, 2018. DOI:10.1109/icip.2018.8451502
- [13] JOHNSTON N, VINCENT D, MINNEN D, et al. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 4385–4393. DOI:10.1109/cvpr.2018.00461
- [14] BALLE J, MINNEN D, SINGH S, et al. Variational image compression with a scale hyperprior [C]//6th International Conference on Learning Representations. Vancouver, BC, Canada: ICLR, 2018

作者简介



陶晓明，清华大学教授、博士生导师；主要从事无线多媒体通信理论及关键技术应用研究；曾获国家自然科学二等奖、国家技术发明二等奖、教育部科技进步一等奖等奖项，获得国家自然科学基金杰青项目资助，以及中国青年科技奖、中国青年女科学家奖等；发表论文 50 余篇，授权专利 40 余项。



杜冰，北京科技大学计算机与通信工程学院讲师；研究方向为无线多媒体通信。



段一平，清华大学电子工程系助理研究员；研究方向为无线多媒体通信。



超高清内容清晰度 用户体验质量评价

Quality of Experience Estimation of Ultra-High Definition Content

朱文瀚/ZHU Wenhan, 翟广涛/ZHAI Guangtao, 陶梅霞/TAO Meixia,
杨小康/YANG Xiaokang, 张文军/ZHANG Wenjun
(上海交通大学, 中国 上海 200240)
(Shanghai Jiao Tong University, Shanghai 200240, China)

摘要:针对多媒体行业对超高清内容清晰度用户体验评价的迫切需求,提出了一种有效的无参考质量评价算法,以预测目标内容的用户感知体验,并区分原始4K和伪4K内容。通过对目标内容进行分割,利用局部方差选择了3个代表性子块代替全局来提高计算效率。针对超高清内容的特性,提取了复杂度特征、频域特征和像素统计特征。采用支持向量回归的方法将这些提取的特征融合为一个质量指标,以预测目标内容的质量分数。实验结果表明,本模型可以有效地评估用户感知体验,并具有良好的辨别真假4K内容的能力。

关键词:用户体验质量;无参考质量评价;超高清;自由能原理;频域分析;自然图像统计

Abstract: In response to the urgent demand for assessing the quality of experience of ultra-high definition content in multimedia industries, a non-reference quality assessment model is proposed to predict the perceptual quality of the target content and distinguish pristine 4K and pseudo 4K contents. Our model segments the image and chooses three representative patches by local variances to improve computing efficiency. According to the characteristics of ultra-high definition content, complexity features, frequency domain features and pixel statistics features are extracted from the representative patches. The support vector regressor is employed to aggregate these extracted features as an overall quality metric to predict the quality score of the target image. The experimental results demonstrate that the proposed method can effectively evaluate quality of user experience and is capable of distinguishing true and pseudo 4K contents.

Keywords: quality of experience; non-reference quality assessment; ultra-high definition; free-energy principle; frequency domain analysis; natural scene statistics

DOI: 10.12142/ZTETJ.202101009

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210114.0948.002.html>

网络出版日期: 2021-01-14

收稿日期: 2020-12-16

随着数字电视与多媒体行业的高速发展,超高清内容已经成为新一代电视、电脑显示器甚至手机屏幕的流行配置。由于超高清图像和视频在改善用户体验方面有着很重要的作用,在适当的观看距离下,4K视

频有生动的细节呈现,清晰度高,能显著增强视觉体验,因此,超高清内容成了时下最火热的话题之一。2012年,国际电信联盟(ITU)发布超高清电视的国际标准——ITU-R BT.2020建议书^[1]。该标准正式规范了4K分辨率为 $3\,840 \times 2\,160$ 像素,宽高比为16:9。此后,各国制定了相应的超高清图像和视频标准,以更加规

范该行业^[2]。例如,中国国家广播电视影视总局发布了用于生产和节目交换的超高清电视系统的参数值(GY/T 307—2017)和针对超高清电视图像质量的主观评估方法(T/CSMPTE 3—2018)。在消费市场上,各种电子设备制造商以4K为卖点,宣称其数字设备支持超高清内容。许多网络视频运营商还推出了超高

基金项目:国家自然科学基金(61831015);国家重点研发计划(2019YFB1405902)

清节目源,例如,Netflix、YouTube、乐视网、优酷和百视通都有4K视频直播服务。此外,智能手机行业将其注意力转向4K,越来越多的高端智能手机可以拍摄和生成4K图像和视频为卖点。

然而,超高清行业的发展同样会带来一些问题。根据Akamai最近的统计数据,只有21%的美国家庭网速在15 Mbit/s以上,这一传输速率被认为是有效播放4K视频的最低门槛。一些调查显示,虽然中国消费了全球约80%的4K电视,但是大部分视频信号仍是高清水平。此外,为了推广4K这一新兴卖点,一些内容提供商或个人在网络上传播大量虚假4K视频。尽管这些“高端”的“4K”视频具有与自然4K内容相同的分辨率,但其往往模糊且缺乏细节,无法满足消费者的需求。这些虚假的4K视频在存储和传输过程中占用了大量的内存和带宽资源,但却无法为用户提供相应的高质量体验。因此,如何将这些伪超高清内容从真实的超高清内容中辨识出来显得尤为重要。

图像质量评价作为一种预测图像的感知质量的方法,在过去的20年中得到了广泛研究^[3]。一般而言,图像质量评价可以分为主观图像质量评价和客观图像质量评价^[4]。其中,主观质量评价被认为是判断图像感知质量的最准确方法。研究者们通过建立许多主观的图像质量数据库来提供各种质量和相应的真实质量分数的图像,以促进客观模型的发展。与主观评价相比,客观评价可以自动、高效地预测失真图像的感知质量,具有可重复性高、速度快的特点,是质量评价领域的研究重点。根据参考图像的可用信息,客观的质量评价算法通常可以分为全参考、半参考和无参考算法。其中,全参考质量评

价模型可以利用参考图像的全部信息。均方误差(MSE)、峰值信噪比(PSNR)和结构相似性算法(SSIM)^[5]是全参考领域的3种最经典的算法。半参考质量评价模型则只能使用一部分参考图像的信息,例如参考图像的个别特征值,但仍可以大大减少传输参考图像时的信息量^[6]。此外,在大多数的现实场景中,由于参考图像并不存在,无参考图像质量评价则可以发挥出作用,这是因为它不需要参考图像就可以准确地评估失真图像的感知质量。根据方法论的不同,无参考质量评价模型大致可以分为3大类:基于自然图像统计的模型^[7]、基于机器学习的模型^[8]和基于人眼视觉系统的模型^[9]。

目前,大多数图像质量评价方法都针对普通的低分辨率图像或人为制作的失真图像。与这类图像不同,超高清图像具有非常高的分辨率,而人眼很难区分真实的超高清图像和通过插值算法得到的伪4K图像。据我们所知,目前还没有专门针对这项任务而设计的算法。因此,预测超高清图像的质量、区分真伪超高清图像是一个全新的挑战。这值得我们去研究现有的无参考质量评价模型是否可以胜任此任务,同时值得我们去研究针对超高清图像质量的新算法。

1 算法设计

1.1 图像分解预处理

超高清图像的分辨率比一般的图像大很多,这会显著增加算法的计算量,造成算法运算时间过长,不利于算法的实际应用。因此,我们首先尝试将一个输入图像切成多个子图像,以获得最具代表性的一个或几个子图像来代表整个输入图像,然后在这些选定的子图像上执行后续的特

征提取,以减少算法的计算量。

在给定一个4K图像 I 的条件下,我们首先将 I 划分为 16×9 个子图像 $I_{i,j}$,其中 $i \in \{1, 2, \dots, 16\}$, $j \in \{1, 2, \dots, 9\}$ 。这使得子图像 $I_{i,j}$ 的宽度像素和高度像素均为240,在随后的计算过程中具有良好的属性。由于人类的拍摄习惯和节目拍摄技巧,最重要和最具吸引力的内容往往集中在图像的中心而不是边缘。因此,为了避免代表性的子图像出现在图像的边缘,例如带有电视台徽标、电视节目名称、字幕和人们不太关注的图像内容的子图像,我们缩小了选择范围:从左侧的第三列到右侧的第三列,以及从顶部的第二行到底部的第二行。

然后,我们依据图像复杂度的特性来选择代表性的子图像。由于具有高复杂度的子图像具有更多样化的内容,这些内容可能更具吸引力并且更加重要,因此,我们采用了局部方差作为依据。局部方差是一个可以有效反映图像结构信息、对高频信息高度敏感而又比较简单的特征。对于子图像 $I_{i,j}$ 中的像素点 (m, n) ,局部方差 $\sigma^2(m, n)$ 可以表示为:

$$\sigma^2(m, n) = \sum_{k=-K}^K \sum_{l=-L}^L \omega_{k,l} (I_{k,l}(m, n) - \mu(m, n))^2, \quad (1)$$

其中, $\omega = \{\omega_{k,l} | k = -K, \dots, K, l = -L, \dots, L\}$ 是一个二维的圆对称高斯加权函数, K 和 L 分别表示该函数窗宽和窗高的尺寸, μ 表示图像的局部均值。接下来,我们计算子图 $I_{i,j}$ 的平均局部方差,并将其作为最终的选择依据,如公式(2)所示:

$$\sigma^2(I_{i,j}) = \frac{1}{M \times N} \sum_{m=1}^M \sum_{n=1}^N \sigma^2(m, n), \quad (2)$$

其中, M 和 N 分别表示图像子块的宽度和高度。将所有子图的局部方差

进行排序后,我们选择了3个拥有最大局部方差的子图作为代表性子图 $I_{1,2,3}$,以用于后续算法的实现。

1.2 复杂度特征提取

在基于人类视觉系统建模的无参考图像质量评价研究中,很多学者研究自由能原理,并取得了良好的研究成果。自由能原理是在脑神经科学领域里被提出的,用于量化人脑的感知、行为和学习的过程^[10]。在图像处理领域中,自由能被证明可以很好地表征图像复杂度特征,并且和图像质量高度相关^[9]。因此,本文中,我们尝试使用自由能原理模型来模拟人脑预测图像的过程,并提取图像复杂度特征。

基于自由能的大脑原理的一个基本前提是,认知过程受人脑内部生成模型的控制。当人的大脑收到一个“惊喜”时,大脑会在其内部生成模型,主动预测有意义的信息并消除残留的不确定性,以生成一个预测结果,来解释大脑的感知。

我们假设用于视觉感知的内部生成模型 G 是参数化的,它通过调整参数向量 θ 来解释视觉场景。具体来说,在给定一个输入图像 I 的条件下,它对大脑产生的“惊喜”,可以通过在模型参数向量 θ 空间上的联合分布 $P(I, \theta|G)$ 来表示:

$$-\log P(I|G) = -\log P(I, \theta|G) d\theta. \quad (3)$$

为了使公式(3)便于理解,我们引入了一个辅助项 $Q(\theta|I, G)$ 分别加入分子和分母中,可以得到公式(4):

$$-\log P(I|G) = -\log Q(\theta|I, G) \frac{P(I, \theta|G)}{Q(\theta|I, G)} d\theta, \quad (4)$$

其中, $Q(\theta|I, G)$ 是这个图像的参数模型的辅助后验分布。该分布可以认

为是人脑计算的模型参数 $P(\theta|I, G)$ 的真实后验的近似。大脑将通过感知图像,调整 $Q(\theta|I, G)$ 的参数 θ ,来最小化近似后验 $Q(\theta|I, G)$ 与真实后验 $P(\theta|I, G)$ 之间的差距,用于更好地解释这个输入信号。为了简化表达,我们在后续分析中删除了生成模型符号 G 。利用詹森不等式,我们可以得到公式(5):

$$-\log P(I) \leq -\int Q(\theta|I) \frac{P(I, \theta)}{Q(\theta|I)} d\theta. \quad (5)$$

然后,基于统计物理和热力学,我们定义公式(5)的右边部分为自由能:

$$F(\theta) = -\int Q(\theta|I) \frac{P(I, \theta)}{Q(\theta|I)} d\theta. \quad (6)$$

可以看出,由于 $-\log P(I) \leq F(\theta)$,因此, $F(\theta)$ 定义了一个图像“惊喜”的上确界。注意到 $P(I, \theta) = P(\theta|I)P(I)$,我们可以继续将公式(6)转化为公式(7):

$$\begin{aligned} F(\theta) &= \int Q(\theta|I) \frac{Q(\theta|I)}{P(\theta|I)P(I)} d\theta = \\ &= -\log P(I) + \int Q(\theta|I) \frac{Q(\theta|I)}{P(\theta|I)} d\theta = \\ &= -\log P(I) + \text{KL}(Q(\theta|I) \| P(\theta|I)), \end{aligned} \quad (7)$$

其中, $\text{KL}(\cdot)$ 表示近似后验分布和真实后验分布之间的 Kullback-Leibler 散度。

本文中,我们利用稀疏表示的方法来近似人脑的内部生成模型^[11]。具体来说,我们利用一个提取算子 $O_s(\cdot)$ 选取输入图像的一个图像块 $x_s \in \mathbb{R}^B$,其中 B 是图像块的尺寸。那么基于一个完备字典 $Y \in \mathbb{R}^{B \times U}$,图像块 x_s 的稀疏表示是通过计算一个向量 $\alpha_s \in \mathbb{R}^U$ 来表示的,如公式(8)所示:

$$\alpha_s^* = \arg \min_{\alpha_s} \frac{1}{2} \|x_s - Y\alpha_s\|_2 + \lambda \|\alpha_s\|_p, \quad (8)$$

其中, α_s 是提取的图像块的表示系数向量, U 表示稀疏表示模型里的原子的个数。 λ 是一个正常数,用于平衡重建保真度的权重约束条件和稀疏惩罚条件。通过计算每一个图像块的表示系数向量,整个输入图像 I 的稀疏表示可以表示为:

$$\hat{I} = \sum_{s=1}^{n_p} O_s^T(Y\alpha_s^*)/O_s^T(1_B), \quad (9)$$

其中, \hat{I} 表示输入图像的稀疏表示,可以被认作人脑对于输入图像的理解。“/”表示两个矩阵按元素对相除, n_p 为图像块的数量, $O_s^T(\cdot)$ 表示 $O_s(\cdot)$ 的转置, 1_B 表示尺寸为 B 的全1矩阵。

根据自由能理论,自由能表示输入图像与人脑内部生成模型得到的最佳预测图像之间的差异。因此,我们可以将自由能定义为输出图像的预测残差的熵。预测残差可以表示为:

$$RE = |I - \hat{I}|. \quad (10)$$

而其熵可以表示为:

$$H = -\sum_{i=0}^{255} p_i \log_2 p_i, \quad (11)$$

其中, p_i 表示预测残差的第 i 个灰度级的概率密度。最终我们可以得到输出图像的自由能的值。本文中,我们采用自由能作为图像复杂度特征。

1.3 频域特征提取

通常,插值方法会平滑目标图像,造成图像中低频信息增加,高频信息减少。因此,频域特征在这项任务中很有效果。本文中,我们采用离散余弦变换(DCT)获得超高清图像的频域特性。在给定一个输入图像 I 的情况下,我们可以获得其DCT系数

D 。由于数据在直角坐标系是二维的,因此,我们将直角坐标系转换到极坐标系中,降低维度,以便于后续的处理。随后,我们计算图像在对数尺度下沿极半径方向的频谱能量 E_ρ ,如公式(12)所示:

$$E_\rho = \log(1 + \sum_p |D|^2 / f^2), \quad (12)$$

其中, ρ 表示极半径方向, D 表示图像 I 的DCT系数。 $f = f_s / N$,其中 f_s 表示采样率。

通过大量的实验,我们发现了真伪4K图像能量谱和累积能量谱上的特征。图1给出了一对真伪4K图像标准化后的能量谱和累积能量谱的示意图。在图1(a)中,黑色曲线 P_1 表示真4K图像,红色曲线 P_2 表示伪4K图像,它们都是从低分辨(例如2K、1080p、720p等)的图像上采样得到的。蓝色实线 P_3 是一条辅助线,经过点 P_1 与 P_2 的交点 P 。 p_x 和 p_y 分别为交点 P 的横坐标和纵坐标。蓝色虚线 P_4 表示一个辅助图像,在整个频率上具有相同的能量,且能量高于或低于 p_y 。图1(b)中, E_i 为 P_i 的累积能量谱($i = 1, 2, 3, 4$)。由于是标准化后的累积能量谱, P_3 和 P_4 为相同斜率的一条过原点的线段。

由频域能量谱与累积能量谱的关系可知:

$$E_i(\omega) = \int_0^\omega P_i(x) dx, \quad (13)$$

$$P_i(\omega) = \frac{\partial E_i(\omega)}{\partial \omega}. \quad (14)$$

通过大量的实验统计,我们对原始分辨率为4K的图像,以及从2K、1080p、720p 3种分辨率插值得到的伪4K图像的累积能量谱进行了拟合,发现在这4种情况下,它们的特性均近似满足: $E_i(\omega) \approx a_i \omega^{b_i}$ 。

由于 E_i 的二阶导数小于零,所以它们都是凹函数。因此,我们可以发现曲线上的单点具有和 E_4 相同的斜率,如黑色曲线上的点 b ,红色曲线上的点 e 。绿色的虚线是与蓝线平行的辅助线。点 b 和点 e 分别为累积能量谱曲线 E_1 和 E_2 与绿色虚线的交点,如图1(b)所示。这些单点在 E_i ($i = 1, 2$)和 E_4 之间的最大距离记为 L_1 和 L_2 。我们利用这些距离作为算法的频域特征。

实际上, E_i ($i = 1, 2, 3, 4$)是图像标准化后的累积能量谱,它们都从原点开始,到点 $(\omega_m, 1)$ 结束,其中 ω_m 是最大频率。由于不同真假4K图像的特征各不相同,因此其频域能谱和辅助图像的频域能量谱 P_4 位置关系可以分为3种情况:(1) P_3 通过 P_1 和 P_2 的交点;(2) P_4 的值大于 p_y ,如图1(a)

所示;(3) P_4 的值小于 p_y 。下面我们将分3种情况讨论所提出特征的有效性。

(1)考虑 $P_4 = p_y$ 的情况,我们有: $P_1 < P_2, (\omega \leq p_x)$ 。同时, E_i ($i = 1, 2, 3$)在 $\omega = p_x$ 处有着相同的斜率,这意味着 E_1 和 E_2 在同一水平坐标中分别拥有单个点,如图1(b)中所示的点 b 和点 a 。因此我们可以推导出公式(15):

$$\int_0^{p_x} P_1 d\omega < \int_0^{p_x} P_2 d\omega, (\omega \leq p_x). \quad (15)$$

在图1(b)中,公式(15)表明线段 ad 的长度大于 bd ,等价于 $ac > bc$ 。根据相似三角形原理,我们有 $L_1 < L_3$ 。即真4K图像的所提特征小于伪4K图像。

(2)考虑 $P_4 > p_y$ 的情况,当 $P_1 = P_2 = P_4$ 时,我们有 $\omega_1 < \omega_2$,其中 ω_1 和 ω_2 分别是 P_1 和 P_4 、 P_2 和 P_4 的两个交点,则有:

$$\omega_a = \omega_b = \omega_d = \omega_1 < \omega_2 = \omega_e = \omega_g. \quad (16)$$

同时,由于 $P_2 \geq P_4, \omega \in [\omega_1, \omega_2]$,我们可以得到: $y_f - y_c < y_e - y_a$,其中 y_i 表示点 i 的纵坐标。因此,线段 ac 的长度大于 ef 。根据相似三角形原则,我们有 $L_3 < L_2$ 。结合情况(1),我们可以得到 $L_1 < L_3 < L_2$,即真4K图像的所提特征小于伪4K图像。

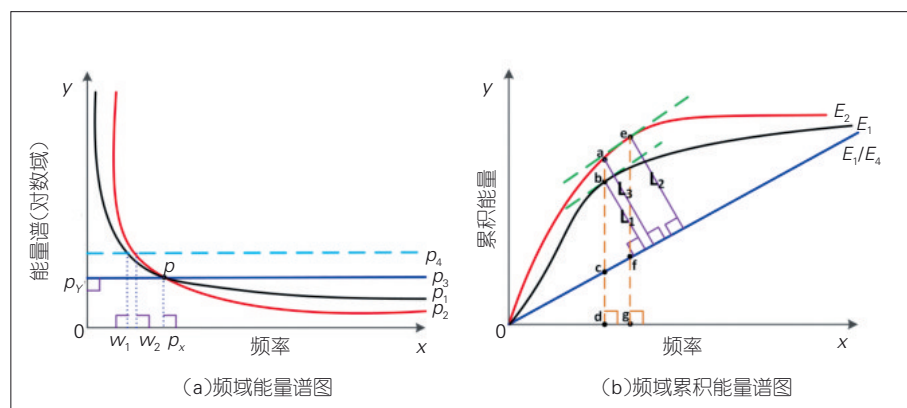
(3)考虑 $P_4 < p_y$ 的情况,由于 $P_1 > P_2$,当 $\omega > p_x$ 时,我们有:

$$\int_{p_x}^{\omega_1} P_1(x) dx > \int_{p_x}^{\omega_2} P_2(x) dx, \quad (17)$$

可以进一步得到:

$$\int_0^{\omega_1} P_1(x) dx > \int_0^{\omega_2} P_2(x) dx. \quad (18)$$

公式(18)说明 $L_1 > L_2$,即真4K图像的所提特征小于伪4K图像。



▲图1 一对真伪超高清图像标准化的频域能量谱和累积能量谱示意图

综上所述,我们所提取的频域成分的特征,可以有效地描述4K图像的真假,敏感于超高清图像的质量。因此,在本文中,我们将其定义为本算法的频域特征。

1.4 像素统计特征提取

作为一种对图像质量很敏感的信息,自然图像统计特征在图像质量评价领域被广泛应用。因此,本算法在像素层面上,也考虑了统计信息特征来提升算法的性能。我们使用了局部的均值去除对比度归一化方法来表征超高清图像的质量变化。

具体来说,对于一张给定的输入图像 I ,我们将其转化为灰度图 Z ,利用公式(19)获得其均值去除对比度归一化系数:

$$\hat{Z}(x, y) = \frac{Z(x, y) - \mu(x, y)}{\sigma(x, y) - 1}, \quad (19)$$

其中, $\hat{Z}(x, y)$ 为坐标点 (x, y) 处的均值去除对比度归一化系数。 $\mu(x, y)$ 和 $\sigma(x, y)$ 分别表示在坐标点 (x, y) 处的局部均值和局部标准差,可以通过公式(20)和公式(21)来计算:

$$\mu(x, y) = \sum_{k=-K}^K \sum_{l=-L}^L \omega_{k,l} Z(x+k, y+l), \quad (20)$$

$$\sigma(x, y) =$$

$$\sqrt{\sum_{k=-K}^K \sum_{l=-L}^L \omega_{k,l} (Z(x+k, y+l) - \mu(x, y))^2}, \quad (21)$$

其中,与1.1节相似, $\omega_{k,l}$ 是一个二维的圆对称高斯加权窗函数。在以前的研究中我们观察到,对于自然图像,均值去除对比度归一化系数值与单位正态高斯特征具有很强的相关性^[12]。我们尝试利用该系数的属性,来判断真实4K图像和从不同原始分辨率插值以及不同插值算法得到的伪4K图像。为了证明在此任务中使用均值去除对比度归一化系数的效果,我们选择了一个真实4K图像,如图2(a)所示。接着,我们分别计算了该图像与其对应的具有不同的原始分辨率和不同插值算法的伪4K版本的均值去除对比度归一化系数值,如图2(b)所示。

可以看出,真实4K图像及其不同4K版本的分布是可区分的。真实4K图像的分布显示出类似高斯的外观,而其他不同的伪4K版本则以自己的方式偏离了这种特性。这表明该系数的属性在区分真实4K图像和伪4K图像中起着积极的作用。为了数学化描述均值去除对比度归一化系数的分布,我们采用广义高斯分布来有效描述真实4K图像和伪4K图

像统计数据谱。零均值的广义高斯分布可定义为:

$$G(x; \alpha, \sigma^2) = \frac{\alpha}{2\beta(1/\alpha)} \exp(-(\frac{|x|}{\beta})^\alpha), \quad (22)$$

其中, $\beta = \sigma \sqrt{\Gamma(1/\alpha)/\Gamma(3/\alpha)}$ 。伽马方程可以表示为公式(23):

$$\Gamma(\varphi) = \int_0^\infty \varphi^{\varphi-1} e^{-\varphi} d\varphi, \quad \varphi > 0, \quad (23)$$

其中, α 和 σ 是两个参数,分别改变此高斯分布的幅度和方差。我们采用该广义高斯分布模型,近似超高清图像的均值去除对比度归一化系数分布。 α 和 σ 这两个参数被提取作为本算法像素统计特征。由于多尺度处理有助于改善质量评价模型的预测分数与人类感知之间的相关性,我们从两个尺度上提取特征,包括原始比例和降低两倍分辨率的图像。

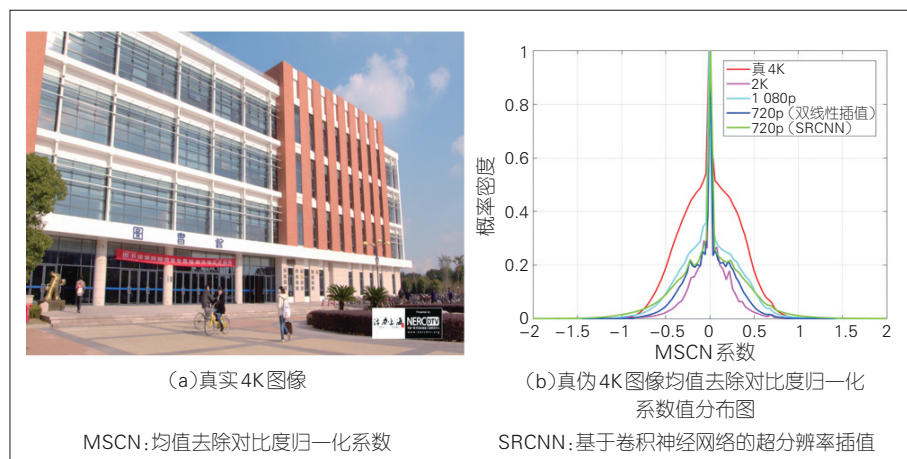
1.5 特征融合和模型表示

为了聚合上述提取的与超高清图像质量相关的特征,并生成质量评价模型以预测目标图像的质量分数,在综合考虑了回归器的有效性和模型的计算速度后,我们利用支持向量回归(SVR)方法聚合提出的特征,并采用LIBSVM软件包来学习有径向基函数(RBF)内核的模型^[13]。

2 实验过程和分析

2.1 实验数据构成

为了测试算法的有效性,我们首先构建了真伪超高清图像的数据库,并从几个现有的超高清视频序列库中获得了50多段视频序列。然后,我们从这些具有不同图像内容的视频序列中提取总共350张真实4K图像,得到了真实4K内容数据集。这些素材内容非常广泛,包括室外场景、室内场景、建筑物、角色、动物、静



▲图2 均值去除对比度归一化系数的效果测试

物、夜景、运动场景、电影和电视剧片段。接着,我们将真实的4K图像下采样为具有2K、1080p和720p 3种分辨率的图像。接着,我们通过14种不同的插值方法将它们都上采样到4K分辨率。总共有2802个伪4K图像构成了伪4K内容数据集。

2.2 实验方案

基于上述所构建的数据库,我们将提出的算法与现有的客观质量评价模型进行了比较。由于本项任务没有用于确定4K图像真实性的参考图像,因此我们仅选择无参考质量评价算法来衡量性能。我们采用了11种最新的无参考质量评价模型与我们提出的模型进行比较以预测准确性。这11种模型分别是NIQE^[7]、QAC^[14]、IL-NIQE^[15]、LPSI^[16]、HOSA^[8]、BRISQUE^[17]、BPRI^[18]、BMPRI^[19]、NFEDM^[9]、CPBD^[20]和GMLF^[21](NIQE、QAC……GMLF是指不同的无参考质量评价模型)。

根据质量评价领域的传统评估方法,我们使用4个通用评估标准来衡量所有比较的无参考质量评价模型的性能,它们分别是斯皮尔曼等级相关系数(SRCC)、肯德尔等级相关系数(KRCC)、皮尔逊线性相关系数(PLCC)和均方根误差(RMSE)。此外,我们还计算了3个准确性指标:精确率(Precision)、召回率(Recall)和准确率(Accuracy),以比较算法的性能和判断4K图像的真实性。

为了对所提出的模型进行训练,我们将测试材料随机分为两组:训练集和测试集,它们分别包含80%和20%的图像。我们使用训练集训练提出的模型,并使用测试集测试其性能。为了保证模型的鲁棒性,我们将此过程重复了1000次。这1000次重复的中值结果被认为是最终性能。

2.3 实验结果和分析

表1给出了所有算法的性能结果。其中,Precision_T和Precision_F分别表示真4K图像和伪4K图像素材组的精确率,而Recall_T和Recall_F分别表示真4K图像和伪4K图像素材组的召回率。由表1可知,在传统指标中,与传统图像质量评价数据库中的性能结果相比,所有算法的性能均不算出色。例如,这些指标中SRCC和PLCC值均不超过0.9,而通常这些指标在传统的质量评价数据库上会超过0.9。造成这种现象的主要原因是真实的4K图像与其对应的伪4K图像之间的差距很小,肉眼难以分辨。对于传统的人为失真来说,这项任务中的差异微乎其微,甚至很多伪4K图像的质量都要优于传统质量评价数据库里的参考图像。从结果上看,我们算法的性能明显优于其他主流的无参考质量评价模型。我们提出的方法的SRCC值超过0.8,

PLCC值接近0.85,而其他算法的SRCC值大都低于0.7,PLCC值低于0.8。

通过分析分类算法中常用的指标精确率、召回率和准确率的结果,我们还可以得出这样的结论:每个模型都具有较强的判断能力,而伪4K图像的判断准确度要优于真4K图像。此外,我们提出的算法具有最佳的性能,综合判断精度超过97%。因此,我们的算法具有优秀的区分真实和伪4K图像的能力,并且这种能力与主观感知分数呈正相关关系。

3 结束语

本文中,我们设计了一种新的无参考质量评价模型来评价超高清内容清晰度的用户体验质量。基于超高清内容的特性,我们在目标内容上分别提取复杂度特征、频率特征和像素统计特征,采用具有最高局部方差的3个子图代替完整的目标图像以改

▼表1 提出算法和对比算法的性能结果

指标	BPRI ^[18]	BMPRI ^[19]	BRISQUE ^[17]	CPBD ^[20]	NFERM ^[9]	GMLF ^[21]
SRCC	0.3506	0.3594	0.6651	0.5963	0.6708	0.2387
PLCC	0.5614	0.6534	0.6696	0.6194	0.6662	0.2376
KRCC	0.2956	0.2308	0.5061	0.4315	0.4990	0.1594
RMSE	13.5928	12.4345	12.2003	12.8950	12.2497	14.2550
Precision_T	0.9097	0.5890	0.8795	0.5522	0.9653	0.3433
Precision_F	0.9727	0.9441	0.9477	0.9349	0.9299	0.9196
Recall_T	0.7771	0.5486	0.5629	0.4686	0.3971	0.3600
Recall_F	0.9904	0.9522	0.9904	0.9525	0.9982	0.9140
准确率	0.9667	0.9074	0.9429	0.8988	0.9315	0.8525
指标	HOSA ^[8]	NIQE ^[7]	IL-NIQE ^[15]	LPSI ^[16]	QAC ^[14]	(pro.)
SRCC	0.7153	0.5223	0.3819	0.5782	0.6866	0.8136
PLCC	0.7296	0.5691	0.3437	0.7629	0.6427	0.8447
KRCC	0.5299	0.3797	0.2593	0.5051	0.5204	0.6472
RMSE	11.4445	13.5061	15.4249	10.6193	12.5836	7.9403
Precision_T	0.6613	0.7550	0.1469	0.7188	0.4868	0.9748
Precision_F	0.9496	0.9442	0.9081	0.9942	0.8920	0.9807
Recall_T	0.5914	0.5371	0.4600	0.7886	0.2114	0.9138
Recall_F	0.9622	0.9782	0.6663	0.9936	0.9722	0.9914
准确率	0.9210	0.9293	0.6434	0.9708	0.8877	0.9781

KRCC:肯德尔等级相关系数
PLCC:皮尔逊线性相关系数

RMSE:均方根误差
SRCC:斯皮尔曼等级相关系数

(pro.):本文提出的算法

善计算效率。支持向量回归的方法被用于回归这些特征到一个整体质量指标上。实验表明,在预测超高清内容清晰度的用户体验质量方面,本方法优于其他最新的无参考质量评价模型,并且具有良好的区分原始和伪超高清图像的能力。本算法的研究将会对超高清内容清晰度用户体验评估领域的发展起到积极的促进作用。

参考文献

- [1] ITU. Parameter values for ultra-high definition television systems for production and international programme exchange: ITU-R BT.2020 [S]. 2012
- [2] SUGAWARA M, CHOI S Y, WOOD D. Ultra-high-definition television (rec. ITU-R BT.2020): a generational leap in the evolution of television standards in a nutshell [J]. IEEE signal processing magazine, 2014, 31(3): 170-174. DOI: 10.1109/msp.2014.2302331
- [3] ZHAI G T, MIN X K. Perceptual image quality assessment: a survey [J]. Science China information sciences, 2020, 63(11): 211301. DOI: 10.1007/s11432-019-2757-1
- [4] ZHU W H, ZHAI G T, MIN X K, et al. Multi-channel decomposition in tandem with free-energy principle for reduced-reference image quality assessment [J]. IEEE transactions on multimedia, 2019, 21(9): 2334-2346. DOI: 10.1109/tmm.2019.2902484
- [5] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE transactions on image processing, 2004, 13(4): 600-612. DOI: 10.1109/tip.2003.819861
- [6] SOUNDARARAJAN R, BOVIK A C. RRED indices: reduced reference entropic differencing for image quality assessment [J]. IEEE transactions on image processing, 2012, 21(2): 517-526. DOI: 10.1109/tip.2011.2166082
- [7] MITTAL A, SOUNDARARAJAN R, BOVIK A C. Making a "completely blind" image quality analyzer [J]. IEEE signal processing letters, 2013, 20(3): 209-212. DOI: 10.1109/lsp.2012.2227726
- [8] XU J T, YE P, LI Q H, et al. Blind image quality assessment based on high order statistics aggregation [J]. IEEE transactions on image processing, 2016, 25(9): 4444-4457. DOI: 10.1109/tip.2016.2585880
- [9] ZHAI G, WU X, YANG X, et al. A psychovisual quality metric in free-energy principle [J]. IEEE transactions on image processing, 2012, 21(1): 41-52. DOI: 10.1109/tip.2011.2161092
- [10] KARL F. The free-energy principle: a unified brain theory? [J]. Nature reviews neuroscience, 2010, 11(2): 127-138. DOI: 10.1038/nrn2787
- [11] LIU Y T, ZHAI G T, GU K, et al. Reduced-reference image quality assessment in free-energy principle and sparse representation [J]. IEEE transactions on multimedia, 2018, 20(2): 379-391. DOI: 10.1109/tmm.2017.2729020
- [12] RUDERMAN D L. The statistics of natural images [J]. Network: computation in neural systems, 1994, 5(4): 517-548. DOI: 10.1088/0954-888X_5_4_006
- [13] SCHÖLKOPF B, SMOLA A J, WILLIAMSON R C, et al. New support vector algorithms [J]. Neural computation, 2000, 12(5): 1207-1245. DOI: 10.1162/089976600300015565
- [14] XUE W F, ZHANG L, MOU X Q. Learning without human scores for blind image quality assessment[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA: IEEE, 2013: 995-1002. DOI: 10.1109/cvpr.2013.133
- [15] LIN ZHANG, LEI ZHANG, BOVIK A C. A feature-enriched completely blind image quality evaluator [J]. IEEE transactions on image processing, 2015, 24(8): 2579-2591. DOI: 10.1109/tip.2015.2426416
- [16] WU Q B, WANG Z, LI H L. A highly efficient method for blind image quality assessment [C]//2015 IEEE International Conference on Image Processing (ICIP). Quebec City, QC, Canada: IEEE, 2015: 339-343. DOI: 10.1109/icip.2015.7350816
- [17] MITTAL A, MOORTHY A K, BOVIK A C. No-reference image quality assessment in the spatial domain [J]. IEEE transactions on image processing, 2012, 21(12): 4695-4708. DOI: 10.1109/tip.2012.2214050
- [18] MIN X K, GU K, ZHAI G T, et al. Blind quality assessment based on pseudo-reference image [J]. IEEE transactions on multimedia, 2018, 20(8): 2049-2062. DOI: 10.1109/tmm.2017.2788206
- [19] MIN X K, ZHAI G T, GU K, et al. Blind image quality estimation via distortion aggravation [J]. IEEE transactions on broadcasting, 2018, 64(2): 508-517. DOI: 10.1109/tbc.2018.2816783
- [20] NARVEKAR N D, KARAM L J. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection [C]//2009 International Workshop on Quality of Multimedia Experience. San Diego, CA, USA: IEEE, 2009: 87-91. DOI: 10.1109/qomex.2009.5246972
- [21] XUE W, MOU X, ZHANG L, et al. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features [J]. IEEE transactions on image processing, 2014, 23(11): 4850-4862. DOI: 10.1109/tip.2014.2355716

作者简介



朱文瀚, 上海交通大学电子信息与电气工程学院、人工智能研究院、教育部人工智能重点实验室在读博士研究生; 研究领域包含图像感知质量评价、图像视频信号处理; 发表国际学术论文17篇。



翟广涛, 上海交通大学电子信息与电气工程学院院长助理、教授、博士生导师, 《Displays》主编, 《中国科学: 信息科学》编委, IEEE电路与系统分会视觉信号处理与通信技术委员会(CAS VSPC)成员、多媒体系统及应用技术委员会(MSA)成员, 中国电子学会青年科学家俱乐部副主席, 上海市图像图形学会副理事长; 研究方向为多媒体信号处理等; 发表国际期刊论文100余篇。



陶梅霞, 上海交通大学电子信息与电气工程学院教授、博士生导师, IEEE Fellow, 中国电子学会信息论分会副主任委员, 曾任《IEEE Transactions on Wireless Communications》《IEEE Transactions on Communications》《IEEE Journal of Selected Areas in Communications》等期刊的编委或客座编委; 获2019年IEEE通信学会马可尼论文奖、2013年IEEE通信学会海因里希兹论文奖; 主要从事无线通信与网络基础研究, 包括无线缓存、边缘计算及5G关键技术等; 发表国际期刊论文80余篇、国际会议论文100余篇。



杨小康, 上海交通大学人工智能研究院常务副院长、人工智能教育部重点实验室主任、教育部“长江学者”特聘教授、国家杰出青年科学基金获得者、国家“万人计划”创新领军人才、IEEE Fellow, 《IEEE Transactions on Multimedia》《IEEE Signal Processing Letters》编委; 研究领域为图像处理与机器学习; 主持国家重点研发专项、“973”项目、国家自然科学基金项目等10余项, 获国家科技进步二等奖、中国电子学会自然科学一等奖、上海市科技进步一等奖等多个奖项; 发表国际学术论文200余篇, 申请发明专利50余项。



张文军, 上海交通大学教授、教育部“长江学者”特聘教授、国家杰出青年科学基金获得者、“973”项目首席科学家、国家自然科学基金委创新群体学术带头人、IEEE Fellow, 曾任国家高清晰度电视功能样机系统研发项目总体组组长、数字电视国家工程研究中心首席科学家、教育部未来媒体网络协同创新中心主任, 国际未来广播电视合作研究计划技术委员会主席; 主要从事图像通信与数字电视、宽带无线传输、系统芯片设计等工作, 获国家科技进步二等奖(2项)、何梁何利基金科学与技术进步奖、上海市科技进步一等奖(4项)、上海市科技功臣奖。

交互式视频质量 评价方法研究进展

A Review of Interactive Video Quality Assessment Methods



李继龙 /LI Jilong¹, 赵雪 /ZHAO Xue², 杨铀 /YANG You³

(1. 国家广播电视总局广播电视科学研究院, 中国 北京 100086;

2. 武汉理工大学, 中国 武汉 430070;

3. 华中科技大学, 中国 武汉 430074)

(1. Academy of Broadcasting Science, National Radio and Television Administration, Beijing 100086, China;

2. Wuhan University of Technology, Wuhan 430070, China;

3. Huazhong University of Science and Technology, Wuhan 430074, China)

摘要: 在交互式视频应用快速发展的同时, 如何评价视频质量成为当前亟待解决的挑战性难题, 其成果对整个多媒体通信系统的各环节技术发展具有关键作用。从主观质量评价、客观质量评价两个角度综述了当前交互式视频质量评价的研究与应用现状, 其中主观质量评价方法包括主观视频质量评价数据库、主观视频质量评价打分与计算机制, 客观质量评价方法则包括视觉信号处理与分析、深度学习机制下的评价与建模方法等。在总结上述研究方法与成果的基础上, 展望了本领域的研究与发展。

关键词: 交互式视频; 视频质量评价; 主观质量评价; 客观质量评价

Abstract: With the rapid development of interactive video applications, the research on interactive video quality assessment becomes an urgent challenge to the community, because it is helpful to the development of other modules in the multimedia communication system. The researches on interactive video quality assessment via both objective and subjective methodologies are surveyed. Extant methods are then reviewed, including databases of subjective video quality assessment, score and computation mechanism of subjective video quality assessment, visual signal processing and analysis of objective video quality assessment, and deep learning based methods of objective video quality assessment. Based on the above surveys, future directions and open problems on the research of interactive video quality assessment are discussed.

Keywords: interactive video; video quality assessment; objective quality assessment; subjective quality assessment

DOI: 10.12142/ZTETJ.202101010

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210114.1456.004.html>

网络出版日期: 2021-01-15

收稿日期: 2020-12-21

在视频通信系统中, 视频源与信道之间如同水源与水管的关系。如何克服信源与信道之间的各种矛盾, 从而给用户提供更加优质的视频服务, 一直以来都是业界追求的目标。然而,

自电视机诞生之后的 100 多年, 视频服务一直是被动接受的模式, 其发展变化无非只是从不同地点的同一时刻接受同一服务(广播电视), 变成在不同地点的不同时刻接受同一服务(基于互联网协议的互联网视频)。人们在观看视频的过程中始终无法主动改变正在播出的视频内容, 这使得该领域的研究与应用给人们所提供的想象

空间非常有限。近些年来, 在线视频服务开始从被动式向主动式转变, 出现了云端虚拟现实(VR)、云游戏等 Cloud VR 业务及以面向在线教育、在线会议的多视点视频业务等, 用户可以在终端通过“人-机-内容”交互的方式主动改变所看到的视频内容。在线视频服务有望在可预见的未来实现“千人千面”的特点。为此, 与这

基金项目: 国家重点研发计划(2017YFC08062 02); 国家自然科学基金(61971203); 武汉市应用基础前沿(2019020701011422、2020020601012222)

类视频有关的质量评价问题开始涌现,成为传统视频质量评价研究领域中的新方向、新课题。

视频的质量评价主要面向终端用户,因此该环节位于多媒体通信系统的最末端,其目的在于为多媒体通信系统前端的采集、处理、编码等环节提供一个可供参考的评价依据,从而构成处理流程上的闭环。视频质量评价的研究对象总体而言可分为两个层面:一个是解决信道质量与显示质量之间的关系,主要考察的是用于描述信道质量的多种因素与视频重建客观质量之间的作用机制,一般称为关键质量指标(KQI);另一个是解决显示质量与用户感受之间的关系,主要考察的是用于描述图像重建质量的多种因素与人类视觉系统响应质量之间的作用机制,一般称为视频质量评价(VQA)。相比而言,由于涉及信道质量,因此关注通信终端应用的学者与企业比较重视对KQI的研究;而视频的信息失真与质量重建更多的是由有损压缩或视频处理的环节所带来的,因此涉及上述领域的学者和企业比较重视对VQA的研究。Cloud VR及多视点视频业务作为产业界中的新生事物,目前在KQI方面的研究较少,尚不构成体系;但是这两种视频形式在学术界的研究中已经历过较长的历程,因此在VQA方面的成果已具有一定规模,本文的工作也主要集中于此。

1 交互式视频质量的主观评价

如前所述,视频的最终接收者是用戶,因此视频质量的好坏理应由人来决定。然而,终端用户因个人知识背景、观看环境,甚至观看时的情绪千差万别,其对视频质量优劣的反应也会各不相同,因此如何对视频质量进行有效的评价是一个极具挑战性的难题^[1]。一般情况下,其研究可

分为主观、客观质量评价两个大的方向。视频主观质量评价采用“自顶向下”“以人为本”的研究模式,探索涉及人本体相关的因素与视频质量之间的联系;客观质量评价采用“自底向上”“以技术为本”的研究模式,探索和构建视频中的视觉信号与视频质量之间的映射关系。两种模式互为支撑,不可相互替代。

视频主观质量评价从技术手段上可通过邀请主观测试人员采取某种规定的打分机制,对具有不同失真类型、等级的视频进行打分,这涉及主观质量评价数据库、主观质量评价打分和计算机制等相关工作。在打分与计算机制方面,国际电信联盟无线电通信部门(ITU-R)和电信标准分局(ITU-T)制定了通用的主观质量打分与计算机制,如ITU-R BT.500-13^[2]和ITU-T P.910^[3]等。在打分的操作过程中,根据刺激方式的不同,主观质量评价方法可以分为单刺激、双刺激和多刺激的方式。单刺激即在一次打分过程中只播放失真视频,双刺激则在一次打分过程中随机播放参考、失真视频。在不同的标准机制中,操作流程略有不同。如ITU-R BT.500-13设计了单刺激连续质量估计方法(SSCQE)、双刺激失真分级方法(DSIS)、双刺激连续质量分级方法(DSCQS)、同时双刺激连续估计方法(SDSCE)等。ITU-T P.910设计了用于评价失真视频的打分方法,包括绝对类别打分法(ACR)、隐藏参考图绝对类别打分法(ACR-HR)、降质类别打分法(DCR)、匹配对比较法(PC)等。打分时可以采用百分制或等级打分制,其中较为常用的等级打分制提供了5个感受等级,即5(优秀)、4(良好)、3(一般)、2(差)、1(很差)。主观测试人员打分后,对异常数据进行处理,便可得到每个视频的平均主观意见得

分(MOS),然后再进一步通过计算失真图像与原始图像的MOS分数差得到差异平均主观意见得分(DMOS)。在绝大多数情况下,通过主观质量评价方法建立起来的主观数据库包含失真图像及其MOS/DMOS,为图像的客观质量评价方法提供了测试依据,而且人们一般也认为主观分数最接近图像的用户对视频质量的感知。目前,上述打分与计算机制是针对传统的非交互式的图像、视频业务的,并没有专门针对交互式视频设计与之相对应的打分与计算机制。虽然如此,大多数科研与工程技术人员认为上述打分与计算机制是与显示内容无关的,因此还可以将这些方法继续沿用至交互式视频的主观评价研究与应用中。在影响交互式视频主观质量的关键因素中,目前尚未有明确的研究成果,一些终端企业一方面参考了立体视频舒适度评价中的如眩晕、分辨率等因素,另一方面也站在企业自身的角度提出了包括黑边、交互延迟、卡顿等方面的因素^[4]。这些工作为本领域未来的研究与发展提供了较好的思路。

主观质量评价数据库的建立是开展质量评价打分的前提,需要就应用过程中典型的情况进行表达,如分辨率、失真类型、失真等级等。目前针对交互式视频的主观质量评价所建立的数据库较少,其建立经历了从立体视频到交互式视频的发展过程。WANG X.等考虑了非对称失真特性对视觉感知质量的影响,建立了双目立体图像主观质量评价数据库^[5]。该数据集包含4种不同的失真类型、10个场景共400组失真图像对。A. K. MOORTHY等针对对称失真,建立了包含20个场景共计365组失真图像对的LIVE-Phase-I数据集^[6]。CHEN M. J.等同时考虑了对称和非对称失真特性的影响,建立了包含8个场景和360组失真图像对

的 LIVE-Phase-II 数据集^[7]。针对立体图像质量评价的客观评价模型的建模需求, WANG J. H. 等建立的 Waterloo-IVC-3D 图像质量数据库^[8], 探索了信号失真分别对单目图像和立体图像视觉感知质量的影响。针对立体视频系统中的编码压缩方案对视觉感知质量的影响, WANG J. H. 等建立了 Waterloo-IVC-3D 视频质量数据库^[9]。随着虚拟现实 (VR) 业务的广泛应用, 3D VR 内容的视觉质量评价得到了广泛关注。近期, CHEN M. 等建立了 LIVE-3D-VR 图像质量数据库^[10], 该数据库包含了 15 个 3D VR 场景、6 种失真类型, 共计 450 组失真图像的用户评分和眼动数据。前述工作主要针对自然场景内容, 未考虑交互视频中存在的虚拟视点绘制等过程对视觉感知质量的影响。在此基础上, YANG Y. 等以交互过程中所产生的虚拟视点为切入口, 建立了虚拟视点视频主观质量评价数据库^[11-12]。该数据库主要考虑了多视点视频在彩色图、深度图压缩联合失真的情况下对虚拟视点图像绘制的相关影响, 重点考察了量化参数 (QP) 从 22 到 47, 且 $\Delta QP=5$ 的条件下, 对 5 个不同分辨率的视频进行的失真处理。上述数据库的建立, 为本领域研究工作奠定了非常重要的基础。但是, 由于主观质量评价数据库的建立是一个极其耗费资源、投入大见效慢的工作, 受到各种外部因素的影响, 该方向的工作在近年来的推进相对迟缓。

2 交互式视频质量的客观评价

客观质量评价的目标在于克服主观质量评价对人本身的依赖, 仅依靠对视频信号的分析与计算即可实现视频质量的评价, 从而使得视频质量评价从分时、分空间的人为操作变成当时当刻的自动计算, 这样可以大大提升多媒体通信系统的处理效率。

近年来, 交互式视频的客观质量评价以 360° VR 视频为主, 分别以该视频的球面映射 (SP)、等距柱面映射 (ERP)、立方体映射 (CMP) 等 3 种不同的方式为载体, 在其基础上提取视觉特征并加以建模, 来实现客观质量的评价。例如, 球面峰值信噪比 (S-PSNR)^[13]、加权峰值信噪比 (WS-PSNR)^[14] 等都是传统的峰值信噪比计算的基础上进行了微调, 以适应 VR 视频的应用。但是, 这些方法还是无法避免视频客观质量评价的典型问题, 即信号的失真不能代表视觉主观感受上的失真程度。为此, CHEN S. J. 在结构相似性度量的基础上提出了球面结构相似性 (S-SSIM) 度量模型, 能够取得比 WS-PSNR 更加贴近人眼主观感受的性能效果^[15]。这种方法较为直观, 主要是将 SSIM 方法应用到了 SP 上, 因此研究人员认为应该还会有更好的处理模式来解决上述问题。在这种思路的影响下, 利用深度学习的方式来进行视频质量评价是一种快速见效的研究手段。如 ZHANG L. 提出了综合局部描述子的图像质量评价方法 (IL-NIQE)^[16]、LIU L. X. 提出了朝向梯度下的图像质量评价方法 (OG-IQA)^[17], 他们都通过反向传播神经网络来将图像的特征映射成为图像的客观质量。此外, 利用信号分析的方法进行 VQA 的也不在少数, 如 XUE W. 提出了用梯度幅值和高斯-拉普拉斯算子进行建模的方法^[18], A. MITTAL 提出了自然场景统计失真的方法^[19], YANG Y. 提出了基于 Counterlet 小波的方法^[11, 20]、相似性评估法^[21]等。这些方法虽然在计算效率上具有较好的性能, 但是它们对 SP 与二维图像之间相互转换时所具有的视觉失真缺乏有效的分析, 因此其最终的表现性能仍然有待提升。为此, H. T. LIM 提出了一种基于对抗生成网络的 VR 视频

质量评价方法, 将多种压缩失真、位置信息、视觉特征进行了融合, 取得了较好的计算效果^[22-23]。

上述研究工作主要针对 SP 模式展开, 对 ERP 和 CMP 的模式研究较少。值得注意的是, SP 是一种平面与球面的相互映射过程, 虽然这种映射符合当前 VR 应用的工程需求, 但是存在着较多的几何失真。在这种本身就具有失真的图像上进行 VQA 计算, 是值得商榷的。相比于 ERP, SP 具有更小的失真, 而 CMP 的失真则几乎可以忽略不计。如何在这两种映射的基础上进行 VQA 的建模与计算, 并与 SP 进行有效关联, 是一个值得探索的方向。

3 交互式视频的发展趋势

在信息传递的各种形式中, 视听信号更容易让人们理解, 因此也成为了现实世界中信息的主要载体。自从视听业务以数字信号播出以来, 音视频信号在数字设备中的应用变得更加便利。这导致视听业务的表现形式越来越丰富, 人们对视听服务的需求不断激增, 这也倒逼着传统的用于承载音视频业务的通信方式不断发展。近些年来, 通信技术的不断发展, 特别是 5G 技术与产品的国际化竞争引起了人们的广泛关注。信道越来越宽, 传输速率越来越快, 通信变得无处不在, 这些都使得信源与信道之间的抱团滚动式发展产生越来越大的影响力。自由视点电视的概念于 1996 年被提出, 它认为观众应该改变观看的视角, 从被动接收到主动改变所观看的内容, 形成千人千面的视觉效果^[24]。虽然上述工作未能带来商业价值, 但是这个交互式媒体的思路与目前低时延、大带宽的通信技术相结合, 在近几年形成了 VR、云游戏、云主机的高交互视听业务, 它和在 2020 年新冠肺炎疫情期间发挥关键作用的在线教育、直

播连麦、在线会议、远程医疗等交互式视听业务模式一起开始逐渐被用户所接纳。未来媒体势必以千人千面为目标,朝着大数据量、大计算量、大通信量的方向发展。上述业务架构具有“云-边-端”协同计算特点,在未来一定会衍生出更丰富的媒体应用。

为了在这些关键应用中保障用户的体验,增强用户对交互式视频的粘滞度,无论是 KQI 还是 VQA,仍有一些问题值得深入研究、探讨。

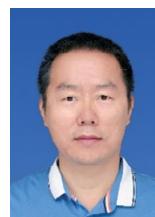
致谢

本文的工作得到深圳大学计算机学院王旭副教授的支持,在此特别表示感谢。

参考文献

- [1] HUYNH-THU Q, GARCIA M N, SPERANZA F, et al. Study of rating scales for subjective quality assessment of high-definition video [J]. IEEE transactions on broadcasting, 2011, 57(1): 1-14. DOI:10.1109/tbc.2010.2086750
- [2] ITU. Methodology for the subjective assessment of the quality of television pictures, Recommendation ITU-R BT.500-13 [EB/OL]. [2020-12-20]. https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-13-201201-1!!PDF-E.pdf
- [3] ITU. Subjective video quality assessment methods for multimedia applications, ITU-T P. 910 [EB/OL]. [2020-12-20]. <https://www.itu.int/rec/T-REC-P.910/en>
- [4] Cloud VR 用户体验与评测白皮书 [EB/OL]. [2020-12-20]. <https://www.huawei.com/minisite/static/cloud-vr-user-experience-evaluation-white-paper-cn.pdf>
- [5] WANG X, YU M, YANG Y, et al. Research on subjective stereoscopic image quality assessment [C]//Multimedia Content Access: Algorithms and Systems III. San Jose, CA, USA: SPIE, 2009: 18-22. DOI:10.1117/12.807641
- [6] MOORTHY A K, SU C C, MITTAL A, et al. Subjective evaluation of stereoscopic image quality [J]. Signal processing: image communication, 2013, 28(8): 870-883. DOI: 10.1016/j.image.2012.08.004
- [7] CHEN M J, SU C C, KWON D K, et al. Full-reference quality assessment of stereopairs accounting for rivalry [J]. Signal processing: image communication, 2013, 28(9): 1143-1155. DOI:10.1016/j.image.2013.05.006
- [8] WANG J H, REHMAN A, ZENG K, et al. Quality prediction of asymmetrically distorted stereoscopic 3D images [J]. IEEE transactions on image processing, 2015, 24(11): 3400-3414. DOI: 10.1109/tip.2015.2446942
- [9] WANG J H, WANG S Q, WANG Z. Asymmetrically compressed stereoscopic 3D videos: quality assessment and rate-distortion performance evaluation [J]. IEEE transactions on image processing, 2017, 26(3): 1330-1343. DOI: 10.1109/tip.2017.2651387
- [10] CHEN M, JIN Y, GOODALL Y, et al. Study of 3D virtual reality picture quality [J]. IEEE journal of selected topics in signal processing, 2020, 14(1):89-102
- [11] YANG Y, DAI Q. Contourlet-based image quality assessment for synthesized virtual image [J]. Electronics letters, 2010, 46(7): 492-494. DOI: 10.1049/el.2010.3522
- [12] YANG Y, WANG X, LIU Q, et al. User models of subjective image quality assessment on virtual viewpoint in free-viewpoint video system [J]. Multimedia tools and applications, 2016, 75(20): 12499-12519. DOI: 10.1007/s11042-014-2321-7
- [13] YU M, LAKSHMAN H, GIROD B. A framework to evaluate omnidirectional video coding schemes [C]//2015 IEEE International Symposium on Mixed and Augmented Reality. Fukuoka, Japan: IEEE, 2015: 31-36. DOI:10.1109/ismar.2015.12
- [14] SUN Y, LU A, YU L. AHG8: WS-PSNR for 360 video objective quality evaluation: ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, JVET-D0040 [S]. 2016
- [15] CHEN S J, ZHANG Y X, LI Y M, et al. Spherical structural similarity index for objective omnidirectional video quality assessment [C]//2018 IEEE International Conference on Multimedia and Expo (ICME). San Diego, CA, USA: IEEE, 2018: 1-6. DOI:10.1109/icme.2018.8486584
- [16] LIN ZHANG, LEI ZHANG, BOVIK A C. A feature-enriched completely blind image quality evaluator [J]. IEEE transactions on image processing, 2015, 24(8): 2579-2591. DOI:10.1109/tip.2015.2426416
- [17] LIU L X, HUA Y, ZHAO Q J, et al. Blind image quality assessment by relative gradient statistics and adaboosting neural network [J]. Signal processing: image communication, 2016, 40: 1-15. DOI:10.1016/j.image.2015.10.005
- [18] XUE W, MOU X, ZHANG L, et al. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features [J]. IEEE transactions on image processing, 2014, 23(11): 4850-4862. DOI: 10.1109/tip.2014.2355716
- [19] MITTAL A, MOORTHY A K, BOVIK A C. No-reference image quality assessment in the spatial domain [J]. IEEE transactions on image processing, 2012, 21(12): 4695-4708. DOI: 10.1109/tip.2012.2214050
- [20] 蒋刚毅, 王旭, 杨铀, 等. 基于 Contourlet 的降维图像质量评价模型 [J]. 光子学报, 2009, 20(5):1658-1662
- [21] 黄大江, 郁梅, 杨铀, 等. 基于相似度的立体图像对中心视点图像质量评价方法 [J]. 光子学报, 2008, 37(8):1673-1697
- [22] LIM H T, KIM H G, RA Y M. VR IQA Net: deep virtual reality image quality assessment using adversarial learning [C]//IEEE international conference on acoustics, speech and signal processing. Calgary, AB, Canada: IEEE, 2018: 6737-6741. DOI: 10.1109/ICASSP.2018.8461317
- [23] KIM H G, LIM H T, RO Y M. Deep virtual reality image quality assessment with human perception guider for omnidirectional image [J]. IEEE transactions on circuits and systems for video technology, 2020, 30(4): 917-928. DOI: 10.1109/tcsvt.2019.2898732
- [24] TANIMOTO M. FTV: free-viewpoint television [J]. Signal processing: image communication, 2012, 27(6): 555-570. DOI: 10.1016/j.image.2012.02.016

作者简介



李继龙, 国家广播电视总局广播电视科学研究院正高级工程师、学术带头人; 主要研究工作包括融合媒体、5G 广播电视、广播电视融合网、无线数字广播、信道编码和调制技术研究等; 曾参与多项国家、部委重要项目研发工作, 作为主要研究人员参与了有线/无线卫星融合网、卫星直播标准和数字音频广播标准的研究与制定; 曾获得广电总局“科技创新奖”一等奖一项、二等奖两项, “王选新闻科学技术奖”一等奖两项、二等奖一项; 发表论文 40 余篇, 出版著作 3 部, 获得授权国家发明专利 6 项。



赵茜, 武汉理工大学信息工程学院在读研究生; 主要从事机器学习、深度学习领域的研究工作。



杨铀, 华中科技大学电子信息与通信学院教授、博士生导师, 中国图象图形学会图象视频处理与通信专委会秘书长; 主要从事以视觉感知与计算为核心的计算机视觉、计算摄影学、立体视频系统等方面的研究工作; 2012 年获教育部高等学校科技成果技术发明一等奖, 2018 年当选英国国际工程技术学会会士 (IET Fellow), 2020 年获 TET 创新技术奖中“通信与信息技术”领域杰出创新奖; 主持和参与包括国家重点研发计划、国家自然科学基金面上项目、“863”项目、国家重大专项、国家重大科技成果转化等在内的项目 20 余项; 发表论文 80 余篇, 获得授权国家发明专利 24 项。

HTTP 自适应流媒体直播系统中的用户体验质量优化

QoE Optimization in HTTP Adaptive Live Streaming System



宋靳铎 /SONG Jinke, 张远 /ZHANG Yuan, 王博 /WANG Bo

(中国传媒大学, 中国 北京 100024)
(Communication University of China, Beijing 100024, China)

摘要: 分析 HTTP 自适应流媒体直播系统对终端用户体验质量 (QoE) 产生影响的各类因素及其相互之间的作用关系, 对基于服务器端、网络传输以及客户端的 QoE 优化策略进行总结。认为 HTTP 自适应流媒体直播系统的 QoE 优化重点在于降低延时, 提出结合网络层和应用层影响因素来降低延时并提升用户 QoE 的建议。

关键词: 流媒体直播; 自适应流媒体; 用户体验质量; 优化策略

Abstract: The influence factors on the users' quality of experience (QoE) in a hypertext transport protocol (HTTP) adaptive streaming system and the interaction between these factors are studied. Then the QoE optimization strategies from the aspects of the server, network transmission and client side are summarized. The QoE optimization of HTTP adaptive live streaming system focuses on reducing delay, and combination with network layer and application layer factors can reduce delay and improve user QoE.

Keywords: live streaming; adaptive streaming; QoE; optimization strategy

DOI: 10.12142/ZTETJ.202101011

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210122.1809.006.html>

网络出版日期: 2021-01-25

收稿日期: 2020-12-15

近年来, 随着互联网与流媒体技术的飞速发展, 游戏直播、在线教育等直播服务发展迅速。基于超文本传送协议 (HTTP) 的自适应流媒体技术 (HAS) 由于其高兼容性、高可扩展性在直播场景中得到了广泛应用。根据 WOWZA 发布的《2019 Streaming Video Latency Report》^[1], 超过 50% 的直播服务提供商采用了基于 HAS 的流媒体直播系统架构。流媒体直播服务的低延时、高质量需求给用户体验质量的优化带来了新的挑战。

1 基于 HAS 的流媒体直播系统框架

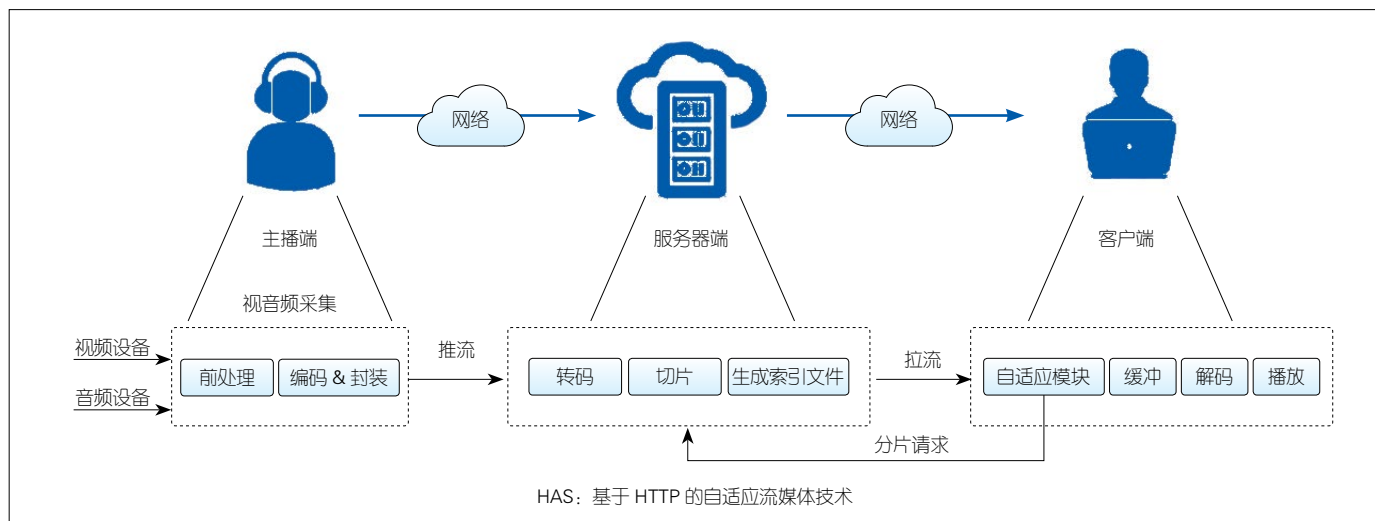
在 HAS 技术中, 媒体数据源被编码成不同码率的媒体切片文件以适应不同的网络状况和客户端设备类型, 客户端根据带宽或缓冲区状态请求合适码率的媒体分片, 以减少卡顿事件的发生, 提升带宽利用率。

如图 1 所示, 基于 HAS 的流媒体直播传输系统框架包括 3 部分: 主播端、服务器端和客户端。主播端主要用来实现媒体采集、前处理、编码和封装的功能, 并将封装好的媒体流推送至服务器。服务器需要对同一媒体

内容准备多种码率的媒体文件, 因此需要对主播端推送的媒体流进行实时转码, 并将每种码率的媒体内容进行切片处理。服务器端存储的媒体分片的码率和时长等信息都被记录在一个随直播进行且实时更新的索引文件中。客户端在从服务器拉取媒体流时会首先拉取索引文件, 再根据索引文件的信息以及当前的估计带宽或缓冲区状态对下一个向服务器请求的分片码率进行自适应决策, 并对已下载的媒体分片进行解码播放。

经典的 HAS 协议包括 Microsoft 公司提出的微软平滑流协议 (MSS)^[2]、

基金项目: 国家自然科学基金 (61971382)



▲图1 基于HAS的流媒体直播传输系统框架

Apple 公司提出的 HTTP 实时流协议 (HLS)^[3], 以及 Adobe 提出的 HTTP 动态流协议 (HDS)^[4]。虽然这些协议遵循的技术框架相同, 但是彼此之间互不兼容; 因此, 动态图像专家组 (MPEG) 与第三代合作伙伴计划 (3GPP) 联合提出了开源的 MPEG- 动态自适应流媒体 (DASH)^[5] 标准。目前业界广泛采用的标准是 HLS 和 MPEG-DASH, 然而直播场景下两种标准的延时均在 6 s 以上。近两年来, 针对用户日益增长的低延时需求, 研究者提出了能够将延时控制在 3 s 以内的低延时 HLS^[6] 以及基于用媒体应用格式 (CMAF) 的低延时 DASH^[7] 解决方案。

2 HTTP 自适应流媒体直播系统中的用户体验质量评估

2.1 QoE 影响因素

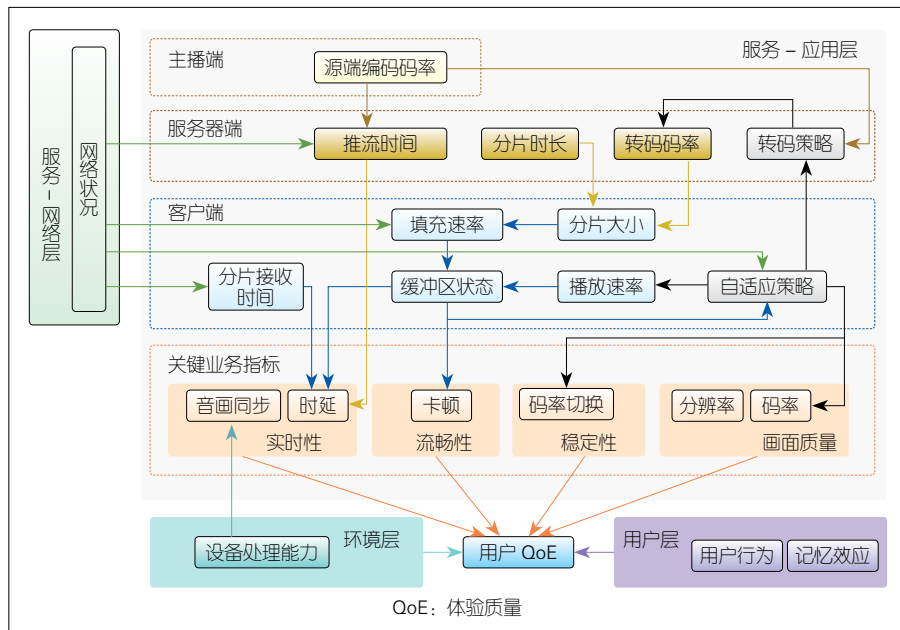
用户体验质量 (QoE) 通常用来评价流媒体直播系统中终端用户对服务的满意程度, 其定义为: 用户在一定客观环境中对使用的服务或者业务的整体认可程度。根据 QoE 的定义, HTTP 自适应流媒体直播系统中的 QoE 影响因素可分为服务、环境、用户 3 个方面, 如表 1 所示。

本文中我们仅考虑服务、环境和用户层面中与技术相关的因素。如图 2 所示, 关键业务指标、用户层、环境层可对用户 QoE 产生直接影响。关键

▼表1 流媒体直播系统中用户体验质量影响因素

影响因素	应用层	网络层
服务	<ul style="list-style-type: none"> 关键业务指标: 码率、码率切换幅度/频次、分辨率、卡顿、延时、音画同步 客户端: 码率自适应策略、缓冲区状态 (缓冲区占用量、容量) 服务器端: 转码策略、分片时长 主播端: 直播内容、音视频采集设备、前处理方法、编码方案 	<ul style="list-style-type: none"> 网络吞吐量 丢包率 往返延时 抖动
环境	<ul style="list-style-type: none"> 软硬件设施: 播放设备的分辨率、CPU 处理能力 自然环境: 地点、噪声、光照等 人文环境: 人文规范、社会观念等 	
用户	<ul style="list-style-type: none"> 用户对服务的期望、主观偏好、年龄性别、身心状态以及文化背景等 	

CPU: 中央处理器



▲图2 直播系统中的用户 QoE 及其影响因素

业务指标可从画面质量、稳定性、流畅性以及实时性 4 个方面来衡量。码率和分辨率直接影响直播画面的清晰度，直播全程的平均码率和分辨率共同决定了流媒体服务的画面质量水平；码率切换的频次和幅度反映了图像质量的波动状况，进而决定了视觉观感的平稳度；卡顿导致视频内容不连续、声音断续等问题，通常以卡顿次数、频率以及持续时间来反映流媒体服务的流畅性；延时是媒体流从主播端发出到客户端接收并观看所需要的时间，是反映流媒体服务实时性的业务指标，高延时极大地降低了用户对服务的满意度；音画不同步是由于网络状况较差（延时、抖动等）及设备处理能力不足造成的视音频不同步，降低了用户的服务体验质量。

由图 2 可知，网络层与主播端、服务器端、客户端应用层中的影响因素相互作用，共同影响 HTTP 自适应流媒体直播服务的关键业务指标，这也成为 QoE 的间接影响因素。例如，客户端的缓冲区状态和自适应策略直接影响了服务的关键业务指标，缓冲区状态示意如图 3 所示。由图 3 可知，客户端的网络状态与分片大小（即分片的数据量）共同决定了缓冲区数据的填充速率。通常，客户端按正常速率播放，但当缓冲区占用量即将耗尽时，可通过客户端的自适应策略调慢播放速率，以等待缓冲区数据充盈，降低卡顿事件发生的概率。而当缓冲区占用量较大时，会带来较大的播放延时，那么则需要加快播放速率以降低播放延时。客户端还可以根据缓冲区状态或当前网络状态来进行码率自适应调节，决定下一请求分片的码率级别，以尽量避免卡顿事件的发生。

在 HTTP 自适应流媒体直播服务中，服务器主要实现对视频进行转码和切片。如图 2 所示，服务器根据主

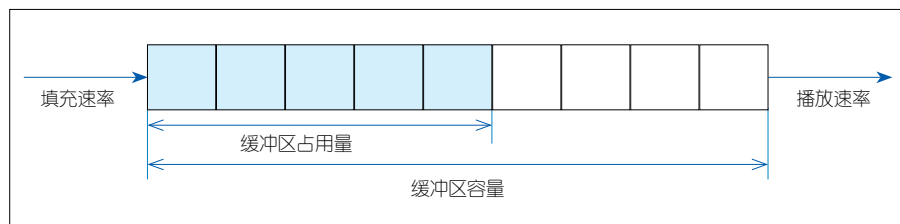
播端编码码率上限和客户端自适应模块决策的分片码率调整转码策略。分片码率与分片时长共同决定了传送到客户端的数据量（即分片大小），而分片大小又与客户端网络状态共同影响了缓冲区的填充速率。除限制服务器最高转码码率外，主播端的编码码率还决定了上行网络的传输数据量，与上行网络状态共同影响源视频推流到服务器的时间，即上行延时。上行延时与下行延时（即视频分片传输到客户端的时间）共同构成了直播系统的传输延时。需要注意的是，由于前处理、转码、切片、解码等阶段的处理延时难以控制，因此图 2 中并没有体现处理延时对整个系统延时的影响。

2.2 QoE 建模

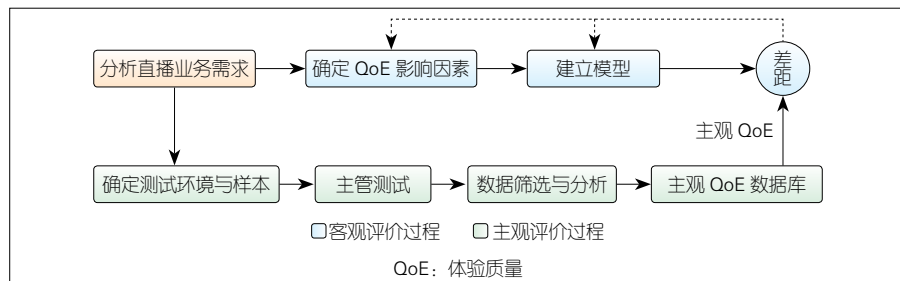
QoE 建模的目标是分析影响 QoE 的各项因素，并建立模型描述这些影响因素与主观测试结果的关系，从而为流媒体直播系统的优化提供参考。现有的大多数针对直播服务的 QoE 建模流程如图 4 所示，首先根据直播业务需求确定主观测试的客观环境以及媒体流样本，再通过对样本进行主观测试和数据筛选来建立主观 QoE 数据

库，数据库应包含媒体流样本素材、样本对应的主观分数以及相应的卡顿、画面质量等客观测试数据。然后，分析直播业务需求并确定 QoE 的影响因素，再对各影响因素与预测 QoE 之间的映射关系建模，并不断调整模型结构、影响因素的选择、影响因素分配权重等，来最小化预测 QoE 分数与主观 QoE 得分之间的差距。在实际建模过程中，通常以均方根误差（RMSE）、斯皮尔曼相关系数（SROCC）以及皮尔逊相关系数（PLCC）来衡量主观 QoE 与预测 QoE 之间的相关程度。

大多数 QoE 建模采用了经典机器学习方法和深度学习方法。如 C. G. BAMPIS 等采用各种回归模型如岭回归（RR）、支持向量回归（SVR）以及随机森林（RF）等来对画面质量、码率下降等 QoE 影响因素与用户主观评分的映射关系建模^[8]；N. ESWARA 等利用级联的长短期记忆网络（LSTM）来捕获用户 QoE 在时间轴上复杂的依赖关系以及对流媒体服务的非线性反应^[9]；D. GHADIVARAM 等将 QoE 影响因素分为 3 类：与卡顿相关的因素、与视频内容相关的因素以及对客户端缓冲区状态的建模，并将这

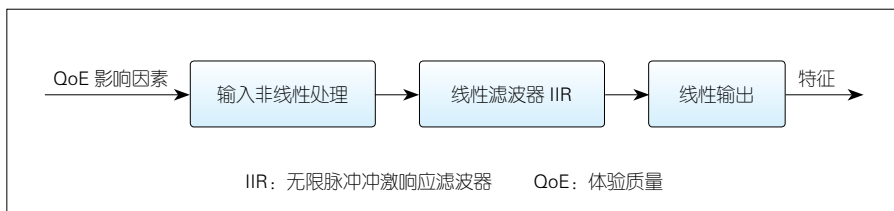


▲图 3 缓冲区状态示意图



▲图 4 QoE 建模流程

些影响因素分别输入到如图 5 所示的 Hammerstein-Wiener^[10]非线性模型中, 以进一步提取影响 QoE 的深层特征, 然后将提取的深层特征组合输入到 SVR 中以预测 QoE, 显著提升了主观 QoE 和预测 QoE 之间的相关性^[11]。



▲图 5 Hammerstein-Wiener 模型

3 HTTP 自适应流媒体直播系统中的 QoE 优化策略

与点播相比, 直播业务对 QoE 中的延时性能要求更高。可从流媒体直播系统中的服务器端、传输网络和客户端进行旨在提高 QoE 性能的优化。

3.1 服务器端优化

对于 HTTP 自适应流媒体直播来说, 服务器的功能是提供适配客户端需求的媒体流, 其重点在于如何根据客户端设备能力和网络条件来快速准备适合的码率分片, 以降低延时, 提高带宽利用率并减小卡顿事件发生的概率。因此, 服务器端的优化主要从调整分片长度和视频编码的码率控制两个方面进行。

在网络状况不稳定时, 固定不变的分片时长会引起客户端加载速率的剧烈变化, 从而导致码率切换频繁, 引起直播延迟与卡顿。针对上述问题, 费泽松等根据客户端请求码率的变化, 使切片长度在预设范围内随加载速率改变, 并以流畅性作为 QoE 优化目标, 实现了基于 HLS 的自适应码率视频直播的 QoE 监视与优化方法^[12]。针对下载完整分片才能播放视频的问题, 低延时直播视频传输平台 (L3VTP) 将视频传输粒度从分片级别降至帧级别, 先将转码视频直接推流到内容分发网络 (CDN), 再直接推流到客户端, 省去了服务器切片处理和客户端周期性请求分片的过程, 从而降低直播系统的播放延时与处理延时^[13]。

在流媒体系统中, 视频编码器的

码率控制算法通常是针对连续码流设计的, 未将视频切片情况考虑在内, 缺乏对视频分片层面的控制, 从而导致生成的分片码率相对设定值波动较大, 引起带宽浪费和卡顿。在低延时、小缓存的直播场景中, QoE 表现得更差。詹亘等针对上述问题提出了基于切片级别进行比特分配的码率控制算法^[14]。在切片时长固定时, 按照设定码率为每个切片分配目标比特数。在给切片内所有帧分配比特时, 对切片的帧类型构成进行预测, 并为不同类型帧分配不同权值, 再根据权值进行帧级别的比特分配。通过建立基于残差变换绝对值和 (SATD) 和量化系数的线性预测模型, 利用模型迭代进行宏块级别的量化系数调整, 从而实现单帧码率的准确控制。实验结果表明, 与 x264 编码器相比, 该编码方案将视频切片码率波动降低了 76%。

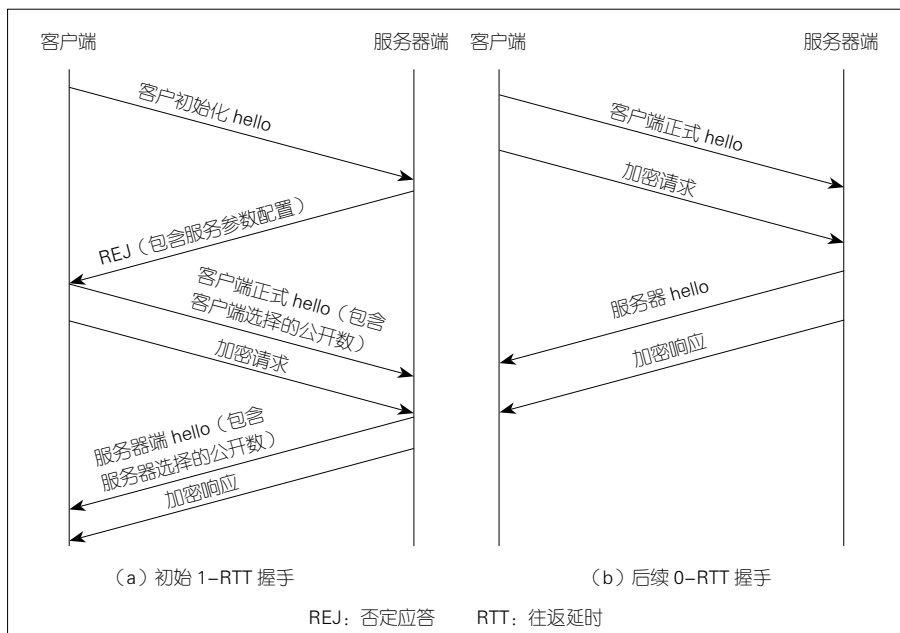
3.2 网络传输优化

针对传输网络优化的研究可分为两类: 通过合理部署网络架构或选择数据中心的方式来优化传输; 通过优化或改变流媒体传输协议来有效降低传输延时。

在保持原有网络架构的基础上, 可使用软件定义网络 (SDN) 来实现网络构架的灵活部署。例如, P. H. THING 等将所有路由协议集成在一个 SDN 环境中, 通过拓扑管理器收集链路拓扑信息, 通过流量管理器检测各路由的网络状况, 并基于上述拓扑信息和网络状况设计了一个动态最优路

径选择算法, 实现了基于 SDN 的流媒体传输最优路径的选择, 提升了链路利用率^[15]。直播平台通常选用租借的云服务器来提供视频转码、传输以及响应用户请求的直播服务, 然而由于直播端与客户端分布广泛, 使用基于云的数据中心来满足用户需求仍然具备挑战性, C. DONG 等提出了一种为主播端和客户端动态选择数据中心的算法, 在保证用户 QoE 的前提下, 为直播服务提供商节省了运行成本^[16]。

大多数直播传输方案在传输层均依赖于传输控制协议 (TCP)。然而, TCP 在数据传输前须完成 3 次握手以建立连接, 如果使用加密 Web 服务, 还须增加一次安全套接字协议 / 安全传输层协议 (SSL/TLS) 握手。与用户数据报协议 (UDP) 相比, 这种按序传输的可靠连接会引入建立连接、丢包重传等延时。针对上述问题, TCP 快速连接协议 (TFO)^[17] 利用 cookie 信息在确认字符回到接收端前发送数据, 从而在建立握手的同时还进行了有效的数据传输, 有效降低了延时, 但是由于兼容性较差未被广泛采用; 谷歌提出快速 UDP 网络连接协议 (QUIC), 通过类似 TFO 的技术使传输握手和加密同时完成, 实现了在一个往返延时 (RTT) 内建立可靠连接的功能^[18]。如图 6 (a) 所示, 客户端在建立会话时可将 cookie 和加密数据直接发送至服务器, 服务器再利用这些信息对客户端进行验证, 验证通过即开始接收数据。之后客户端可在本地缓存加密认证信息, 从而在与服务



▲图 6 一种基于用户数据报协议的可靠低延时传输协议中的握手延时优化

器恢复会话时实现零 RTT 的连接延迟，如图 6 (b) 所示。此外，SHI H. 等针对延时敏感业务，提出了延时敏感性传输协议 (DTP)，将若干数据包组成的数据单元抽象为数据块，根据数据包的接收截止时间和优先级、数据块的剩余大小以及链路网络状态来决定数据包发送的先后顺序^[19]。

3.3 客户端优化

相对于自适应流媒体点播系统，直播的流媒体数据不是事先制作并保存在服务器中的，而是从主播端实时采集并推流到服务器端的。所以直播客户端的码率自适应在考虑点播 QoE 影响因素的基础上，还增加了延时因素，通常综合卡顿、码率质量、码率切换以及延时等影响因素来进行客户端的码率自适应决策。

客户端的码率自适应决策算法可分为启发式算法和模型法。启发式算法是一种基于直观或经验构造的算法。如 XIE L. 等认为在直播场景的客户端小缓冲区前提下，带宽以及带宽波动的识别至关重要，因此提出了一种基

于缓冲区阈值、带宽瞬时值以及带宽波动状态的码率自适应算法，以降低卡顿率为目标，实现了低延迟条件下的无缝播放，减少了码率切换的频次，提高了用户体验质量^[20]。针对码率自适应问题的模型法可以分为：控制理论、效用最优化问题以及强化学习方法等。近年来，客户端的码率自适应优化大多基于强化学习的方法，并采用动态奖励函数来实现直播流媒体码率的自适应优化。例如 BitLat^[21] 和 HD3^[22] 基于延时、缓冲时间、分片码率等影响因素建立 QoE 线性预测模型，并将预测 QoE 作为动态奖励函数来实现直播场景下的码率控制和延迟控制。

除了对下载分片码率进行自适应调整外，还可以通过调整播放速率来实现基于客户端的混合自适应控制策略。播放速率自适应 (AMP) 算法根据网络状况和客户端缓冲区占用量实时调整客户端视频的播放速度，以降低客户端的播放延时。如 ZHANG G. 等依据直播的低延时要求，提出了一种联合调整分片码率和播放速率的自适应算法，最终使得直播视频在保证一

定画面质量的同时降低了播放时延^[23]。此外，客户端通常还以丢帧的形式来引入延时控制机制，以满足直播对实时性的要求。如 MILLER K. 等利用基于客户端的带宽预测和对预测误差分布的估计，通过跳过部分视频片段来降低延时，并利用码率过渡函数来避免码率切换幅度过大带来的突兀感^[24]；HONG R. 等通过跳帧的机制来最小化基于直播端到端延时的缓冲区阈值，提升了平均 QoE^[25]；Vabis^[26] 是一种跨服务器端和客户端的 QoE 优化策略，实现了基于帧级别的细粒度码率控制，并在客户端引入了 3 种延时机制，将直播平均延时降低了 32% ~ 77%，并提高了 28% ~ 67% 的平均 QoE。然而，目前针对端到端的自适应流媒体直播系统的优化还比较少，因此在这方面有待于进一步研究与探讨。

4 结束语

本文首先介绍了 HTTP 自适应流媒体直播系统框架，并对其 QoE 影响因素进行分析，最后针对 HTTP 自适应流媒体直播系统中的 QoE 优化策略进行了总结。HTTP 自适应流媒体直播系统的优化目前还存在以下问题：

(1) 由于自适应机制不可避免地引入包括转码、切片、自适应决策在内的处理延时，因此不适用于“电商带货”等对超低延时有需求的直播场景。

(2) 现有的 QoE 优化大多针对客户端应用层的影响因素，或者将传输优化和客户端优化单独考虑，没有考虑网络层和应用层相互作用的关系对 QoE 的影响，而网络传输延时是整个直播系统延时的重要组成部分；因此，想要从本质上降低整个系统的延时，有必要深入研究网络层与应用层之间的复杂关系，综合考虑网络层与应用层影响因素，实现对自适应流媒体直播系统的跨层 QoE 优化。

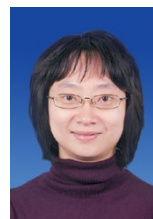
参考文献

- [1] WOWZA. 2019 Video streaming latency report [EB/OL]. [2020-12-10]. <https://www.wowza.com/wp-content/uploads/Streaming-Vid-eo-Latency-Report-Interactive-2020.pdf>
- [2] Smooth streaming protocol [EB/OL]. [2020-10-30][2020-12-10]. https://docs.microsoft.com/en-us/openspecs/windows_protocols/ms-sstr/8383f27f-7efe-4c60-832a-387274457251
- [3] HTTP live streaming [EB/OL]. [2020-12-10]. https://developer.apple.com/documentation/http_live_streaming
- [4] Configure HTTP dynamic streaming and HTTP live streaming using AMS [EB/OL]. [2020-12-10]. https://helpx.adobe.com/adobe-media-server/dev/configure-dynamic-streaming-live-streaming.html#overview_of_http_dynamic_streaming_and_http_live_streaming
- [5] SODAGAR I. The MPEG-DASH standard for multimedia streaming over the Internet [J]. IEEE multimedia, 2011, 18(4): 62-67. DOI:10.1109/mmul.2011.71
- [6] Enabling low-latency HLS [EB/OL]. [2020-12-10]. https://developer.apple.com/documentation/http_live_streaming/enabling_low-latency_hls
- [7] DASH-IF. Low-latency modes for DASH [EB/OL]. [2020-12-10]. <https://dashif.org/docs/CR-Low-Latency-Live-r8.pdf>
- [8] BAMPIS C G, BOVIK A C. Learning to predict streaming video QoE: distortions, rebuffering and memory [EB/OL]. [2020-12-10]. <https://arxiv.org/abs/1703.00633>
- [9] ESWARA N, ASHIQUE S, PANCHBHAI A, et al. Streaming video QoE modeling and prediction: a long short-term memory approach [J]. IEEE transactions on circuits and systems for video technology, 2020, 30(3): 661-673. DOI:10.1109/tcsvt.2019.2895223
- [10] GHADIYARAM D, PAN J, BOVIK A C. A time-varying subjective quality model for mobile streaming videos with stalling events [C]//SPIE Optical Engineering + Applications. International Society for Optics and Photonics. USA: SPIE, 2015
- [11] GHADIYARAM D, PAN J, BOVIK A C. Learning a continuous-time streaming video QoE model [J]. IEEE transactions on image processing, 2018, 27(5): 2257-2271. DOI:10.1109/tip.2018.2790347
- [12] 费泽松, 王飞, 孙尧, 等. 一种自适应码率视频直播的 QoE 监控和优化方法: CN105357591A [P]. 2016
- [13] YI G, YANG D, WANG M W, et al. L3VTP: a low-latency live video transmission platform [C]//Proceedings of the ACM SIGCOMM 2019 Conference Posters and Demos on - SIGCOMM Posters and Demos, 19. New York, NY, USA: ACM Press, 2019: 138-140. DOI:10.1145/3342280.3342336
- [14] 詹巨, 肖晶, 陈宇静, 等. 面向自适应码率视频直播的码率控制算法 [J]. 计算机工程, 2019, 45(3): 268-272
- [15] THINH P H, DAT N T, NAM P N, et al. An efficient QoE-aware HTTP adaptive streaming over software defined networking [J]. Mobile networks and applications, 2020, 25(5): 2024-2036. DOI:10.1007/s11036-020-01543-1
- [16] DONG C, WEN W, XU T, et al. Joint Optimization of data-center selection and video-streaming distribution for crowdsourced live streaming in a geo-distributed cloud platform [J]. IEEE Transactions on Network and Service Management, 2019:1-1
- [17] CHENG Y, CHU J, RADHAKRISHNAN S, et al. TCP fast open [R]. RFC Editor, 2014
- [18] CUI Y, LI T X, LIU C, et al. Innovating transport with QUIC: design approaches and research challenges [J]. IEEE Internet computing, 2017, 21(2): 72-76. DOI:10.1109/mic.2017.44
- [19] SHI H, CUI Y, QIAN F, et al. DTP: deadline-aware transport protocol [C]//Proceedings of the 3rd Asia-Pacific Workshop on Networking 2019. New York, NY, USA: ACM, 2019: 1-7. DOI:10.1145/3343180.3343191
- [20] XIE L, ZHOU C, ZHANG X G, et al. Dynamic threshold based rate adaptation for HTTP live streaming [C]//2017 IEEE International Symposium on Circuits and Systems (ISCAS). Baltimore, MD, USA: IEEE, 2017. DOI:10.1109/iscas.2017.8050574
- [21] WANG C, GUAN J F, FENG T T, et al. Bit-Lat: bitrate-adaptivity and latency-awareness algorithm for live video streaming [C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2019: 2642-2646. DOI:10.1145/3343031.3356069
- [22] JIANG X L, JI Y S. HD3: distributed dueling DQN with discrete-continuous hybrid action spaces for live video streaming [C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2019: 2632-2636. DOI:10.1145/3343031.3356052
- [23] ZHANG G H, LEE J Y B. LAPAS: latency-aware playback-adaptive streaming [C]//2019 IEEE Wireless Communications and Networking Conference (WCNC). Marrakech, Morocco: IEEE, 2019: 1-6. DOI: 10.1109/wcnc.2019.8885622
- [24] MILLER K, AL-TAMIMI A K, WOLISZ A. QoE-based low-delay live streaming using throughput predictions [J]. ACM transactions on multimedia computing, communications, and applications, 2017, 13(1): 1-24. DOI: 10.1145/2990505
- [25] HONG R Y, SHEN Q W, ZHANG L, et al. Continuous bitrate & latency control with deep reinforcement learning for live video streaming [C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2019: 2637-2641. DOI:10.1145/3343031.3356063
- [26] FENG T T, SUN H F, QI Q, et al. Vabis: video adaptation bitrate system for time-critical live streaming [J]. IEEE transactions on multimedia, 2020, 22(11): 2963-2976. DOI:10.1109/tmm.2019.2962313

作者简介



宋新铎, 中国传媒大学信息与通信工程学院在读硕士研究生; 主要研究方向为流媒体系统的 QoE 优化。



张远, 中国传媒大学媒体融合与传播国家重点实验室副主任、教授; 主要研究方向为多媒体通信、智能媒体分析与处理。



王博, 中国传媒大学信息与通信工程学院在读硕士研究生; 主要研究方向为实时音视频通信。



小视频内容分析技术发展探讨

Short Video Content Analysis Technology

薛向阳 /XUE Xiangyang, 李斌 /LI Bin

(复旦大学, 中国 上海 200433)
(Fudan University, Shanghai 200433, China)

DOI: 10.12142/ZTETJ.202101012

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210118.1639.002.html>

网络出版日期: 2021-01-18

收稿日期: 2020-12-15

摘要: 小视频数量呈爆炸式增长态势, 并引发了许多技术需求, 包括小视频的编辑、搜索、推荐、溯源、审查和监管等。介绍了小视频数据的主要特点和小视频内容分析技术面临的挑战, 并对目标检测、追踪、重识别等小视频内容分析技术的研究进展做了综合分析。认为只有构建一个整合多种不同算法的系统, 才能够更准确、更鲁棒地解决分析问题, 才能系统性地完成小视频内容分析任务。

关键词: 小视频; 内容分析技术; 视频目标检测; 多目标追踪; 人物重识别

Abstract: The number of short videos has increased explosively, which has led to more technical requirements, such as editing, searching, recommendation, sourcing, censoring, and monitoring of short videos. The main features of short video data and the challenges faced by the short video content analysis technology are introduced. The research progress of short video content analysis technologies such as object detection, tracking, and re-identification is comprehensively analyzed. It is considered that only by building a system that integrates multiple algorithms, can the analysis problems be solved more accurately and robustly, and the short video content analysis task can be completed systematically.

Keywords: short video; content analysis technology; video object detection; multi-object tracking; person re-identification

1 小视频数据类型与特点

1.1 小视频数据类型

随着抖音、快手、腾讯微视、西瓜视频等小视频应用平台的兴起, 小视频已经随处可见。在激烈的竞争下, 市场上涌现出了不同类别的小视频内容。

(1) 社交生活类

以快手、抖音、腾讯微视等为代表的平台, 鼓励用户拍摄、制作、上传小视频, 分享自己的生活点滴, 这方便了用户拓宽自己的社交范围。此类小视频主题多为生活记录, 如拍摄宠物、烹饪、服饰等。通过分享生活点滴, 用户可以找到与自己趣味相投的朋友, 拓宽社交圈。

(2) 内容服务类

以西瓜视频、梨视频为代表的平台, 依靠大数据分析为用户提供精准内容服务, 如感兴趣的话题、认识的朋友、关心的产品等。此类小视频主题多为行业热点资讯、育儿经验或家教信息、“双十一”优惠活动等。

(3) 剪辑技术类

以小咖秀等为代表的平台, 为对视频制作感兴趣的用户提供制片剪辑等功能, 使用户以更灵活幽默的方式记录自己的生活。此类小视频主题多为宣传视频、纪念视频、情景短剧以及其他具有特殊意义的视频(如高考加油视频)等。

1.2 小视频数据特点

小视频数据除了具有规模海量这一特点之外, 其余还包括类型繁多、

特效复杂、姿态多变等。

(1) 类型繁多

类型繁多是小视频数据的一大特点。小视频数据包含的物体类别为开集, 除人物之外, 还涵盖宠物、电子产品、音乐器材、体育用品等。此外, 与图像数据集(ImageNet)^[1]的1 000类和目标检测数据集(COCO)^[2]的80类相比, 小视频数据的类别更丰富, 包含更多的子类, 如不同品种的猫和狗、不同品牌的电子产品等。

(2) 特效复杂

与其他视频相比, 小视频往往包含更多的特效, 以使自身更具有吸引力和娱乐性, 如各种幻灯片转场、人物美颜特效、多屏镜面特效等。这对目标检测和追踪等分析任务而言, 是一个不可忽视的巨大挑战。

(3) 姿态多变

在小视频中,各目标的外观姿态往往变化较大。小视频记录生活点滴,包含大量特写镜头。一段小视频主题可能聚焦于人、动物、产品等。这些目标围绕的主题包含较多姿态和外观变化,例如人的换装小视频、宠物成长记录小视频等。

除前文提到的3种特点外,由于小视频的拍摄设备多为智能手机,故小视频数据的特点还包括画面清晰度相对较低、镜头抖动、视野较窄等。

2 小视频分析技术面临的挑战

学术界对视频内容分析技术已进行大量且系统的深入研究。例如,针对视频盗用转载和重复出现问题的视频拷贝检测技术,对视频进行分割以提取感兴趣或关键场景的镜头分割技术,对视频中主要物体进行检测、分类和追踪的语义提取技术等。其中,小视频语义提取是最受关注的技术,是后续各种应用的基础。

在对小视频中的主要物体进行语义抽取时,涉及的技术模块主要包括视频目标检测、多目标追踪、人物重识别(也称 Person ReID)等。视频目标检测是指,从视频图像帧中自动定位事先定义好的类别集合中的物体,并推断其类别。多目标追踪是指,利用目标的外观特征和位置信息来将相邻帧中的相同目标关联起来,以构成目标序列,实现对目标的持续追踪。人物重识别是指,在多个非重叠摄像头拍摄的场景下,在一段视频或者某个图片集合中筛选出感兴趣的人物。当然,重识别技术也可以用于筛检某一动物、某一物品等。

2.1 小视频目标检测

目前,人们对视频目标检测的研究主要集中在类似 ImageNet VID^[3](VID

指视频目标检测)的数据集上。这些数据集合往往包含相对较少的物体类别,背景相对简单,前景物体容易与背景区分。小视频场景下的目标检测任务面临的巨大挑战具体包括:(1)类别繁多。小视频中出现的物体类别数以万计,且物体类别的分布呈现长尾效应。大量物体类别严重缺乏训练数据,极大地影响了目标检测算法的性能。(2)剪辑与特效带来较大干扰。镜头切换和视频特效使得物体外观信息被严重干扰,前后帧中主要物体的外观连续性被严重破坏。(3)背景复杂、物体运动难预测。小视频来自用户上传,其背景和人物姿态变化往往更复杂。

2.2 小视频多目标追踪

考虑到业界的实际需求,传统的多目标追踪任务主要聚焦于交通监控等应用场景中对行人和车辆的追踪。这导致目前学术界广泛研究的数据集更多是通过监控设备来采集的,并且主要针对行人目标进行追踪。目前,多目标追踪算法解决的焦点主要是监控场景中的常见问题,如行人目标密集、遮挡等。

在小视频场景中,多目标追踪任务面临前所未有的挑战。与交通监控场景相比,小视频创作偏爱近景。人物在视频上占据区域较大,很难被简单地视为刚体。人物姿态变化直接影响追踪效果。除此以外,频繁的镜头切换也打破了物体帧间位置连续性的假设。

因此,小视频目标追踪任务面临的挑战可归纳为:(1)镜头切换。这使得时空连续性只能在局部窗口内有效。(2)场景不确定性。目标的距离、大小难以预测,很难依据先验信息进行算法性能优化。(3)制作特效问题。小视频有电脑特效或叠加字幕,这给

目标追踪带来很多干扰。

2.3 目标重识别

通用目标的重识别是一个十分困难的研究课题,主要是因为每类目标的特征各不相同。在对小视频分析时,我们通常从人物等特定类别目标重识别开始研究,而这面临的挑战包括:每个镜头中人物的入镜区域存在很大不同,上一个镜头出现的是一个完整的人物,下一个镜头中可能只有上半身入镜;人物在小视频画面中的复杂运动姿态与传统监控画面中的行走姿态有很大差别。这些挑战使得小视频场景下的目标重识别与相机固定监控场景下的行人重识别有很大的不同。

针对小视频场景的人物重识别任务主要包括两点:(1)视频内人物重识别。根据某段小视频前几帧出现的主要人物目标,将后续帧出现的相同人物目标与之一一匹配起来。这类任务的挑战主要是人物局部入镜、姿态变化大、遮挡情况复杂多样(如障碍物遮挡、人物相互遮挡、随机字幕遮挡)。(2)视频间的人物重识别。根据(1)中得到的某个人物图片序列,搜寻其他小视频中出现的相同着装的该人物。这类任务的挑战主要是解决人物着装变化、背景风格差异大、面部遮挡模糊等问题。

2.4 算法性能需求

(1) 计算速度

对于现有海量规模的小视频数据,如果算法处理不够快,对用户请求的响应不及时,用户的使用体验将极大降低。以小视频搜索为例,如果搜索算法能为用户即时提供新的热点视频,用户体验无疑将会得到提升。

(2) 算法精度

由于小视频包含的物体种类繁多,且姿态外观等变化较大,如果分

析算法的精度不够高,用户体验将受到显著影响。这对小视频内容分析算法提出了很高的要求,即必须在面临各种挑战的情况下保持稳定且很高的精度,才可获得良好的应用效果。

(3) 泛化能力

小视频类别很多,其包含的物体类别也是开放的,这对分析技术的泛化能力提出更高要求。小视频分析算法只有具备了良好的泛化能力,才能很好地适应各种应用场景,从而才能真正满足用户时刻变化的应用需求。

3 小视频分析技术研究进展

本章分别从小视频分析任务涉及的技术研究进展,和针对第2章所述的小视频数据特殊难点的解决方案出发,对相关方法进行详细介绍。

3.1 视频目标检测

目标检测从计算机视觉兴起时便一直是基础性的研究任务。随着2015年面向视频目标检测任务的数据集ImageNet VID的发布,深度学习在目标检测研究中开始发挥巨大作用。当前学术界主流研究思路有:

(1) 将检测与追踪相结合

基于检测与追踪结合的方法在图像级别的目标检测结果的基础上,辅以目标追踪方法来将各帧中相同物体的检测框关联起来。2017年由KANG K.等提出具有卷积神经网络的小管(T-CNN)^[4]的方法,通过图像目标检测器对输入视频完成目标检测,再通过目标追踪算法得到目标的检测框序列。2019年由LUO H.等提出的分布式对象技术(DoT)^[5]框架则进一步地对视频目标检测任务进行有选择性地检测和追踪,充分利用检测算法和追踪算法各自的优点,在速度和质量上取得平衡。

(2) 利用光流信息

光流可描述物体的运动状态和轨迹。2015年和2017年P. FISCHER等分别提出了光流网络(FlowNet)^[6]和FlowNet 2.0^[7],通过卷积神经网络直接计算出光流,用来代替目标追踪模块。ZHU X.等在2017年提出的流引导特整体聚合(FGFA)^[8]算法,利用光流描述的运动轨迹将相邻帧的特征聚合到当前帧的特征上,可得到更鲁棒的物体特征,能明显减少由于视频中物体运动模糊和亮度变化带来的影响。光流适用于对局部时空域内的物体运动进行建模,但难以对全局时空域内的物体特征进行整合。

(3) 利用循环神经网络

视频是一种典型的序列数据,用循环神经网络来对帧序列和物体的运动进行建模是一种常见的选择。2017年,LU Y.等提出关联长短期记忆(LSTM)^[9]结构,对视频目标检测任务中的相邻帧间物体的关联信息进行专门建模。通过与检测网络相结合,该方法可直接回归获得物体的位置和类别,同时还能将物体在不同帧之间的特征在时空上都关联起来,最终可得到融合了时序运动信息的关联特征。然而,这类方法的缺点是大量增加了模型训练难度和计算耗时。

(4) 利用全局帧特征融合

WU H. P.等不仅考虑到从局部时域中提取物体的运动信息,还更加关注物体在全局时域上的时序信息,并在2019年提出了序列级语义聚合(SELSA)^[10]算法。该算法在整个视频的完整序列内提取各帧所有感兴趣区域的特征,通过一个聚类模块和变换模块将不同帧之间具有相似语义信息的候选框匹配,从而得到一个全局时域内综合的特征,随后与各帧中提取得到的局部特征相聚合,可得到一个更鲁棒的特征。CHEN Y. H.等在2020年提出基于记忆增强的全局-局

部整合(MEGA)^[11]算法,同时利用局部时域和全局时域内物体的时序信息,即在局部更加关注物体的运动信息,在全局更加关注物体的外观信息,并将两者结合得到最终的融合特征。

3.2 视频目标追踪

目前,视频多目标追踪主要分为3个模块:目标检测、特征提取/运动预测、亲和力计算与关联。

(1) 目标检测模块

目标检测模块负责提供目标位置信息,并将其作为后续处理的先验信息。检测模块提供位置信息,用于确定目标的外观特征,为运动预测提供目标初始位置信息。针对目标检测的研究已经取得长足进步:从传统的可变形部件模型(DPM)^[12]到深度学习方法,从视觉几何网络(VGGNet)^[13]到最新的高分辨率网络(HRNet)^[14],ImageNet数据集的精度不断被刷新,位置预测方式从一阶段的快速区域卷积神经网络(Faster R-CNN)^[15]到两阶段的YOLOv4(指对象检测算法)^[16],在精度和速度上都取得了巨大突破。

(2) 特征提取/运动预测

特征提取/运动预测模块主要负责从外观特征提取高层语义特征和充分利用运动信息。多目标跟踪算法DeepSort^[17]利用简单残差网络构成的重识别(ReID)模型,大幅度改善Sort^[18]算法的性能。而HRNet等方法则采用姿态评估模型来挖掘目标姿态等更为丰富的信息。在运动预测方法中,目前采用比较多的是简单高效的卡尔曼滤波算法。卡尔曼滤波算法可预测接近匀速直线的运动,也有些方法采用更为复杂的粒子滤波,以拟合目标的复杂运动。

(3) 亲和力计算与关联

亲和力计算模块从物体区域的特征信息中计算出匹配对,即当前检测

区域与预测结果区域之间的相似度, 以此作为依据来进行关联计算。关联模块从相似度矩阵中求解出最佳的匹配方式, 尽量将同一目标的检测区域匹配到对应的轨迹上, 通过关联形成新的轨迹。网络流算法、匈牙利匹配算法、多假设追踪算法等都是通过以降低全局匹配为代价来提升匹配效果的。此外, 基于深度学习的方法也有所进展: 多趟近邻排序 (MPN)^[19] 算法以及深度多目标跟踪 (DeepMOT)^[20] 算法利用卷积神经网络分别模拟传统的网络流算法和匈牙利匹配算法来实现关联匹配, 并取得了出色的效果。

3.3 视频物体重识别

对于小视频场景下的通用物体重识别, 学术界目前还没有找到很好的解决方法。对于复杂场景下的人物等特定物体重识别来说, 我们一般将人物局部入镜的重识别问题定义为局部人物重识别, 即利用局部人物图片来检索其完整的人物图片。此外, 还有不少关于遮挡人物重识别的研究工作, 下面我们将分别进行介绍。

(1) 局部人物重识别

早期处理局部人物重识别的方法是将局部人物图片和完整人物图片缩放到同样尺寸, 这会导致特征不对齐等问题。有的研究则采用滑动窗口方法, 利用局部人物图片大小相同的滑动窗口在完整人物图片上进行区域检索, 找到最相近的区域进行相似度计算。当局部人物图片的宽度大于完整人物图片时, 这类方法就会失效, 同时也耗费了很多计算资源。

为了解决局部人物重识别的问题, HE L. X. 等提出了一种深度空间特征重构 (DSR) 的方法^[21]。该方法首先利用全卷积网络生成固定尺寸的特征图, 然后利用字典学习模型中的重建误差来计算不同特征图的相似度。

SUN Y. F. 等提出一种自监督的方法^[22]来解决局部人物重识别的特征不对齐问题。该方法将图片划分为上、中、下3个抽象模块区域, 得到每个区域中像素点的区域标签, 并以此来训练模型对每个区域的观察能力。在推理阶段, 模型通过预测区域可见得分, 判断图片是否发生了身体部位的缺失, 进而通过自监督的注意力机制实现对人物图片间对应区域的相似度比较。

(2) 遮挡人物重识别

不同于局部人物重识别, 遮挡人物重识别主要的问题在于图片中包含的遮挡区域会使得直接提取的全局特征包含大量的干扰噪声, 进而影响两张图片的相似度计算结果。针对这一点, MIAO J. X. 等^[23]通过引入额外的姿态检测模型来获得人体关键点信息, 进而引导重识别模型关注人物的非遮挡区域。具体思路是, 首先通过关键点的位置信息来提取人物的局部特征, 然后利用关键点的置信度信息来判断哪些关键点是处于遮挡区域的。在重识别的推断阶段, 模型只会计算两张图片未被遮挡的区域之间的相似度, 以此来消除遮挡噪声的干扰。

3.4 针对小视频的研究工作

目前, 学术界专门针对小视频特点的研究工作比较少。本文中, 我们挑选一些比较突出的相关研究工作进行分析介绍。

(1) 针对小视频复杂特效问题的研究

针对不同镜头间添加的视频特效导致物体外观信息不匹配问题, ZHONG Z. 等于2018年在行人重识别领域提出了相机风格自适应^[24]算法。该算法假定, 在不同相机风格下拍摄所得的人物数据属于不同的数据域, 同时通过引入循环生成对抗网络 (CycleGAN)^[25], 对每一对具有不同风格

的同一人物图像, 生成图像到图像的风格转移模型。生成不同相机风格下的人物图像为重识别模型提供额外的训练数据。为了防止重识别模型受到由 CycleGAN 风格转移得到的伪图像中噪声的影响, 算法引入了一个标签平滑修正 (LSR) 机制, 以降低在重识别模型损失函数中对伪图像评判的权重。

(2) 针对小视频物体类别繁多的研究

针对物体类别繁多所带来的长尾分布效应, POOJAN O. 与 VISHAL P. 于2019年在图像分类领域提出了基于多任务的开集物体识别 (MLOSRI)^[26]算法。该算法通过使用权值共享的分类网络和解码网络, 同时进行分类与重构任务。此外, 算法依据极值理论^[27]通过一个极值模型来对重构误差分布的尾部部分建模, 使得模型对未出现在训练集中的类别更为敏感。

(3) 针对小视频镜头切换频繁的研究

针对不同镜头下的物体空间位置变化不连续问题, HSU H. M. 等于2019年在目标追踪领域提出一个多摄像机目标追踪系统^[28], 将多个摄像机下的目标追踪问题划分为镜头内的目标追踪问题和镜头间的目标追踪问题。对于镜头内的目标追踪问题, 该研究团队采用跟踪网络追踪器 (TNT)^[29]。对于镜头间的目标追踪问题, 该研究团队首先将镜头内追踪得到的跟踪片输入到 Mask R-CNN^[30]网络中, 以得到去除背景后的结果, 然后再通过一个时间注意力模型, 对各跟踪片提取跟踪片级别的特征, 最后通过比较特征相似度的方式来匹配不同摄像机下的同一物体。

4 小视频内容分析系统

要系统性完成小视频内容分析任务, 单纯依靠某一个算法模块是困难

的。只有构建一个整合多种不同算法的系统,才能够更准确、更鲁棒地解决分析问题。本文在此抛砖引玉,提出一个小视频内容分析系统的构成框图。结合此前提到的小视频数据的特点,以及当前对于视频分析技术的研究成果,我们认为小视频内容分析系统至少应包括镜头分割、视频目标检测、视频目标追踪、视频目标重识别等模块,如图1所示。

对于输入的小视频,首先,镜头分割模块将不同镜头分割开来,使得每个镜头内物体运动能基本满足帧间位置连续性假设;接着,目标检测模块获得各帧内物体的定位框和物体分类结果,并将结果输入到后续镜头内的目标追踪模块,同时属于同一物体的检测框在相邻帧中将被关联起来;最后,系统再进行跨镜头目标重识别,得到各物体在小视频中完整的时空运动轨迹。小视频内容分析系统的输出结果可被应用到后续更多的应用处理中,例如实现视频结构化、完成以视频搜索视频等任务。

视频结构化应用的主要目标是,

仅从无结构视频数据中解析主要物体的语义属性和时空轨迹等结构化的语义信息,就可以实现人车信息检索以及行为研判等,为交通安全和社会治安提供风险评估和事件预警。以视频搜视频是小视频的一大类应用。常规文字、图片搜索等不能完全满足用户需求,而以视频搜索类似视频的功能在各大应用软件的出现,有助于提升用户体验。小视频内容分析结果使小视频搜索成为可能。此外,小视频查重、溯源等也是类似应用。基于小视频内容分析的各种衍生应用正在日益增多,这将大大改善小视频的用户体验。

5 结束语

小视频应用的兴起是互联网技术发展的必然结果,也是人工智能技术广泛服务人们生活的发展趋势。目前,越来越多的巨头公司和科研机构开始研发小视频内容分析技术,旨在更好地应用人工智能技术分析海量视频数据,以更好地服务社会。随着小视频研究和应用的不断发展,在为受众提供更高服务质量的同时,对小视频数

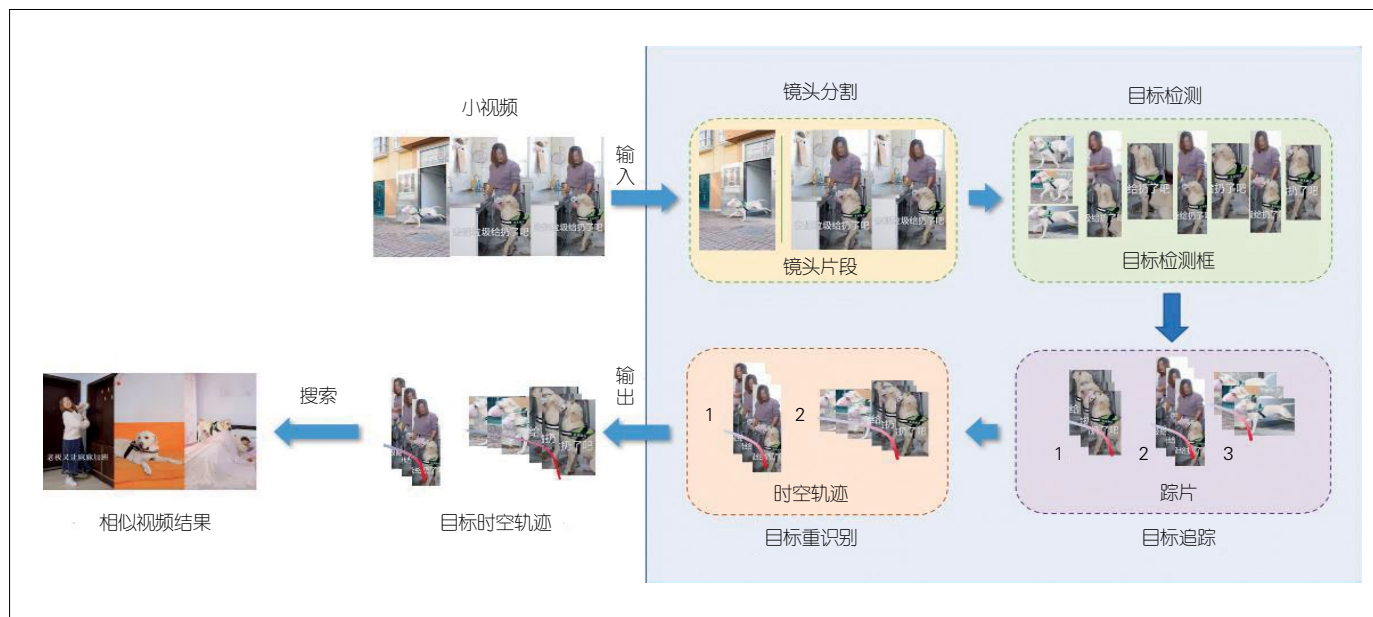
据规范化利用、确保个人隐私和数据安全,正在成为社会大众非常关注的热点问题。

致谢

感谢复旦大学计算机科学技术学院邱泰儒、徐僖禧、王浔彦、陈冠先等为本文写作而做出的大量贡献。

参考文献

- [1] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database [C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL, USA: IEEE, 2009: 248–255. DOI: 10.1109/cvprw.2009.5206848
- [2] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]//European conference on computer vision. Zurich, Switzerland: Springer, 2014: 740–755. DOI: 10.1007/978-3-319-10602-1_48
- [3] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge [J]. International journal of computer vision, 2015, 115(3): 211–252. DOI: 10.1007/s11263-015-0816-y
- [4] KANG K, LI H S, YAN J J, et al. T-CNN: tubelets with convolutional neural networks for object detection from videos [J]. IEEE transactions on circuits and systems for video technology, 2018, 28(10): 2896–2907. DOI: 10.1109/tcsvt.2017.2736553



▲图1 小视频内容分析系统构成框图

- [5] LUO H, XIE W X, WANG X G, et al. Detect or track: towards cost-effective video object detection/tracking [C]//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, HI, USA: AAAI, 2019, 33: 8803–8810. DOI: 10.1609/aaai.v33i01.33018803
- [6] DOSOVITSKIY A, FISCHER P, ILG E, et al. FlowNet: learning optical flow with convolutional networks [C]//2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile. IEEE, 2015: 2758–2766. DOI: 10.1109/iccv.2015.316
- [7] ILG E, MAYER N, SAIKIA T, et al. FlowNet 2.0: evolution of optical flow estimation with deep networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 2462–2470. DOI: 10.1109/cvpr.2017.179
- [8] ZHU X, WANG Y, DAI J, et al. Flow-guided feature aggregation for video object detection [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 408–417
- [9] LU Y, LU C, TANG C K. Online video object detection using association LSTM [C]//Proceedings of the IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2344–2352
- [10] WU H P, CHEN Y T, WANG N Y, et al. Sequence level semantics aggregation for video object detection [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019: 9217–9225. DOI: 10.1109/iccv.2019.00931
- [11] CHEN Y H, CAO Y, HU H, et al. Memory enhanced global-local aggregation for video object detection [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 10337–10346. DOI: 10.1109/cvpr42600.2020.01035
- [12] FELZENSZWALB P, MCALLESTER D, RAMANAN D. A discriminatively trained, multi-scale, deformable part model [C]//2008 IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, USA: IEEE, 2008. DOI: 10.1109/cvpr.2008.4587597
- [13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2020–12–05]. <https://arxiv.org/abs/1409.1556v1>
- [14] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 5693–5703. DOI: 10.1109/cvpr.2019.00584
- [15] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 39(6): 91–99. DOI: 10.1109/tpami.2016.2577031
- [16] BOCHKOVSKIY A, WANG C Y, LIAO H M. YOLOv4: Optimal speed and accuracy of object detection [EB/OL]. [2020–12–05]. <https://arxiv.org/abs/2004.10934>
- [17] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric [C]//2017 IEEE International Conference on Image Processing (ICIP). Beijing, China: IEEE, 2017: 3645–3649. DOI: 10.1109/icip.2017.8296962
- [18] BEWLEY A, GE Z, OTT L, et al. Simple online and realtime tracking [C]//2016 IEEE International Conference on Image Processing (ICIP). Phoenix, AZ, USA: IEEE, 2016: 3464–3468
- [19] BRASÓ G, LEAL-TAIXÉ L. Learning a neural solver for multiple object tracking [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 6247–6257
- [20] XU Y H, SEP A, BAN Y T, et al. How to train your deep multi-object tracker [C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 2020: 6787–6796. DOI: 10.1109/cvpr42600.2020.00682
- [21] HE L X, LIANG J, LI H Q, et al. Deep spatial feature reconstruction for partial person Re-identification: alignment-free approach [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT: IEEE, 2018: 7073–7082. DOI: 10.1109/cvpr.2018.00739
- [22] SUN Y F, XU Q, LI Y L, et al. Perceive where to focus: learning visibility-aware part-level features for partial person Re-identification [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 393–402. DOI: 10.1109/cvpr.2019.00048
- [23] MIAO J X, WU Y, LIU P, et al. Pose-guided feature alignment for occluded person Re-identification [C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, South Korea: IEEE, 2019: 542–551. DOI: 10.1109/iccv.2019.00063
- [24] ZHONG Z, ZHENG L, ZHENG Z D, et al. Camera style adaptation for person Re-identification [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 5157–5166. DOI: 10.1109/cvpr.2018.00541
- [25] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2223–2232. DOI: 10.1109/iccv.2017.244
- [26] OZA P, PATEL V M. Deep CNN-based multi-task learning for open-set recognition [EB/OL]. [2020–12–05]. <https://arxiv.org/abs/1903.03161>
- [27] DE HAAN L, FERREIRA A. Extreme value theory: an introduction [M]. Springer Science & Business Media, 2007
- [28] HSU H M, HUANG T W, WANG G, et al. Multi-camera tracking of vehicles based on deep features re-ID and trajectory-based camera link models [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: IEEE, 2019: 416–424
- [29] WANG G A, WANG Y Z, ZHANG H T, et al. Exploit the connectivity: multi-object tracking with TrackletNet [C]//Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM, 2019: 482–490. DOI: 10.1145/3343031.3350853
- [30] HE K M, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision (ICCV). Venice, Italy: IEEE, 2017: 2961–2969. DOI: 10.1109/iccv.2017.322

作者简介



薛向阳, 复旦大学计算机科学技术学院教授、博士生导师; 主要从事计算机视觉、视频大数据分析、机器学习等研究; 发表论文 200 余篇, 其中 90 余篇发表在《IEEE Transactions on Pattern Analysis and Machine Intelligence》《IEEE Transactions on Image Processing》等) 和顶级国际会议 (如 ICCV、CVPR、ICML、NeurIPS、ACM MM、IJCAI、AAAI 等) 上。



李斌, 复旦大学计算机科学技术学院青年研究员、博士生导师, 上海高校特聘教授 (东方学者); 研究领域为机器学习、类脑人工智能及其在机器视觉与大数据分析中的应用; 在《IEEE Transactions on Knowledge and Data Engineering》《IEEE Transactions on Cybernetics》等知名期刊与 ICML、NeurIPS、IJCAI、AAAI 等一流机器学习和人工智能会议上发表论文 60 余篇。



构建智能实时网络， 使能 5G 视频业务繁荣

Building Smart Real-Time Networks to Enable Prosperity of 5G Video Services

吕达 /LYU Da

郑清芳 /ZHENG Qingfang

(中兴通讯股份有限公司，中国 深圳 518057)
(ZTE Corporation, Shenzhen 518057, China)

摘要：5G 将促进视频业务的大繁荣，包括极大地改善现有的视频业务体验和催生新型的视服务形态。为应对 5G 视频业务所面临的超低时延、高可靠及高体验质量等方面的挑战，中兴通讯提出构建智能实时视频网络（SmartRTN）的理念，并围绕这一理念，创新性地研发出一系列技术和方案，包括基于内容智能分析的低码高清视频编码技术、超低时延的网络传输、基于深度学习的内容处理与增强、结合边缘计算以及网络切片的组网方案和智能调度策略等。这些技术和方案被应用于视频业务端到端各个环节，有效地解决了困扰 5G 视频业务发展的技术瓶颈问题。

关键词：低码高清；实时通信；超分辨率；智能调度；体验质量

Abstract: 5G is expected to bring prosperity of video applications, including significantly improving existing applications and bringing forth new exciting applications. To meet the challenges of 5G video applications, such as ultra low latency, high reliability and high quality of experience, ZTE proposes the concept of constructing smart real time video network (SmartRTN). Based on this concept, ZTE innovatively develops a series of technologies and solutions, including low bitrate high quality video compression based on content intelligent analysis, ultra low latency video transportation, smart video processing and enhancement based on deep learning, networking solutions and intelligent scheduling strategies combining edge computing and network slicing. These technologies and solutions have been applied in all end-to-end video service processes, effectively solving the technical bottlenecks that beset 5G video service development.

Keywords: low bitrate high quality; real time communication; super resolution; smart scheduling; quality of experience

DOI: 10.12142/ZTETJ.202101013

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210126.1653.002.html>

网络出版日期: 2021-01-27

收稿日期: 2020-12-08

1 5G 视频业务概述

1.1 5G 促进视频业务持续发展

5G 具有大宽带、低时延的特点，它解决了视频业务发展的关键瓶颈问题，大大促进了视频业务的发展。5G 不仅使传统视频业务，如安防、视频会议、点播、直播等，获得了迅速发展，还使由视频业务衍生的远程教育、远程医疗等远程交互业务也获得了巨大发展。更进一步地，面向家庭

和娱乐场景的超高清视频、沉浸式视频、全景视频、3D 视频也获得了高速发展的机会。

根据 Cisco 可视化网络指数（VNI）预测，到 2022 年，全球互联网协议（IP）视频流量将占总流量的 82%，如图 1（a）所示。所有形式的 IP 视频（包括互联网视频、IP 视频点播、视频流游戏、视频会议和基于文件共享的视频文件）的总和将继续保持在总 IP 流量的 80%~90%。2017—2022 年，全球

视频流量的复合年增长率为 26%。随着网络的广泛部署以及市场竞争的发展，移动视频业务发展迅猛，2022 年移动视频流量将占据总移动数据业务的 79%，并保持 46% 的年复合增量率，如图 1（b）所示。

1.2 新型视频业务下的端到端技术指标

视频业务的形态不断增加，对端到端的技术指标提出差异化要求：准实时直播时延的可接受范围为 1~3 s；实

时互动直播的时延要控制在 500 ms 以内；视频会议要求端到端时延需要在 200 ms 以内、编码时延需在 100 ms 以内、操作指令时延在 30 ms 以内；而对于实时性要求较强的增强现实（AR）/虚拟现实（VR）业务及云游戏业务，端到端时延一般需要控制在 100 ms 以内、编码时延需要在控制在 10 ms 以内。

1.3 5G 视频业务端到端的质量仍需提升

5G 给网络状况带来的提升只是视频业务繁荣的必要非充分条件，我们还必须从视频业务端到端全流程的角

度来设计完整的技术体系。5G 只是为视频的高效传输提供底层网络支撑。如何协同利用人工智能（AI）、云计算、边缘计算等新技术，来构建端到端的视频技术体系，以及如何从智能实时网络、智能化处理、端云边协同高性能计算及存储、智能部署等多个角度、业务全流程，来提升视频业务采集、预处理、编码、传输、解码、后处理、渲染各个环节的处理效率和业务质量，以给用户提供更清晰、高流畅度、低时延、强交互的极致用户体验，是中兴通讯正在努力的方向。

2 中兴通讯打造下一代智能实时网络（SmartRTN）

中兴通讯基于多年技术积累和产品研发工作，从信源、信道、用户体验、业务部署及运维等多方面综合考虑，提出通过构建智能实时视频网络来使能视频业务繁荣的理念。围绕这一理念，中兴通讯创新性地研发出一系列技术，并将这些技术成功应用于视频业务端到端各环节，例如：

（1）在信源方面，中兴通讯结合业界最新视频编码标准的进展，通过对视频内容的智能分析，合理地分配码率，尽可能在保证较高画质体验的前提下提升数据压缩比；进一步地引入基于 AI 的图像生成技术，使特定场景内容（如人脸等）取得了极致的压缩比。

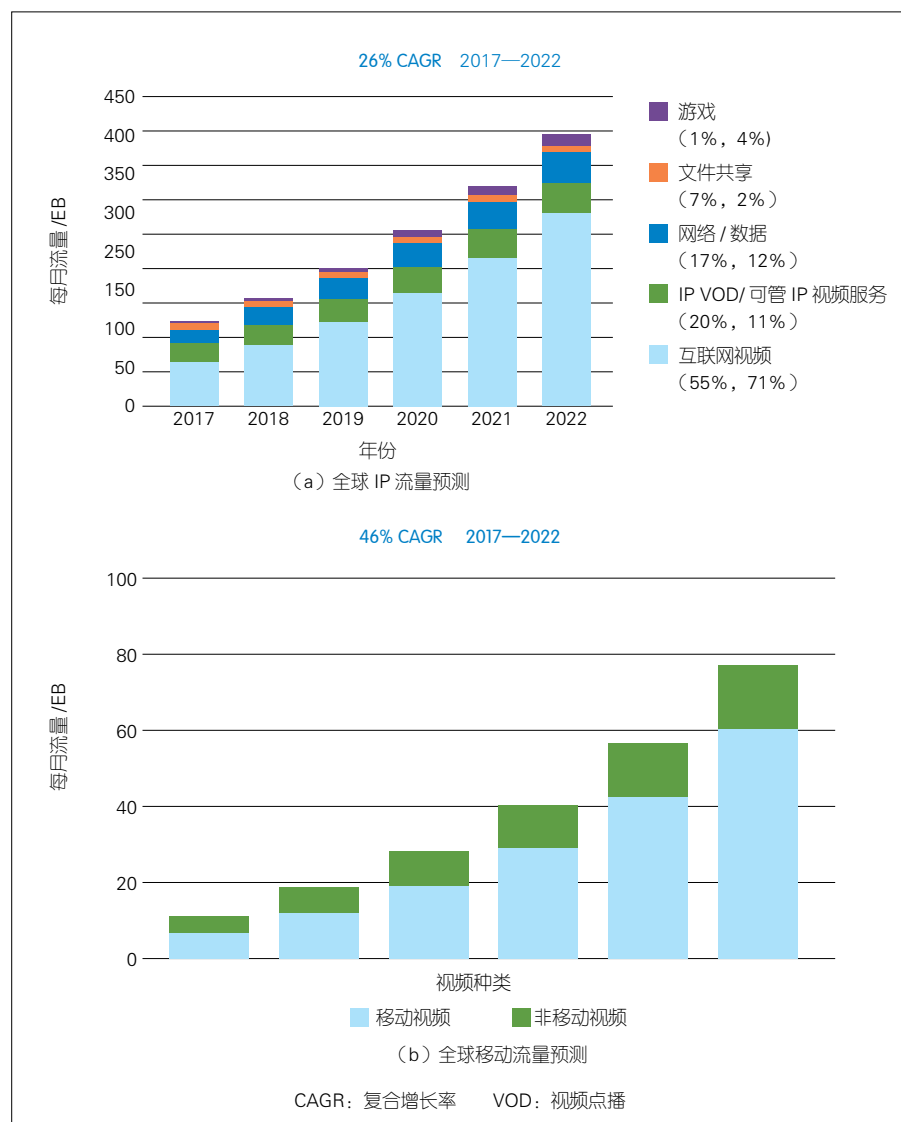
（2）在信道方面，中兴通讯自主研发传输协议，通过控制视频编码与传输之间的协同机制，有效降低了传输时延。精巧设计的抗丢包策略，实现了弱网环境下的可靠传输。

（3）在用户体验方面，中兴通讯研发了一系列技术，以对不同环节予以改善。例如，在成像环节，增强在不同光照条件下的画面清晰度；在显示环节，通过虚拟背景技术保护用户隐私；在会议场景中，通过对人脸以及人物动作的识别，使会场管理更加便利。

（4）在业务部署和运维方面，中兴通讯借助 5G 的网络切片，实现了用户服务质量（QoS）差异化保障；使用边缘计算，实现业务的就近接入和媒体的下沉处理；通过智能路由，实现最优路径的选择；通过智能用户体验质量（QoE）检测，及时发现故障并无感修复。

2.1 低码高清

视频低码高清是指在保证视频画



▲图 1 Cisco 可视化网络指数全球 IP 流量预测和全球移动流量预测（2017—2022 年）^[1-2]

面质量的前提下, 尽可能提升压缩比、降低视频码率, 它可以从视频编码、视频前后处理等多个维度进行提升。视频编码主要分为基于现有成熟编码的优化和新一代编码技术的引入。

2.1.1 挖掘现有视频编解码标准的最大潜力

基于目前产品广泛使用的 H.264/H.265 编码, 我们实现针对不同场景的内容感知编码 (CAE) 优化:

(1) 基于感兴趣区域 (RoI) 编码优化

在典型的视频通信场景中, 人们的主要关注点在于人脸及周边区域, 而非背景区域。如图 2 所示, 在视频通信发送端引入实时人脸检测和基于 RoI 的编码算法, 并对不同区域设置不同的码率, 可使最终实现的 RoI 编码在保持画面主观质量不下降的前提下, 实现 20% 的码率节省。

(2) 基于屏幕内容特性的压缩编码优化

无论是视频通信还是云电脑的应用, 视频的内容来源主要包括两类: 屏幕内容分享和摄像头视频。屏幕内容和摄像头采集生成的视频内容有本质差别, H.265 已有专门针对屏幕内容的高效视频压缩编码 (HEVC) - 屏幕图像编码 (SCC) [3] 压缩标准。如图 3 所示, 考虑到现有大规模部署的 H.264 系统, 针对视频会议的辅流文档共享、云电脑的屏幕内容分享场景, 中兴通讯对屏幕内容进行分类压缩, 采用调色板、文字特征提取等压缩方式, 在确保文字区域无损清晰的前提下, 使图像传输带宽降低 10% 以上。

(3) 基于动态帧率的编码优化

在视频通信或云电脑的实际使用场景中, 经常会出现阶段性画面无变化的情形, 比如, 在视频交互通信中播放幻灯片 (PPT) 文档内容、云电脑

用户操作不太频繁。动态帧率的编码优化能够根据场景的运动剧烈程度来动态实时调整帧率, 比如在 PPT 分享或屏幕应用静止时, 可以通过自动降低帧率实现至少 10% 的综合带宽降低效果。

2.1.2 研发新一代编码技术

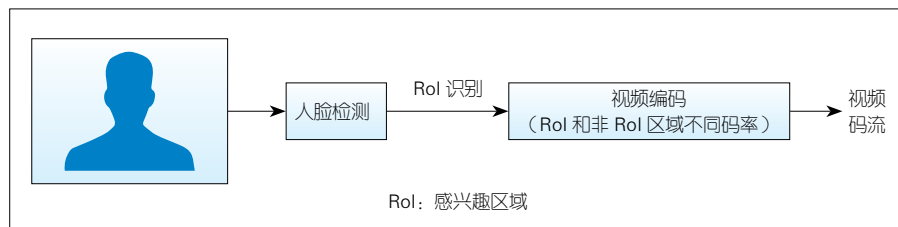
除了前述基于现有 H.264/H.265 进行码率、帧率等方面的编码优化外, 中兴通讯还积极参与研发新一代视频编解码技术。目前, 全球最新视频 Codec 标准主要以多功能视频编码 (VVC, 也称 H.266) [4]、开放媒体联盟视频标准 (AV1) [5] 和第 3 代数字音视频编解码技术标准 (AVS3) [6] 为主流,

同时基本视频编码 (EVC) [7] 和低复杂度增强视频编码 (LCEVC) [8] 针对特定场景 (如降低编码复杂度、充分利用现有硬件等) 也有一定的应用空间。部分最新视频编码码率降低效果对比结果具体如图 4 所示。

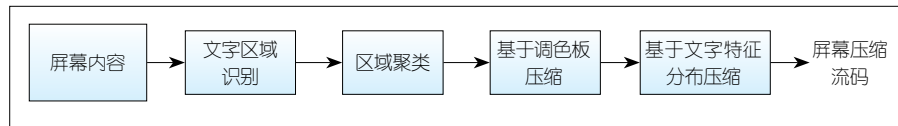
中兴通讯持续参与 HEVC、VVC 标准的制定工作, 并在当前动态图像专家组 (MPEG) 的两个特别工作组 (AHG) 中担任领导职位。

2.1.3 AI 进一步提升压缩比

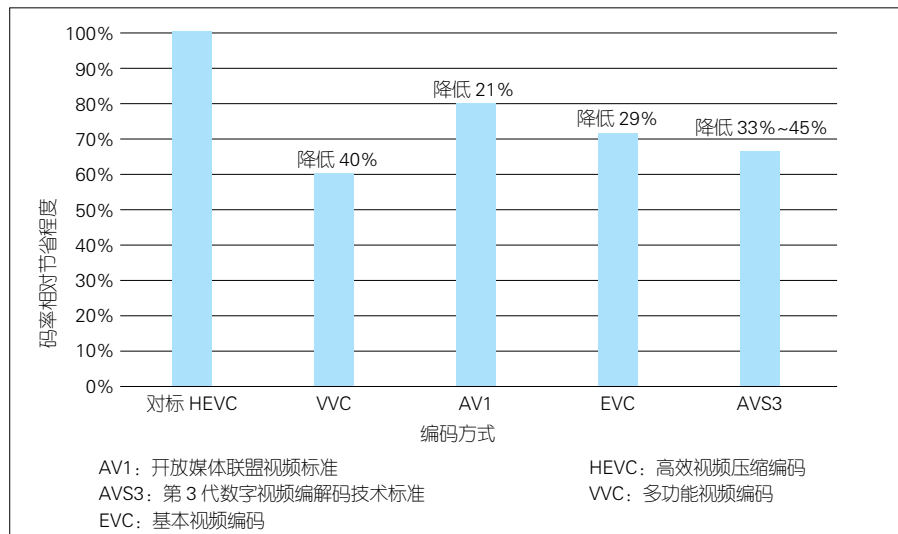
AI 在各个领域中的应用得到了迅猛发展, 并在特定的业务场景中, 带来了新的解决方法。关注用户真正的场景需求有可能颠覆传统的视频编



▲图 2 感兴趣区域编码示例



▲图 3 中兴通讯针对屏幕内容的文字压缩技术



▲图 4 最新 Codec 码率降低对比图 (相对 HEVC)

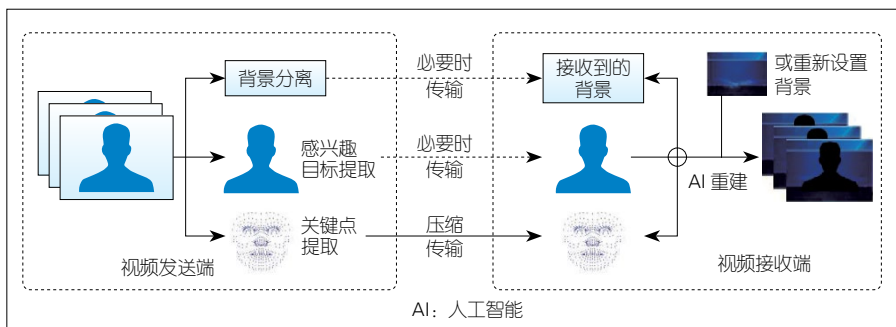
解码技术, 并带来极致的压缩比。例如, 在 SmartRTN 网络中, 针对个人视频通信这种场景, 传输的视频帧主要由变化很小的背景图片和运动的人脸构成, 用户的关注点主要是表情的交流。基于生成式对抗网络 (GAN) 的人脸生成技术可以对摄像头获取的内容的关键信息进行重构, 以形成新的解决方案, 具体如图 5 所示。与传统的基于像素域的信号处理技术相比, 中兴通讯提出基于感兴趣目标和稀疏关键点检测的方法, 对视频信息进行编码。基于运动驱动感兴趣目标, 并结合该场景下背景信息重构压缩后的视频帧, 该方法使码率得到了更加有效的压缩——可以达到传统算法码率的 1/10。同时, 重构后的视频可以任意切换光照模式和视角, 在虚拟会场中可以实现统一的光照模式及物体的任意视角, 为个人视频通信业务提供更具沉浸式的临场感和更加真实的眼神交流体验。

2.2 超低时延传输

2.2.1 不同视频传输协议

针对不同的视频业务场景, 常见的视频传输协议有实时消息传输协议 (RTMP)、通用媒体应用格式 (CMAF)、低时延 HTTP 实时流媒体 (LHLS)、Web 实时通信 (WebRTC) 等, 其技术特性对比如表 1 所示。

为了满足实时音视频通信对低时延传输的需求, 中兴通讯研发了超低时延传输协议, 在传输层参考快速用户数据报网络连接 (QUIC)^[11] 协议的基础上做了大量的重新设计, 例如加密机制、多路径、前向纠错 (FEC) 支持、优先级管理、可配置的拥塞控制算法等, 以满足实时流传输场景的需求。在媒体传输的应用层上, 中兴通讯研发的协议与 WebRTC、RTMP 等协议兼



▲图 5 中兴通讯基于人工智能的编码解决方案

▼表 1 不同传输协议对比

协议	理想延迟区间/s	降低时延机制	应用场景
RTMP/FLV	1~2	流式传输 + 协议栈优化	泛娱乐直播
CMAF ^[9]	2~3	索引预生成及 HTTP 编码块传输	OTT、体育直播
LHLS ^[9]	2~3	索引预生成及 HTTP 编码块传输	低时延直播
RTP/WebRTC ^[10]	0.2~1	链接及传输协议优化, 尽可能低的抖动缓冲区 (jitterbuffer)	超低时延直播、视频通话

CMAF: 通用媒体应用格式

FLV: Flash 视频格式

HTTP: 超文本传输协议

LHLS: 低延迟 Http 实时流媒体

OTT: 指越过运营商的互联网 (视频) 业务

RTMP: 实时消息传输协议

RTP: 实时传输协议

WebRTC: Web 实时通信

容, 可适配各种不同的实时音视频应用场景的需求。

2.2.2 融合编码和传输技术

实时通信系统需要考虑视频编码器和传输协议的协同控制。传输协议和编解码器不同步或网络条件不稳定, 容易引发延迟现象或故障。

(1) 有两份视频编码时, 选择最合适的一份以避免拥塞。

斯坦福的 Salsify 项目^[10]创新性地体现了新的组合方式——编解码器速率控制和传输拥塞控制。Salsify 的编解码器可保证发送者不会在网络拥塞时发送帧 (必要时会丢弃已经编码的帧), 且不固定帧的发送速率。同时, 编解码器还可被允许生成更接近可用网络容量的帧, 且生成每个帧的两个版本: 一个质量略高于先前的成功案例, 另一个则质量略低。应用程序在查看每个选项的实际压缩大小后, 从这些选项中进行选择 (或不选)。官方的测试结果表明^[11], Salsify 比现有

的商业系统 (如 Skype、FaceTime 和 WebRTC) 在时延控制和视频质量上更为优秀。

(2) 采用编码与传输的管道机制, 边编码边传输

音视频采集、编码、传输、解码、渲染等流程是相互联动和影响的。采集、编码与传输形成管道, 可以有效降低时延。例如, 视频编码编完一个切片后, 在编下一个切片的同时, 可传输刚编完的切片数据; 若采用 SVC 或 LCEVC 编码, 则可以编完一个层, 且在编下一个层的同时, 立刻传输已编完的层数据。

2.2.3 拥塞控制技术

实际的网络状态是复杂多变的, 丢包、延时和网络带宽都在时刻变化, 这就对网络拥塞控制算法提出了很高的要求。网络拥塞是指发送的数据超过了网络所能承载的传输能力。尽管基础通信设施在不断地完善, 网络拥塞的情况在 5G 时代还是有可能出现。

针对实时音视频传输的拥塞控制, 中兴通讯提出适应多场景的拥塞控制模块, 包括传统的基于传输控制协议 (TCP) 的瓶颈带宽和往返时延 (BBR)^[12]、基于用户数据报协议 (UDP) 的谷歌拥塞控制 (GCC)^[13] 和基于机器学习的拥塞控制功能。这些拥塞控制模块可以被选择部署在云端或者集成在发送端。

(1) 针对视频专网等高可靠环境, 通信双方可以采用 TCP 方式传输实时音视频数据。此时发送端自动采用基于 TCP 的控制模块。目前主要采用的拥塞控制算法是 BBR 系列。

(2) 对于弱网不可靠环境, 通信双方采用 UDP 方式传输实时音视频数据, 发送端则自动采用 UDP 系列的控制算法, 如 GCC。

(3) 另外, 中兴通讯提出的拥塞控制模块还包括支持基于大数据驱动的智能拥塞控制决策模块。该模块通过收集发送端、传输网络、接收端等多方的信息, 形成对网络拥塞程度的预测, 从而推动发送端选择不同的编码参数、不同的传输协议、拥塞控制参数 (详细技术原理可参考本文 2.4.3 节)。

2.2.4 FEC、自动重传请求 (ARQ) 等弱网对抗技术

FEC 也叫前向纠错码, 是视频业务系统网络保证可靠传输质量的重要方法。FEC 可以对 n 份原始数据增加 m 份数据, 并能通过 $n+m$ 份中的任意 n 份数据, 还原原始数据, 即如果有任意小于等于 m 份的数据失效, 仍然能通过剩下的数据还原出来。当前的 FEC 算法使用范特蒙矩阵或者柯西矩阵, 来实现纠错码的功能。通过在传统 FEC 算法上做自适应改进, 中兴通讯的视频 FEC 方案可根据网络条件, 实现延时自调整、网络自适应、冗余自增减等功能。

ARQ 也是抵抗网络丢包的一种重要手段。中兴通讯视频系统使用的是基于否定确认包 (NACK) 的丢包重传技术。NACK 是一种通知技术, 其触发通知的条件刚好与确认包 (ACK) 相反。在未收到消息时, NACK 通知发送方“我未收到消息”, 即通知未达。NACK 在接收端检测到数据丢包后, 发送 NACK 报文到发送端。发送端根据 NACK 报文中的序列号, 在发送缓冲区找到对应的数据包, 并将其重新发送到接收端。ARQ 和 FEC 配合使用, 可以在不大幅增加网络冗余的条件下, 实现较好的抗丢包效果。在实际应用中, 中兴通讯视频系统能抵抗 80% 的网络丢包, 满足 95% 以上的使用场景。

2.3 视频智能分析

2.3.1 暗景增强实现低光照下的视频画质提升

在视频通信场景中, 由于场地变换、光照摄像头角度变化等因素, 通常会出现由关键人脸部分光照不均匀导致的暗影现象, 这影响了用户体验。中兴通讯通过对大量 3D 人脸在不同光照模式下的数据进行模拟训练, 实现了基于 2D 图像对光照条件的预测, 并通过光照条件的映射实现了自然光照场景下人脸图像的非线性变换模拟, 使之达到了光照均匀的效果, 提升了暗光场景下的人脸画质。

2.3.2 人像分割及背景虚化

视频通信可以随时随地通过移动终端接入。虽然这极大地方便了客户使用, 但同时也导致客户个人私密信息出现在视频中。因此, 基于语义分割的背景和背景虚化功能就成为了视频通信产品不可或缺的功能。

中兴通讯基于神经网络架构搜索技术构建了轻量级模型, 在自收集的

Portrait 数据集上进行训练, 实现了端侧的语义分割算法, 并通过网络模块轻量化设计、模型剪枝及模型蒸馏等提速方案, 得到了 300 kB 大小的轻量级语义分割模型。通过端侧的部署加速和前后端处理的项目流程优化, 我们在骁龙 845 手机芯片上实现了高达 33 帧/秒的实时推理过程。

背景替换和虚化技术是基于实时人像分割技术的应用。在使用轻量化深度神经网络对输入图完成人像分割任务之后, 所得的人像分割网络输出背景为 0、人像为 1 的图像, 并与输入图进行相乘可保留人像信息。背景替换的图片可以首先将网络输出的图像取反, 然后进行相乘生成替换的背景图像, 最后将人像信息和背景图像合成一张图片, 即可得到所需的背景替换, 具体如图 6 所示。

2.3.3 人脸识别

中兴通讯基于大规模私有人脸数据集、深度卷积神经网络的人脸特征编码模型以及度量学习方法, 在人脸识别领域有着长期的技术积累。特别地, 在视频人脸识别处理中, 中兴通讯提出综合视频空域信息的代表帧融合和特征增强方法, 相应的处理流程如图 7 所示。表 2 给出了中兴通讯视频人脸识别方案在标准测试集 YouTube Faces 上的准确率比较。该方法大大提高了人脸特征的泛化性, 同时提高了对运动/失焦模糊、低分辨、视频编解码噪声的耐受力, 并在多个开源测试集上达到了较高的准确率。

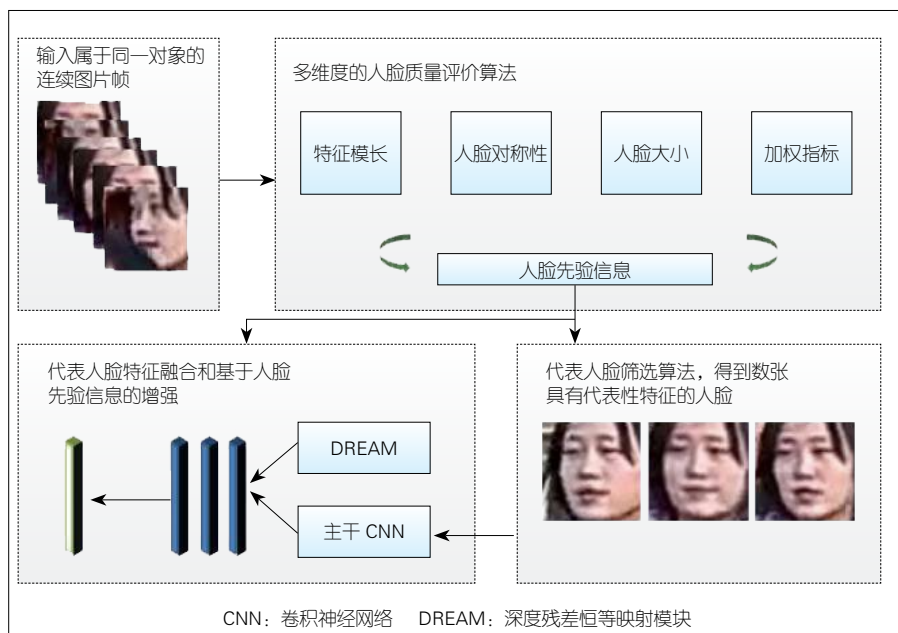
2.4 智能部署

2.4.1 融合移动边缘计算 (MEC) 和网络切片

基于 5G 端到端网络切片技术, 对专用网络进行优化, 可实现视频服



▲图6 暗景增强、背景虚化、背景替换效果对比



▲图7 中兴通讯视频人脸识别处理流程^[14]

▼表2 中兴通讯视频人脸识别方案的准确率比较

测试方法	准确率 / %	所用帧数	融合方法
代表帧 + 特征增强	95.98	3	平均池化
特征融合	95.40	1	/
代表帧	95.80	3	平均池化
单帧	95.04	1	/
FaceNet ^[15]	95.12	100	平均相似度
DAN ^[16]	94.28	所有帧	平均池化
DeepFace ^[17]	91.40	100	平均池化
NAN ^[18]	95.72	所有帧	加权池化
C-FAN ^[19]	96.50	所有帧	加权池化

C-FAN: C-特征聚集网络 DAN: 判别聚集网络 NAN: 神经聚集网络

务加速、视频服务网络与其他网络业务隔离服务, 解决网络拥塞和时延问题。支持 5G 接入侧的 MEC 视频服务下沉, 不仅可实现媒体就近接入、就

近处理, 为用户带来更低时延的视频体验, 还可同时降低对骨干网带宽占用。更进一步地, 融合 5G 网络切片和 MEC 可对基站、频率专享等组成 5G

虚拟专网, 可以满足高端客户的高安全、高可控、高性能要求, 如图 8 所示。

2.4.2 智能路由调度

由于 RTN 网络服务用户的网络条件和质量各异, 基于强大的大数据分析和 AI 预判能力的支持, 中兴通讯实现了实时的智能路由调度, 具体如图 9 所示。针对统一接入调度模块, 用户侧接入调度除了选择就近边缘接入外, 核心的网络路由可以选择进行如下操作:

(1) 基于大数据提取多维度网络路由质量评价指标, 生成当前路由优劣评分;

(2) 基于现有的评价模型, 实现了未来 5~10 min 内网络质量的预判;

(3) 实时统计网络各个节点、不同粒度的质量参数 (如带宽、往返时延等), 并综合前两者的评分结果, 实现当前路由表的实时动态调整。

另外, 为了保证低时延, 网络架构设计与传统的内容分发网络 (CDN) 分层设计稍有差异。其中, 核心中继服务器采用扁平 Mesh 组网架构, 内部链路更短、更灵活, 可支持采用动态选路的方式来调整构建的网状结构。中继服务器之间采用优化过的 QUIC 协议实现数据传输, 使内部链路延迟达到 30 ms 左右。

2.4.3 智能 QoE 监测

实时视频服务的 QoE 受到实时音视频采集、前处理、编码、传输、解码、后处理、渲染各个环节的影响。一旦某一个环节出现问题, 如传输过程中的网络丢包、采集环节中的系统不兼容, 都会直接导致实时音视频服务出现质量问题, 影响用户体验。因此, 我们需要建立端到端的实时音视频服务智能 QoE 监测和优化系统。

如图 10 所示, 实时音视频服务

智能 QoE 监测和优化系统分为数据收集、健康度评估和智能优化 3 个部分。

(1) 数据收集。该部分主要收集端到端的全链路实时音视频通信数据, 包括终端设备数据、网络环境数据等。

• 终端设备数据: 设备机型、用户 IP、视频流的分辨率、帧率, 在前处理、编码、解码、后处理、渲染过程中的 CPU 使用率, 图形处理器 (GPU) 使用率以及内存使用率等;

• 网络环境数据: 上下行网络丢

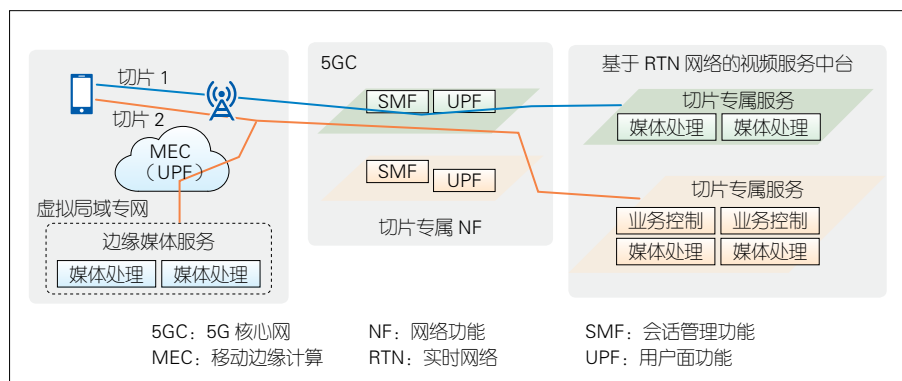
包、抖动、时延等。

(2) 健康度评估。该部分的核心思想是对收集到的监控数据进行过滤、汇聚、实时计算, 并进行实时音视频通信质量评估, 快速识别和感知实时音视频通信中的问题。中兴通讯通过构建 / 更新一组机器学习模型, 判断当前全链路服务状态的健康程度, 并将其作为后续智能优化阶段的触发条件。具体来说, 该部分包括: 首先, 基于“异常状态监控指标与正常状态监控指标处于不同分布”的假设, 选用 QoS 指标 (时延、码率、CPU 等)^[20] 和无参考视频质量评估得分, 构造样本特征空间; 然后, 在此基础上构造多个异构自动编码器^[21], 并利用它们在训练集上的预测残差值进行正则化模型筛选; 最后, 通过模型筛选的多个自动编码器的投票结果, 将被作为当前状态健康度的评估值。

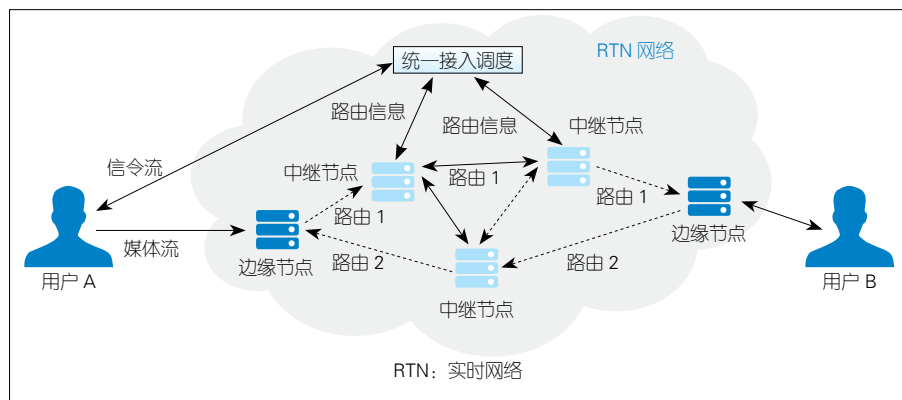
(3) 智能优化。当前状态的健康度低于阈值时, 就需要进行智能优化。这里我们将智能优化过程视为马尔科夫决策过程, 利用强化学习求解当前状态下的最优策略。具体来说, 我们将健康度的前后提升比率定义为奖励, 将网络状态的可观测信息 (时延、丢包、阻塞情况) 定义为状态空间, 将网络参数组合的可调选项 (纠错策略、重传策略、缓冲器的缓冲值和缓冲区间大小) 定义为动作空间, 利用动作探索的奖励反馈实时更新深度策略网络^[22], 并逐步实现当前状态下的最佳网络配置组合。

3 结束语

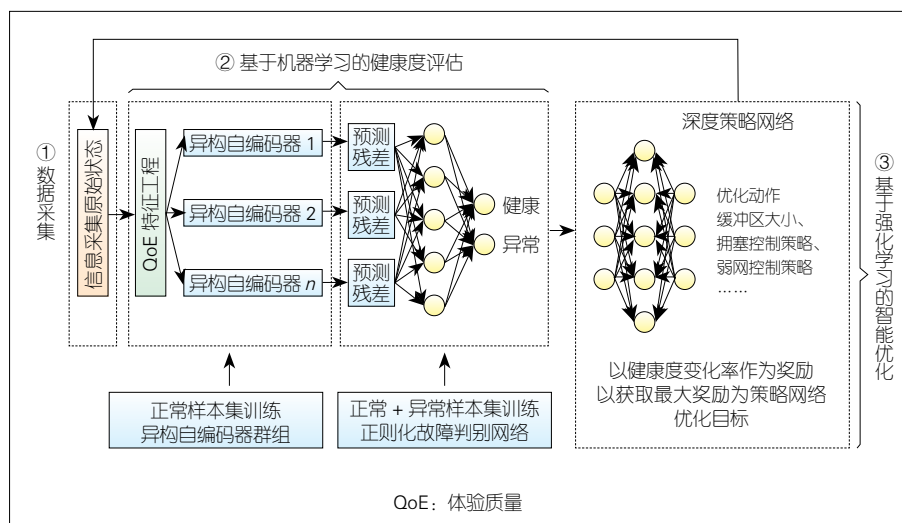
随着 5G 商用落地以及相关设施的完善, 视频的使用体验将不断升级, 视频的业务形态将不断创新, 视频的应用场景也将不断延展。“万物视频化”的趋势对底层技术支撑体系提出了新的、更高的要求。为了使能视频



▲图 8 融合 5G 网络切片、移动边缘计算的实时音视频网络及服务示例



▲图 9 智能路由调度



▲图 10 实时音视频服务智能用户体验质量监测

业务繁荣, 中兴通讯提出了构建智能实时视频网络的理念, 基于自身长期在网络通信、视频多媒体、AI 等领域的持续耕耘和积累沉淀, 创新性地研发了一系列技术和产品, 并使之应用于视频业务端到端流程的各环节。中兴通讯构筑 SmartRTN 综合技术体系, 着眼于改善最终用户的体验, 有效解决了内容增强、高效压缩、可靠传输以及智能运维等问题, 为 5G 视频业务的不断演化和纵深拓展提供了牢固的基础。

参考文献

- [1] Cisco. Cisco visual networking index (VNI) complete forecast update, 2017—2022 [EB/OL]. (2018-12)[2020-12-05]. https://www.cisco.com/c/dam/m/en_us/network-intelligence/service-provider/digital-transformation/knowledge-network-webinars/pdfs/1213-business-services-ckn.pdf
- [2] Cisco. Cisco visual networking index (VNI) global and americas/EMEAR mobile data traffic forecast, 2017—2022 [EB/OL]. (2019-03)[2020-12-05]. https://www.cisco.com/c/dam/m/en_us/network-intelligence/service-provider/digital-transformation/knowledge-network-webinars/pdfs/190320-mobility-ckn.pdf
- [3] ITU. HEVC-SCC [EB/OL]. [2020-12-05]. <http://www.itu.int/rec/T-REC-H.265>
- [4] MPEG. VVC [EB/OL]. [2020-12-05]. <https://mpeg.chiariglione.org/standards/mpeg-i/ver-satellite-video-coding>
- [5] Alliance for Open Media. AV1 [EB/OL]. [2020-12-05]. <http://aomedia.org/>
- [6] AVS Work Group. AVS3 [EB/OL]. [2020-12-05]. <http://www.avs.org.cn/>
- [7] MPEG. MPEG5-EVC [EB/OL]. [2020-12-05]. <https://mpeg.chiariglione.org/standards/mpeg-5/essential-video-coding>
- [8] MPEG. MPEG5-LCEVC [EB/OL]. [2020-12-05]. <https://mpeg.chiariglione.org/standards/mpeg-5/low-complexity-enhancement-video-coding>
- [9] 视频传输延迟分析及解决方案: CMAF, LHL [EB/OL]. (2018-09-21)[2020-12-05]. <https://cloud.tencent.com/developer/article/1346159>
- [10] WebRTC 1.0: Real-Time Communication Between Browsers [EB/OL]. <https://www.w3.org/TR/webrtc/>
- [11] QUIC, a multiplexed stream transport over UDP [EB/OL]. <https://www.chromium.org/quic>
- [12] Salsify. Video is better when the codec and transport work together [EB/OL]. [2020-12-05]. <https://snr.stanford.edu/salsify/>
- [13] Salsify. Salsify 测试结果 [EB/OL]. [2020-12-05]. <http://web.mit.edu/6.829/www/currentsemester/materials/slides-salsify-lecture.pdf>
- [14] DING Z Z, ZHENG Q F, HOU C H, et al. Improving face recognition in surveillance video with judicious selection and fusion of representative frames [C]//ACM Multimedia Asia(MMASia' 20). NY, USA: ACM, 2021. DOI:10.1145/3444685.3446259
- [15] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: a unified embedding for face recognition and clustering [C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 815-823. DOI:10.1109/cvpr.2015.7298682
- [16] RAO Y M, LU J W, ZHOU J. Learning discriminative aggregation network for video-based face recognition [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 3781-3790
- [17] TAIGMAN Y, YANG M, RANZATO M, et al. DeepFace: closing the gap to human-level performance in face verification [C]//2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014: 1701-1708. DOI:10.1109/cvpr.2014.220
- [18] YANG J L, REN P R, ZHANG D Q, et al. Neural aggregation network for video face recognition [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA: IEEE, 2017: 4362-4371. DOI:10.1109/cvpr.2017.554
- [19] GONG S X, SHI Y C, KALKA N D, et al. Video face recognition: component-wise feature aggregation network (C-FAN) [C]//2019 International Conference on Biometrics (ICB). Crete, Greece: IEEE, 2019: 1-8. DOI:10.1109/icb45273.2019.8987385
- [20] InfoQ. 音视频质量评估白皮书 [EB/OL]. (2019-08-26)[2021-12-05]. <https://www.infoq.cn/article/xt9vNLcC6dlkSvu9l6M2>
- [21] HINTON G, OSINDERO S, YW T. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18(7): 1527-1554. DOI: 10.1162/neco.2006.18.7.1527
- [22] 桑顿, 巴图. 强化学习: 第2版 [M]. 俞凯, 译. 北京: 电子工业出版社, 2019

作者简介



吕达, 中兴通讯股份有限公司云视频与能源研究院院长、高级工程师; 研究方向为通信技术和协议、互联网技术、云计算技术、视频技术、数字家庭网络及业务等; 先后从事数字程控交换机、固网软交换、IPTV、视频会议、通信网络供电等产品架构设计与研发管理工作, 曾主持完成数字程控交换机、多媒体视讯、视频会议等重大产品项目; 发表论文多篇, 申请专利 8 项。



郑清芳, 中兴通讯股份有限公司云视频首席科学家; 研究方向为人工智能、计算机视觉、视频编解码、视频通信、人机交互、多媒体芯片与系统等; 先后从事视频智能编目系统、视频搜索系统、手机 3D 成像系统及应用、车载成像与识别系统、人脸识别、3D 立体视觉芯片、视频会议等系统及产品的架构设计与核心技术研发; 发表论文多篇, 申请专利 2 项。



面向视频云微服务系统的智能运维技术

Intelligent Operation and Maintenance Technology for Video Cloud Microservice System

摘要: 提出了一种基于人工智能(AI)的保障视频云体验质量(QoE)的系统架构。该系统针对多个维度创建运维知识图谱,例如运行数据、运行环境、运维数据,以用于建模、感知、映射和分析。在对系统的微服务保障中,运用了图神经网络(GNN)等方法进行分类和预测。通过知识图谱和机器学习,该系统可实现实时监控、自愈恢复、智能预测和主动运维,从而实现 QoE 的智能保障。

关键词: 视频云系统;数据挖掘;知识图谱;图神经网络;智能运维

Abstract: Based on artificial intelligence (AI), a system architecture that guarantees video cloud quality of experience (QoE) is proposed. According to multiple dimensions, the system creates an operation and maintainance knowledge map for modeling, sensing, mapping, and analyzing, such as operating data, operating environment, and operation and maintenance data. In the microservice system, methods such as graph neural network (GNN) are introduced for classification and prediction. Based on the technology of knowledge graph and machine learning, the system can perform realtime monitoring, self-healing, intelligent predicting and active operation and maintainance to implement intelligent guarantee of QoE.

Keywords: video cloud system; data mining; knowledge graph; graph neural network; AI Ops

徐代刚 /XU Daigang

姜磊 /JIANG Lei

梅君君 /MEI Junjun

(中兴通讯股份有限公司, 中国 深圳 518057)
(ZTE Corporation, Shenzhen 518057, China)

DOI: 10.12142/ZTETJ.202101014

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20210128.1824.002.html>

网络出版日期: 2021-01-29

收稿日期: 2020-11-15

随着 5G 时代的到来,增强型移动宽带(eMBB)、海量机器类通信(mMTC)、超可靠低时延通信(URLLC)3大应用场景,尤其是相关的音视频应用,得到了快速发展。此外,由于受到新冠肺炎疫情等不确定性因素的影响,企业或团体远程办公、远程会议的场景和需求日益增多,相应的通信系统也变得日益复杂。以视频会议为例,它需要支持双流会议、多流会议,除了要具备编解码、链路控制、会议控制等多种基本功能外,还要能够管理用户的安全接入、权限控制等。因此,一个好的视频系统,不仅要满足用户的业务需求,支撑音视频编解码、播放、合成、录制、扩展现实(XR)以及会议等多种业务,

还要支撑用户的质量需求,从听得清、看得清到听得懂、看得懂,直至听得真、看得真,以提升用户体验质量(QoE)。对于一个企业级视频系统而言,为了提供优良的用户感知,系统除了提供音视频编解码、XR渲染等技术外,还要能够支持多租户、高并发、大流量。整个系统应稳定可靠且具有高安全性,不仅能灵活控制终端的接入和退出,还可支持云边协同,使系统部件可弹性伸缩。

视频云系统是一个基于微服务架构的云化视频业务系统。它拥有良好的软件设计和硬件架构,支持多业务、多租户、大流量,并支持灵活控制、云边部署、协同协调,可以满足平滑扩容弹缩,并有严格的安全策略设

计,以保证终端用户接入和系统管理的安全性。从视频云系统的角度来看,QoE的保障不仅要求对具体业务指标,如丢包率、抖动和时延等,进行调参调模等,还要求服务端的软件系统具有较好的稳定性和连续性(如基于微服务的视频云系统)^[1]。因此,一旦微服务本身的运行状况出现劣化,上层业务的QoE将会受到影响。基于经验来看,当前期系统调试运行稳定后,中后期系统(如微服务模块)运行状况会出现瓶颈问题,从而导致QoE下降。在系统运行过程中,有两种情况会导致QoE指标下降,且最终影响用户感知:一种是软件代码质量问题,如代码漏洞、场景考虑不周全和压力不足等,或者是系统架构设计有问题,

如无法应对数据风暴；另一种是系统本身策略性问题，如对网络感知出现异常，在需要相应微服务模块进行弹性伸缩时，策略执行动作执行得不够迅速。

因此，我们有必要为视频云系统建立一个智能运维系统。这是因为智能运维系统不仅能够监控具体业务性能指标劣化，还能监控视频业务运行的软件系统劣化，并在此基础上协助分析、定位异常和系统瓶颈，甚至能够主动运维使系统自愈，以更好地提高用户感知。

对于早期的系统运维，运维人员和软件模块开发人员根据巡检和发生的告警，分析日志和代码，来定位和解决问题，其中大多属于事后分析和人肉运维。大量的人工参与，不仅耗时而且容易出现错误，因此，如何能减少人工操作并实现故障自愈走上运维的舞台。随着人力成本的不断增高和业务场景的日益复杂，自动化运维应运而生。自动化运维不仅能够端到端地解决重复、简单而又经验化的低阶运维工作，而且在提取相关经验形成知识库后，还能根据策略定义实施故障自愈。策略闭环和专家经验知识库是自动化运维的两大基石。对于复杂的视频云系统来说，故障自愈不仅是锦上添花，更是一个必备的保障功能，因为它把运维能力从低阶提升到了中阶。

但是，随着 5G 时代来临，由于业务、场景、数据急剧增加，系统架构变得更加复杂，自动化运维已经无法有效满足数据的海量性、流程的复杂性和应用的新颖性需求。因此，在算法、大数据和算力的支撑下，借助人工智能的运维逐渐成熟起来^[2-3]。通过数据挖掘和概率统计来分析海量数据，包括业务专家和流程专家标签异常数据，并通过机器学习（包括深度学习、强化学习）训

练学习异常来提炼规则并融入知识图谱，使自动化运维进一步迈向高阶的智能运维（AIOps）^[4]。

1 视频云系统云边部署架构

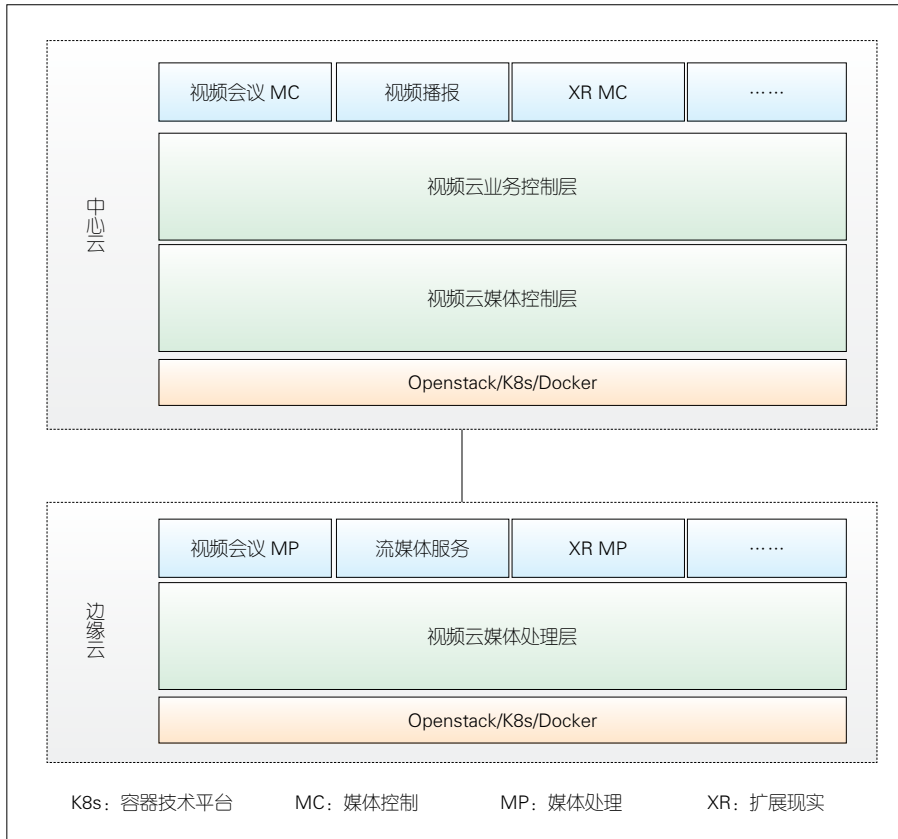
从业务角度来看，视频云系统主要提供媒体应用、媒体控制、媒体处理等服务。其中，媒体应用包括会议电视、音视频会议和直播等；媒体控制除了要对媒体应用的接入进行解析外，还需要进行业务调度和控制业务逻辑，如在视频会议中对不同用户的加入/退出、画面以及静音等的控制；媒体处理要进行音视频编解码、XR 识别等操作。

为了更快速地提供音视频服务，目前视频云系统服务端一般采用云边部署方式^[5]，即媒体面靠近用户边缘部署，控制面在云端中心部署，如图 1 所示。

中心云部署以业务控制和媒体控制为主，包括会议应用服务器、会议控制、资源控制、数据协作、消息群组、水印服务、实时录制控制。边缘云部署以媒体处理为主，其中媒体处理包括多流媒体处理、视频转码合成、实时媒体推流、实时媒体录制，可提供音视频算法库、转码和 QoE 通信引擎等。基于 Kubernetes（也称 K8s）的视频云系统支持微服务的弹性伸缩。

2 视频云系统智能运维架构

借助云边部署和云边协调，系统在中心管理控制并提供业务支持，同时计算下沉到边缘侧，能够更快速地响应和反馈，减少骨干传输网以及上层核心网的资源占用，可以很好地保障用户感知。然而，由于 QoE 的影响可能是来自边缘侧，也可能是来自中心云，甚至可能是两者交互后的结果，因此运维监



▲图 1 视频云系统云边部署示意图

管需要对云边进行统一运维。

如图2所示,左侧是视频云系统的中心云和边缘云,右侧是AIOps智能运维系统。中心云和边缘云都嵌入智能代理,分别通过代理模块将数据上报给运维系统,并先通过分析定位后再通过代理模块分别给云边下发指令进行修复自愈。整个系统实现云边运维闭环自愈。

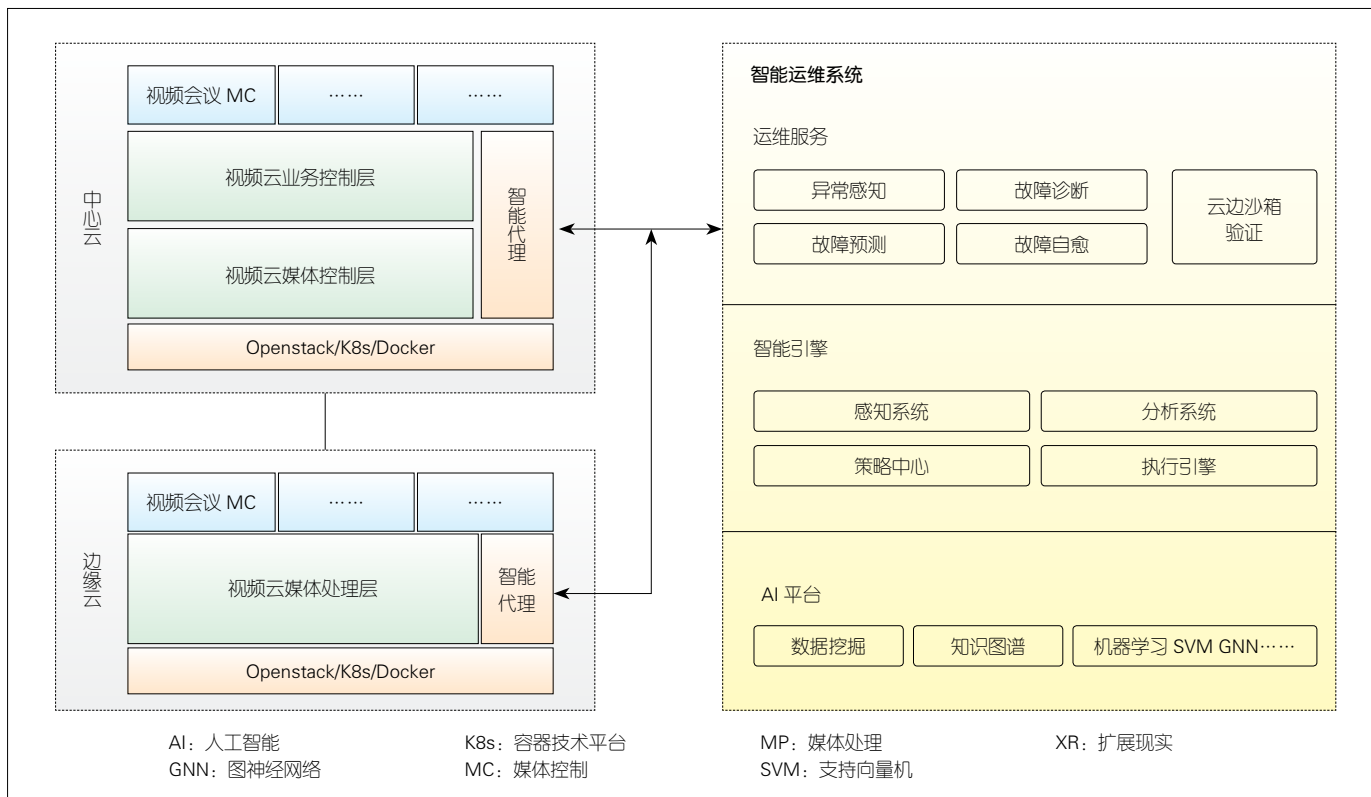
为了保障QoE,智能代理对边缘云和中心云需要采集至少两方面数据。一个方面是运维数据,它不仅包括告警数据、性能数据、日志数据和拓扑资源数据等,还包括微服务的运行数据,如微服务的中央处理器(CPU)和内存等数据;另一个方面是业务数据,即那些会影响QoE的数据。运维数据和业务数据统称为感知数据。其中业务数据具体又分为3类:(1)网络数据,如收发端带宽、丢包率、抖动、时延等;(2)音视频参数,如帧率、

分辨率等;(3)直接体验QoE的终端数据,如手机型号、手机系统、会议室终端、机顶盒终端等。另外,由于视频云系统本身非常复杂,一些模块或者网络在调参(甚至硬件调整)后,需要重新模拟验证,因此一些视频云系统利用数字孪生进行仿真验证。智能代理模块也可以配置探针,采集相关影响QoE的孪生数据并将其上传到保障系统,以进行云边统一沙箱验证。

由图2右侧可以看出,运维保障系统大致包括3层:上层提供运维业务服务,中间框架层提供基本架构和运行支撑,底层是AI支撑。

上层不仅提供具体视频云系统的QoE运维,还提供异常感知、故障诊断、故障预测和故障自愈。异常感知是对感知数据进行异常检测,具体包括两种检测方法:一种是在数据直接异常时,有严重告警或者指标异常,例如抖动超过阈值、微服务CPU直接

超限等,这种异常一般可以通过阈值定义直接检测;另外一种综合判断,比如对某个时段,虽然没有指标超过阈值,但整体已经劣化。此时,可先通过人工直接标注是否异常,然后通过机器学习来学习和推理。故障诊断可对故障进行定界定位,比如,通过关联分析和根因分析,系统发现声音激励切换(VAS)模块导致终端会议在视频切换时出现卡顿等。同时,这些分析的结果将被直接注入知识图谱以供后续推理使用。故障预测可通过对历史数据进行机器学习来推断实时的运行情况,具体包括两种预测:一种是微观预测,例如根据历史数据学习来判断某微服务是否会出现内存异常;另一种是宏观预测,例如当视网络处理单元(NPU)内存消耗较高时,机器学习认为终端会议可能不会出现问题,但XR可能会出现问题。因此,把当前终端类型和NPU内存等维度输



▲图2 视频云系统智能运维架构

入机器学习,可推理预测是否会出现劣化。故障自愈则是根据诊断定位故障原因,或者根据预测判断 QoE 是否发生劣化,并根据知识图谱学习的规则通过策略实施自愈措施,把恢复命令(如微服务弹缩或者微服务迁移指令)通过代理发送给相应系统。

中间的框架层包括感知系统、分析系统、策略中心和执行引擎。感知系统接收代理上传的数据(包括云和边的感知数据),并对这些数据进行分类、清洗和归一化。如果数据是非离散的,归一化可以按照高斯分布或者伯努力分布对其进行处理,例如高斯分布可按照马氏距离来处理。分析系统提供数据分析,如告警、性能以及日志等数据的关联分析和根因分析。策略中心定义相关策略和动作。例如,当 NPU 内存消耗达到 80% 且持续时间超过 60 s 时,系统就会执行弹缩扩容。再例如,如果 AI 预测 NPU 的内存会在视频终端用户数超过 10 个时就会冲高,那么制定策略会将其设定为当用户数超过 8 个时就弹缩(百分比、时间和数量等数据是非标准数据,在此仅做举例说明)。执行引擎保障系统自动化闭环执行,例如对数据清洗及归一化、再挖掘分析、预测、执行策略弹缩、发送指令、全程自动化。

底层 AI 支撑提供 AI 算法和训练推理框架。AI 算法包括数据挖掘和机器学习的算法,并提供知识图谱框架。其中机器学习除了涉及常规分类算法外,如支持向量机(SVM)、逻辑回归(LR)等,还用到了深度学习中的图神经网络(GNN)。知识图谱框架则提供知识图谱的创建、融合和推理等。相对于传统人工专家知识库来说,知识图谱具有更好的可视性、准确性和扩展性。

从异常感知、故障诊断到故障预测、故障自愈,是一个从被动运维到

主动运维的过程。其中,虽然故障自愈的场景不多,但是对于复杂的视频云系统来说,故障自愈是一个非常必要的功能。下面我们以故障自愈的流程来详细说明系统的运作流程。

3 视频云系统故障自愈流程

一般来说,视频云故障自愈有以下 3 种情况:(1)当系统发生故障时,如检测到硬件温度过高,则进行微服务迁移,把有问题的硬件上的微服务迁移到备份机上;(2)当微服务性能关键性能指标(KPI)异常时,相关的微服务可能会进行弹性扩容或者提高服务等级;(3)根据业务指标超常进行微服务弹缩^[6]。

传统自动化自愈等做法就是设计门限阈值或者相关事件,制定相应策略进行迁移、扩容或者熔断等以保障自愈闭环。以微服务弹缩扩容来说,自动化运维最大的问题在于不确定性。这可能是因为指标(如 CPU、内存和带宽)抖动冲高但时间没有达标而不弹缩,也可能是无法一步弹缩到位,即会按照策略多次弹缩才能最终满足要求。还有一种可能就是策略设置错误,比如在一个周期内当最高值达到门限时就会弹缩。然而,由于内存瞬间冲高可能会回落,误弹缩的情况也会发生,一旦发生将浪费不必要的资源。

更高阶的智能化运维能够通过机器学习来寻找比较好的解决方案。基于机器学习的智能自愈,可建立在对历史数据进行机器学习、统计和分析诊断的基础上,生成相关学习模型和知识规则,在异常发生时,能够根据学习到的模型和规则,按照策略定义实施自愈手段,实现闭环自愈。此外,基于机器学习的智能运维,可以在预测服务或者系统必然劣化时,就提前采取措施实施主动运维,即把被动运

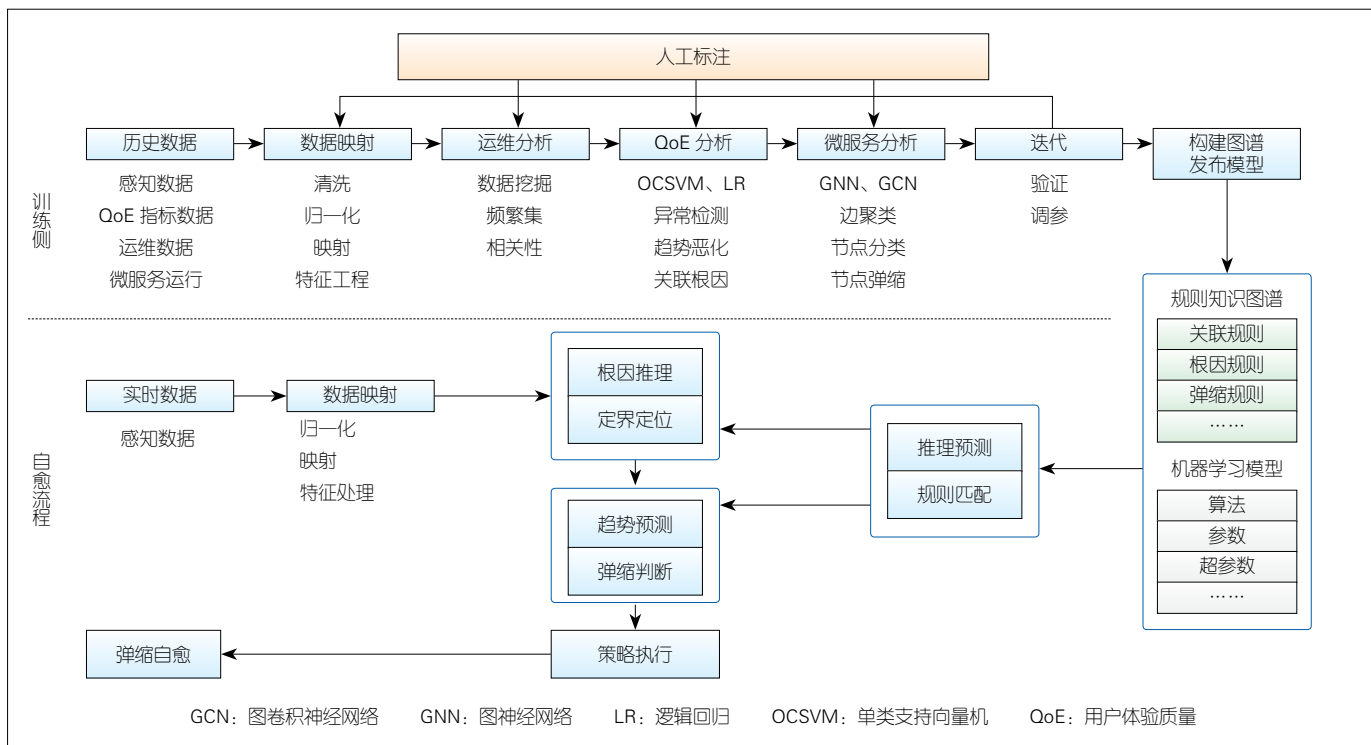
维和主动运维相结合,以保障更好的用户体验效果。

下面我们对视频云系统的故障自愈流程做一个总体描述。如图 3 所示,自愈流程分为上下两部分:上面是训练侧,下面是推理侧。

在训练侧,历史数据就是感知数据,包括 QoE 指标数据、运维数据和微服务运行数据。在采集数据后,系统会首先对数据进行清洗、归一化和映射处理,把运维数据进行频繁集挖掘并分析得到相关性,再通过机器学习对 QoE 进行分析,通过回归或分类来判断趋势是否会恶化(也可以定位相关恶化指标的关联根因)。接着,GNN 学习微服务运行趋势,得到边和节点的关系,以及是否需要弹缩的回归模型和分类模型,并进行验证。在这一阶段,如果效果不好就需要调参进行再次迭代。最终系统形成知识图谱和机器学习模型。值得注意的是,在训练侧,需要人工干预,即不仅在数据挖掘和机器学习阶段要标注,在验证阶段也要交互反馈。

在推理侧,当获得实时监控数据后,如果数据映射感知到异常,系统将进行定位诊断,即根据知识图谱得到问题根因模块,并根据机器学习模型判断是否可能会劣化、是否需要弹缩等。紧接着,系统将依据策略定义执行微服务弹性伸缩或者迁移等具体动作,最终达到自愈。

可以看出,在分析阶段进行数据挖掘后,系统采用传统的机器学习方法进行 QoE 分析,如使用单类支持向量机(OC SVM)进行异常检测,使用 LR 进行趋势是否恶化的分类判断等。由于数据挖掘技术已经比较成熟,比如工业界大多采用的频繁模式树(FP-Growth)和最大频繁项集(FPMAX)等算法,加之传统机器学习和知识图谱构建也比较成熟,本文不再赘述。



▲图3 视频云故障自愈流程示意图

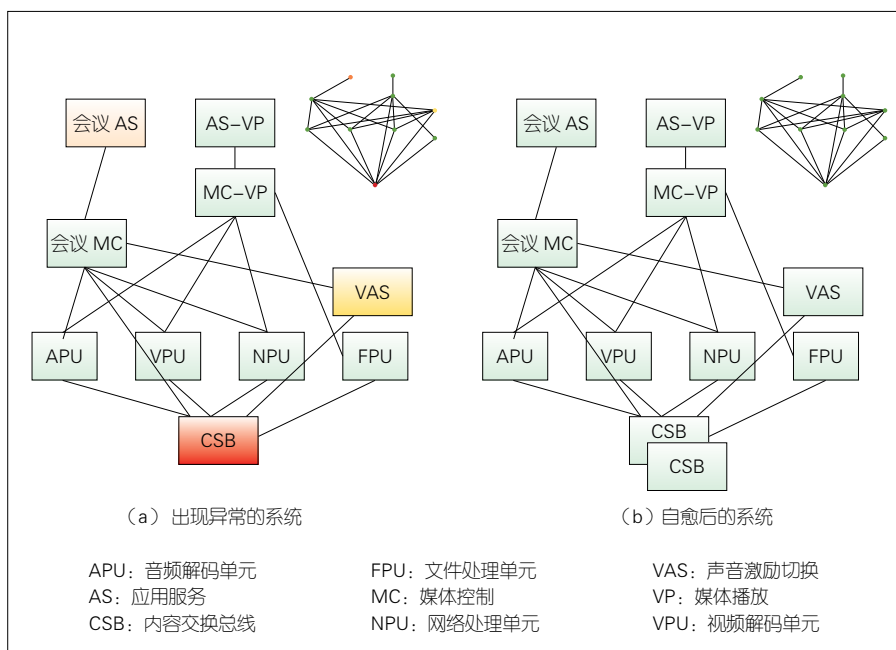
下面我们将重点介绍用于微服务判断和预测的图神经网络算法。

4 视频云系统智能运维模型算法研究

4.1 微服务系统故障分析

在讨论 GNN 算法之前，我们首先介绍基于微服务架构的视频云系统。在视频云系统中，不同微服务组件既有一定的独立性，又有一定的相关性和耦合性。它们的组织架构更像一张图，具体如图 4 所示。

视频云系统表示一个视频会议所需要的微服务局部子图。其中，应用服务（AS）为业务逻辑提供服务。媒体控制可调用不同的媒体处理单元。网络处理单元（NPU）负责收发和接收网络媒体包，同时保障媒体 QoS。内容交换总线（CSB）提供媒体数据的分发。在视频会议中，声音激励切换（VAS）可使发言者的画面被展示



▲图4 视频云微服务部署示意（终端会议和媒体播放）

出来。这些组件与终端会议 AS、媒体播放应用服务（AS-VP）、音频解码单元（APU）、视频解码单元（VPU）均以微服务的形式部署。需要注意的是，图 4 仅是示意图，其中的 APU 和

VPU 等都以各自不同的微服务形式部署。而在实际部署时，为了实现更快的模块间交互，APU、NPU 和 CSB 可能会被部署在同一个微服务中。这是因为如果视频会议出现问题，就需要

快速定位是哪个服务出现了问题。例如,在图4(a)中,如果切换发言出现视频卡顿,终端会议AS将会出现延迟(图4中橘色模块),同时VAS的日志将显示调用缓慢(图4中黄色模块)。一般来说,造成这种现象的原因可能是VAS、VPU或NPU出现了问题,也可能是CSB出现了问题(图4中红色模块)。此时,系统可借助模块间的快速交互,实现问题的准确定位:CSB内存超限影响VAS,导致切换激励出现延迟,进而导致AS出现卡顿。

深度神经网络(DNN)、卷积神经网络(CNN)、LR、SVM+梯度直方图(HOG)等,都是传统的机器学习和深度学习方法。虽然它们在提取欧氏空间数据的特征方面取得巨大的成功,并在线性分类、图像分类、声音处理等相关应用上有着非常优秀的表现,但是在处理非欧式空间数据时(如社交网络、交通网络和化学分子式),由于数据中每个节点的边或者邻域是不固定的,所以这些深度学习方法无法对此建立模型,即无法使用同样尺寸的卷积核来表达或者泛化。而GNN,如图卷积网络(GCN),可同时结合图和卷积的特点,能够取得比较好的效果^[7]。例如,在参考文献[8]中,CHAI D.等利用GCN和多层全联接神经网络(MLP)模型,预测了共享单车流量问题;在参考文献[9]中,作者提出R-GCN模型,并分别将其运用到联系预测和实体分类两项任务上,在关系图的多个推理步骤中使用编码器模型来积累信息,显著改进了链路预测模型;在参考文献[10]中,YING R.等把GCN运用在推荐系统中,拼趣公司(Pinterest)在此论文基础上把GCN运用在商业系统中以推荐图片给不同用户。

如前所述,微服务方式部署实际

上是一个典型的图结构。图4中的节点就是各个微服务,边是各个服务之间的调用关系。例如,假如AS和VAS没有直接的调用关系,那么它们两个节点之间就没有相对应的边。因此,视频云运维系统在对图进行学习和预测时,不仅采用了传统的机器学习算法(例如OCSVM和LR等),还结合了GCN等算法。需要注意的是,在真实视频云系统中,微服务数量有上千个,不同服务之间的关联更加复杂,微服务部署的复杂度远远超过图4所示的结构。因此,简单的图搜索和图嵌入都不能进行有效的推理和根因定位。

4.2 GNN

GNN是一个很宽泛的概念。一般来说,GNN就是图+神经网络。目前,与GNN相关的模型算法有GCN、图注意力网络(GAT)等。本文中,我们以GCN为例来做具体讨论。

一般来说,GNN可被用来处理3类任务:

- 节点层面任务。该任务主要包括对微服务节点进行分类和预测,例如判断这个节点和其他节点是否属于同一类,或者当该节点的业务数据有异常时,是否需要微服务弹缩或

迁移等。

- 边层面任务。该任务与微服务直接相关,比如微服务调用链、根因分析等。

- 图层面的任务。该任务可对整个模块或者子图进行分类预测,例如预测整个视频云是否正常。

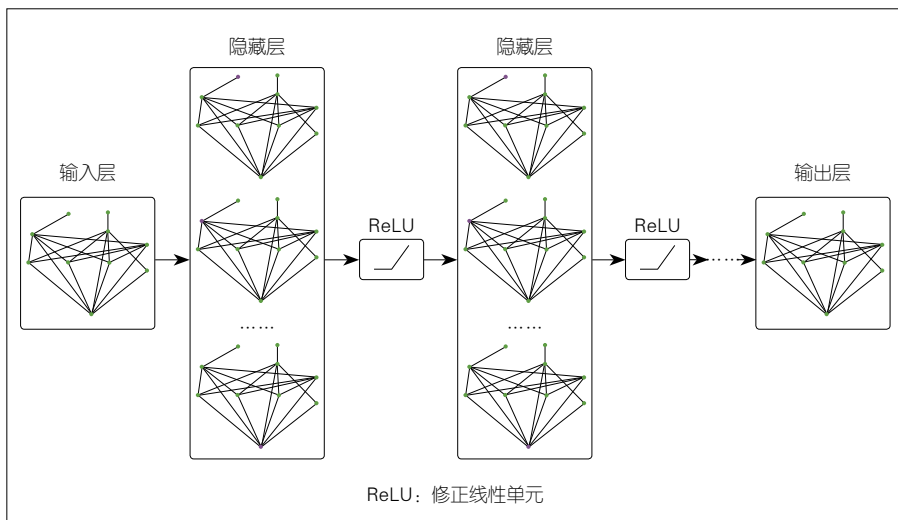
我们提出的视频云保障系统主要考虑节点层面任务和边层面任务。我们需要首先对微服务的组成架构进行图表示,然后再进行标注和训练以获得模型。

我们构建微服务图,其中GCN把整个图 G 、每个节点 V 、每条边 E 转化为稠密向量。当然,并不是每次都要把 G 、 V 、 E 进行向量化的,哪部分需要向量化取决于实际的应用场景。

以图4中的微服务为例,我们搭建了一个GCN网络,如图5所示。其中,左边为输入层,右边为输出层,中间有2个或3个隐藏层,并且每个隐藏层的激活函数均为修正线性单元(ReLU)。

以图4的子图数据为例,我们对其进行迭代传播训练。相关传播公式如公式(1)~(3)所示。

$$H^{(l+1)} = \sigma(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)}), \quad (1)$$



▲图5 神经网络

$$\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}, \quad (2)$$

$$\tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}, \quad (3)$$

其中, σ 是非线性激活函数(这里为 ReLU), \mathbf{D} 是度矩阵, \mathbf{I} 是单位矩阵, \mathbf{W} 是卷积权重矩阵。传播过程中有 N 个节点(在图 4 中, $N=10$), 每个节点都有自己的特征, 包括 CPU 值、内存大小、业务指标、KPI 等 M 个维度的数据。这些节点的特征数据经过归一化处理后首先组成一个 $N \times M$ 维的矩阵 \mathbf{H} , 然后各个节点之间的关系也会形成一个 $N \times N$ 维的矩阵 \mathbf{A} (也称为邻接矩阵)。这里, \mathbf{H} 和 \mathbf{A} 是我们模型的输入。通过 \mathbf{D} 和 \mathbf{A} 可计算并形成拉普拉斯矩阵。

以图 4 的子图为例, 图 6 是邻接矩阵和度矩阵的实际数据。通过这些数据计算得到的拉普拉斯矩阵, 随后将进行公式(1) — (3) 中的卷积运算。

整个学习过程为有监督学习。系统先根据业务异常对节点进行标签, 然后把整个数据按 7 : 3 的比例拆分成训练数据和验证数据。损失函数可以是比较通用的交叉熵损失函数。通过训练学习, 系统可以得到不同节点之间的分类、关联程度和节点是否异常的模式。

通过学习我们可以看出, 由于受到相邻和其他节点的影响(相邻关系越近, 影响越大), 图中的每个节点都在时刻改变着自己的状态, 直至平衡。这里, 我们仍然以图 4 的子图为例, VAS 和 CSB 就是同一类节点, 因此系统可以通过它们来确定关联关系和根因关系。

此外, 有 3 点需要注意: (1) 前文所述的传播表达方式是基于谱域方法而提出的。但是对整个网络来说, 由于节点和边的关系非常复杂, 除了

内存消耗巨大外, 对 \mathbf{D} 和 \mathbf{A} 的求逆和行列式计算也将非常耗时, 因此很多优化方法目前被引入进来, 比如基于空域的方法和节点采样的方法等。

(2) 和 DNN 近似无限表达能力不一样, GNN 的表达能力是比较受限的^[11], 当然, 在云视频微服务系统中, GNN 的表达能力是完全能够覆盖远不足万计的微服务。(3) 在一般的网络搭建中, GCN 后面会再设置一层 MLP 以协助业务判断。由于这也是通用模型, 不是本文讨论的重点, 故这里不再赘述。

5 视频云系统智能运维能力的提升效果

视频云微服务系统是基于 K8s 集群进行部署的, 它主要通过配置运维策略、实时采集指标、可视化监控、故障工单等, 来实现信息技术运维(ITOps)。AIOps 巧妙地将机器学习和知识图谱相结合, 取得了更好的运维效果。下面我们以微服务弹缩自愈场景来进行对比说明。一般来说, 微服务指标包括 CPU 值、内存、输入输出(IO)以及 Java 虚拟机(JVM)内存等。传统 ITOps 对 CPU 指标超限和

邻接矩阵

#AS1	MC1	APU	VPU	NPU	VAS	FPU	CSB	AS2	MC2	
[0,	1,	0,	0,	0,	0,	0,	0,	0,	0],	#AS1 会议 AS
[1,	0,	1,	1,	1,	1,	0,	1,	0,	0],	#MC1 会议 MC
[0,	1,	0,	0,	0,	1,	0,	1,	0,	1],	#APU
[0,	1,	0,	0,	0,	1,	0,	1,	0,	1],	#VPU
[0,	1,	0,	0,	0,	1,	0,	1,	0,	1],	#NPU
[0,	1,	1,	1,	1,	0,	0,	1,	0,	0],	#VAS
[0,	0,	0,	0,	0,	0,	0,	1,	0,	1],	#FPU
[0,	1,	1,	1,	1,	1,	1,	0,	0,	1],	#CSB
[0,	0,	0,	0,	0,	0,	0,	0,	0,	1],	#AS2 AS-VP
[0,	0,	1,	1,	1,	0,	1,	1,	1,	0],	#MC2 MC-VP

度矩阵

#AS1	MC1	APU	VPU	NPU	VAS	FPU	CSB	AS2	MC2	
[1,	0,	0,	0,	0,	0,	0,	0,	0,	0],	#AS1 会议 AS
[0,	6,	0,	0,	0,	0,	0,	0,	0,	0],	#MC1 会议 MC
[0,	0,	4,	0,	0,	0,	0,	0,	0,	0],	#APU
[0,	0,	0,	4,	0,	0,	0,	0,	0,	0],	#VPU
[0,	0,	0,	0,	4,	0,	0,	0,	0,	0],	#NPU
[0,	0,	0,	0,	0,	5,	0,	0,	0,	0],	#VAS
[0,	0,	0,	0,	0,	0,	2,	0,	0,	0],	#FPU
[0,	0,	0,	0,	0,	0,	0,	7,	0,	0],	#CSB
[0,	0,	0,	0,	0,	0,	0,	0,	1,	0],	#AS2 AS-VP
[0,	0,	0,	0,	0,	0,	0,	0,	0,	6],	#MC2 MC-VP

APU: 音频解码单元
AS: 应用服务
CSB: 内容交换总线

FPU: 文件处理单元
MC: 媒体控制
NPU: 网络处理单元

VAS: 声音激励切换
VP: 媒体播放
VPU: 视频解码单元

▲图 6 邻接矩阵和度矩阵数据

一问题能够处理得很好,但是对于内存超限却很难判断,这是因为内存会出现回收的情况。下面我们将对比测试这两种情况。

视频云的微服务蓝图主要包括:逻辑主机(Pod)数量共1个,CPU为2核,内存为2GB。传统ITOps定义策略为:感知时间为1min,当采集间隔时间为5s,即1min采样12次,弹缩消耗时间为1min。在感知时间内,采样中平均CPU消耗达85%时弹缩1个Pod,内存消耗超过80%时弹缩1个Pod。测试时,我们按最终弹缩4个Pod为例,具体测试效果如图7所示。

由图7可知,传统ITOps的弹缩是台阶式弹缩,即感知、弹缩、再感知、再弹缩,最终感知没有异常时,则弹缩到位。在一次感知后,AIOps根据机器学习到的模型,直接进行一步到位的弹缩。图7的测试是在假设感知时间是一致的条件下进行的。而实际上,AIOps的感知时间是按过去的时间窗口来预测的,它的感知时间会远远小于1min。但即便是对于简单的场景,AIOps的效果也会远远好于基于策略的ITOps。

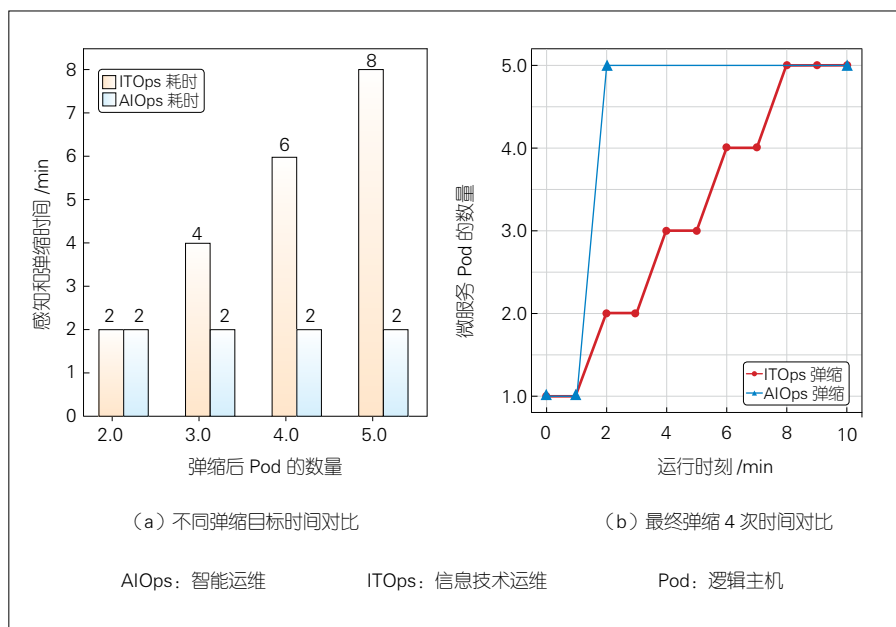
从分类是否正确的角度来看,即是否应该弹缩和是否不弹缩来看,ITOps根据策略定义来分类,AIOps根据机器学习来分类。这里,我们定义TP为应该弹缩,实际弹缩;FP为不应该弹缩,实际弹缩;FN为应该弹缩,实际未弹缩;TN为不应该探索,实际未弹缩。测试显示,对于ITOps来说,TP=67,FP=7,FN=6,TN=20;对于AIOps来说,TP=65,FP=22,FN=8,TN=5。表1给出了在测试环境中100次弹缩是否准确的评价数据统计(统计公式见参考文献[12])。

由表1可知,AIOps在分类方面的表现比完全根据策略定义的ITOps更好。

除了微服务弹缩,在整个系统运维保障能力方面,AIOps比ITOps也有极大的提升,包括异常感知、故障诊断、故障自愈和故障预测等技术指标。以图4的终端会议和媒体播放局部场景为例,该场景共有61类不同告警。故障平均修复时间(MTTR)包括故障感知、故障定位、故障诊断和故障恢复。由表2可知,与ITOps

相比,AIOps至少提升了60%的故障修复效率。

除了在质量保障和效率方面的提升外,AIOps在成本优化方面也有较大的提升。由于成本优化是一个比较复杂的课题,而本文的阐述重点是质量保障和故障自愈,因此,本文中我们仅以视频云微服务系统的K8s集群的成本,来做对比说明下。如表3所



▲图7 AIOps 和传统ITOps 自愈时间对比

▼表1 AIOps 和传统ITOps 分类评价

运维系统	准确率/%	精确率/%	召回率/%
ITOps	70	84	89
AIOps ^[19]	87	90	92

AIOps: 智能运维 ITOps: 信息技术运维

▼表2 AIOps 和传统ITOps 故障处理效率对比

运维系统	故障感知时间/min	故障定位时间/min	故障诊断时间/min	故障恢复时间/min
ITOps	15 ~ 20	> 60	> 50	15
AIOps	< 5	< 15	< 20	15

AIOps: 智能运维 ITOps: 信息技术运维

▼表3 AIOps 和传统ITOps 对K8s 集群的成本优化对比

运维系统	Pod 资源优化/%	非忙时间资源缩减/%	云存储成本	GPU 服务器	调整集群大小
ITOps	11	10	< 5%	< 5%	< 5%
AIOps	14	17	22%	21%	13%

AIOps: 智能运维 GPU: 图形处理器 ITOps: 信息技术运维 Pod: 逻辑主机

示,通过策略和人工经验的 ITOps,仅在 Pod 资源优化和非忙时间资源缩减方面有一定成效而 AIOps 在云存储成本、GPU 服务器和调整集群大小等方面表现更显著。这是因为 AIOps 通过历史数据的机器学习和数据挖掘,在成本优化方面做得更科学、更深入,而不是简单依靠人工经验等给出策略来控制成本。

6 结束语

在面向视频云系统的云边端复杂部署场景时,系统的运维保障变得极为复杂。AIOps 技术能够有效地提升视频云的运维能力和 QoE 等级。与传统 ITOps 相比, AIOps 技术不仅能够节约运维成本、提升运维效率,还能实现更加精准的运维服务、提升用户体验。

致谢

本研究得到中兴通讯股份有限公司胡锐、付迎春、周祥生、弄庆鹏等的帮助,谨致谢意!

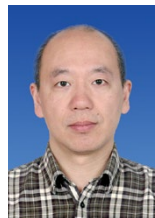
参考文献

- [1] ZHANG X, XIE L, GUO Z. Quality assessment and measurement for internet video streaming [J]. ZTE communications, 2019, 17(1): 12–17. DOI: 10.12142/ZTECOM.201901003
- [2] 董德尊, 欧阳硕. 分布式深度学习系统网络通信优化技术 [J]. 中兴通讯技术, 2020, 26(5): 2–7. DOI: 10.12142/ZTETJ.202005002
- [3] 李建飞, 曹畅, 李奥, 等. 算力网络中面向业务体验的算力建模 [J]. 中兴通讯技术, 2020, 26(5): 34–38. DOI: 10.12142/ZTETJ.202005007
- [4] AIOps 标准工作组. 企业级 AIOps 实施建议白皮书 [R]. 2018
- [5] 高志鹏, 尧聪聪, 肖楷乐. 移动边缘计算: 架构、应用和挑战 [J]. 中兴通讯技术, 2019, 25(3): 26–27. DOI: 10.12142/ZTETJ.201903004
- [6] 龚正, 吴治辉, 王伟, 等. Kubernetes 权威指南: 从 Docker 到 Kubernetes 实践全接触 [M]. 北京: 电子工业出版社, 2017: 93–171
- [7] WU Z H, PAN S R, CHEN F W, et al. A comprehensive survey on graph neural networks [EB/OL]. [2020–11–10]. <https://arxiv.org/abs/1901.00596>
- [8] CHAI D, WANG L Y, YANG Q. Bike flow prediction with multi-graph convolutional networks. [EB/OL]. [2020–11–10]. <https://arxiv.org/abs/1807.10934>
- [9] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks [C]//European Semantic Web Conference. Crete, Greece: Springer, 2018: 593–607
- [10] YING R, HE R Y, CHEN K F, et al. Graph convolutional neural networks for web-scale recommender systems [EB/OL]. [2020–11–10]. <https://arxiv.org/abs/1806.01973v1>
- [11] RYOMA S. A survey on the expressive power of graph neural networks [EB/OL]. [2020–11–10]. <https://arxiv.org/abs/2003.04078>
- [12] 李航. 统计学习方法 [M]. 北京: 清华大学出版社, 2012: 10–19

作者简介



徐代刚, 中兴通讯股份有限公司网络智能平台研发总工、OES 运营技术专家委员会主任及首席架构师; 研究方向为电信运营系统微服务架构和云化网络智能技术。



姜磊, 中兴通讯股份有限公司网络智能平台高级架构师、OES 运营技术专家委员会委员; 研究方向为电信网络智能运维、数据挖掘、机器学习和深度学习。



梅君君, 中兴通讯股份有限公司视频云平台研发总工、大视频技术专家委员会委员; 研究方向为视频能力基础设施云原生架构和低延时高性能实时音视频通信技术。



用于人工智能的硅基光电子芯片

Silicon Photonic Chips for Artificial Intelligence

摘要: 提出了利用硅基光电子芯片进行人工神经网络计算处理的方法。硅基光电子芯片凭借光子的独特性质,能够在人工神经网络的计算处理中发挥高带宽、低时延等优势。在处理深度学习中大量的矩阵计算的乘加任务时,硅基光电子芯片拥有更高的处理速度和更低的能耗,从而有利于深度学习中的神经网络计算速度和性能的提升。

关键词: 人工神经网络;硅基光电子芯片;人工智能;深度学习

Abstract: Silicon photonic chips are used to perform artificial neural network computation. Because of the unique properties of photons, silicon photonic chips have the advantages of high bandwidth and low delay in the computation and processing of artificial neural network. When dealing with the multiplication and addition task of a large number of matrix calculations in deep learning, silicon photonic chips have higher processing speed and lower energy consumption, which is beneficial to the improvement of the computational speed and performance of artificial neural network in deep learning.

Keywords: artificial neural network; silicon photonic chips; artificial intelligence; deep learning

白冰 / BAI Bing
裴丽 / PEI Li
左晓燕 / ZUO Xiaoyan

(北京交通大学, 中国 北京 100044)
(Beijing Jiaotong University, Beijing 100044, China)

DOI: 10.12142/ZTETJ.202101015
网络出版地址: <http://kns.cnki.net/kcms/detail/34.1228.TN.20200421.1831.002.html>

网络出版日期: 2020-04-22
收稿日期: 2020-02-20

人工智能发展的着眼点之一是强大的大型数据集处理工具。这就要求计算机在没有获得明确指令的条件下,能快速高效地学习并组合分析大量信息。人工神经网络就是可以进行学习的数据处理计算机,而以人工神经网络为基础的深度学习算法因其在图像识别、问题决策、语言翻译、自动驾驶^[1]、医疗辅助^[2]等方面的应用而受到学术界和工业界的关注。

目前,人工神经网络几乎全部依赖于传统的电域集成芯片,包括中央处理器(CPU)、图形处理器(GPU)、现场可编程门阵列(FPGA)、专用集成电路(ASIC)。微电子芯片因其结构上无法规避的缺陷,在处理大量的矩阵运算时,面临带宽低、功耗大、速度慢等问题,但人工神经网络实现的基础就是大量的矩阵运算;因此,要想实现深度的人工神经网络,就需要更多的时间和能耗成本。我们可以通过不断提高芯片集成度,进行存内

计算等方法解决这个问题;但与此同时,晶体管尺寸不断缩小,晶体管的性能也越来越受到量子效应的影响,这限制了集成度的不断提高。另外,存内计算的方法与现有的人工神经网络算法匹配度不高也限制了存内计算这种方法的应用。

为了解决上述问题,学术界和工业界越来越多地致力于开发新的硬件架构,以适应人工神经网络和深度学习的应用。借助光子器件优势(带宽大、速度快),业界提出将一部分信息承载和计算处理用于改善电域芯片存在的问题。相比于传统的三五族或砷酸铟光器件,硅基光电子芯片上的光器件集成在同一硅衬底上,集成度更好且基本与成熟的互补金属氧化物半导体(COMS)工艺兼容。

利用光电子技术实现的人工神经网络主要包括前馈神经网络(FNN)、循环神经网络(RNN)、脉冲神经网络(SNN)3种类型。马赫·曾德尔干

涉仪(MZI)和微环谐振器(MRR)具有干涉、谐振等物理特性,可以实现调制器、滤波器等多种器件功能,被广泛地用于通信、传感等领域。目前相对比较完善的、主流的硅基集成通信芯片是基于MZI和MRR的两种类型,因此人工神经网络芯片也主要基于这两种类型。本文中,我们围绕这两种类型对硅基光电子人工智能芯片的进展进行简要阐述,并对未来的发展态势进行展望。

1 利用光网络进行矩阵运算

人工神经网络的思路是首先将输入的事物转化为矩阵,然后经过大量的矩阵运算,最终得到所需要的结果。不同算法的处理流程可能会有一些差异,但是都会包含大量的矩阵运算。矩阵运算的基础就是乘积累加运算(MAC)。在光子领域实现MAC操作并不会在本质上消耗能量,这是光子集成电路的优势之一。

1.1 集成 MZI 进行矩阵运算的原理

利用 MZI 进行片上矩阵运算的原理是基于 M. RECK 等于 1994 年提出的酉矩阵分解方法^[3]的。在该方法中, 可调反射率和透过率的分束器和可调的移相器组成基本单元, 并通过电压控制分束器的分光比和移相器的相位实现控制输出端口的光强, 如图 1 所示^[3]。酉矩阵运算中输入的 $N \times 1$ 列向量元素大小由输入光强大小来表示, 位置由输入端口的的位置表示, 输入的多束光分别从入口端 MZI 的一个臂进入 MZI 阵列, $N \times N$ 酉矩阵中的元素使用 MZI 阵列中每一个 MZI 包含的 2 个移相器和分光器的参数来表示。这使得光通过这些 MZI 时, 相位和幅度会发生改变, 进而达到计算效果。最

后根据输出端的 $N \times 1$ 个输出光强大小来计算结果列向量元素大小, 元素位置由输出端口的的位置表示。在进行输入光强的调制和输出光强的探测后, 利用光网络可实现酉矩阵的计算。

图 1 为光路酉矩阵分解结构。其中, 较大黑横矩形表示分光比可调分束器, 小黑斜矩形表示移相器, 上方细长黑矩形为全反射镜面。

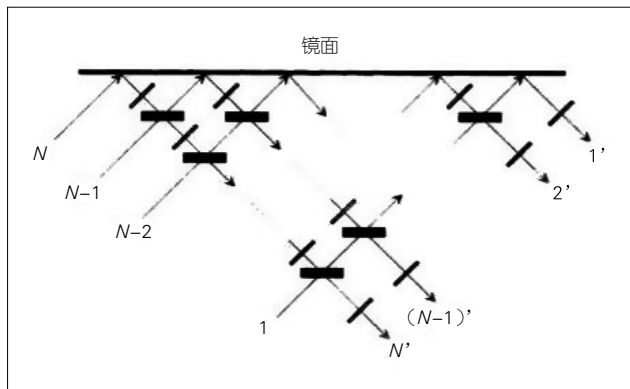
在酉矩阵实现之后, 我们可以利用奇异值分解 (SVD) 的方法对任意矩阵进行分解, 即 SVD 将矩阵分解为 2 个酉矩阵和 1 个对角矩阵酉矩阵, 光路对对角矩阵的模拟用衰减器即可完成, MZI 也可以做衰减器。这样就实现了利用 MZI 进行矩阵运算。

中输入的 $N \times 1$ 列向量中的元素大小用光强大小表示, 列向量中元素的位置由不同的波长表示 (因为输入列向量来自于电域信号, 所以需要通过调制器进行电光转换)。 $M \times N$ 矩阵的每一列元素用同一个波长表示, 不同列用不同的波长表示, 也就是说同一列的 MRR 耦合的是同一个波长。矩阵的每一行用一个公共波导以及耦合在其上的 MRR 表示, 然后每行上的 MRR 根据谐振波长的不同, 分别对输入的不同波长的光信号进行强度调制以实现乘法器, 强度的大小表示的是此行元素的大小, 然后 MRR 再将不同波长的光信号耦合入公共的波导实现加法器。最后矩阵运算结果是一个 $M \times 1$ 列向量, 其元素大小通过光强大小来表示, 然后经光电探测器进行光电转换后, 再通过测量电流大小后得到。

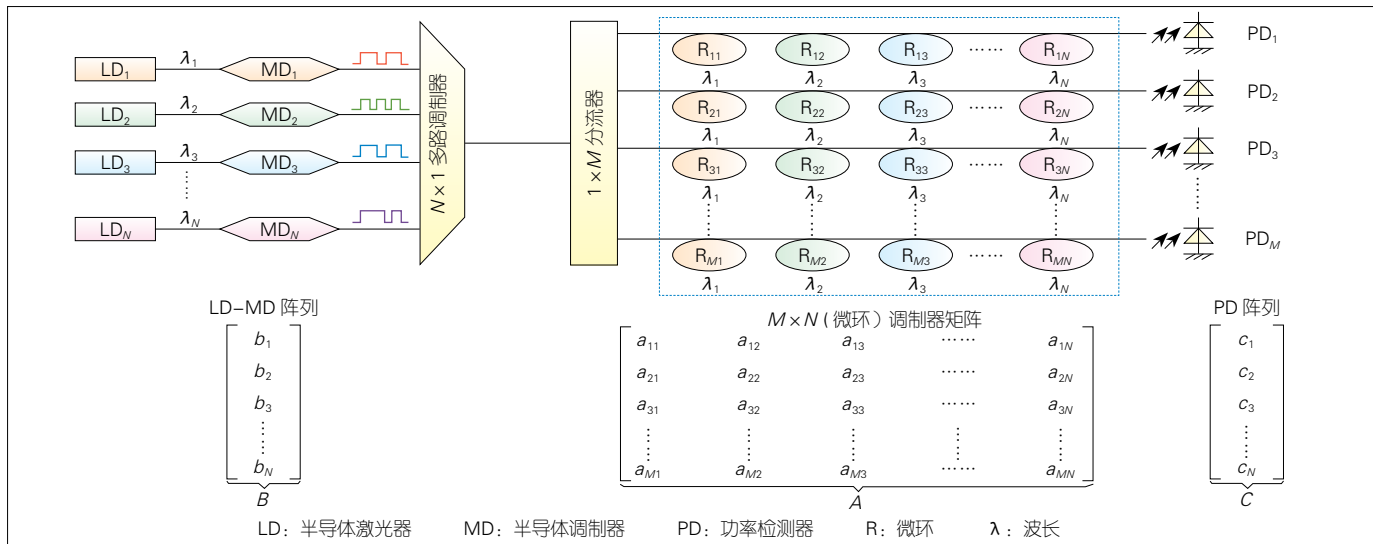
1.2 集成 MRR 实现矩阵运算的原理

MRR 可以先将特定波长的光信号耦合到环上进行调制, 然后再耦合进直波导。MRR 实现矩阵运算的原理为: 通过透过率的调节来实现矩阵的表示。首先矩阵运算

图 2 所示的是 YANG L. 等提出的一种利用 MRR 来实现矩阵运算的方法^[4]。这种 MRR 光网络结构可以执行一个 $M \times N$ 矩阵 A 和一个 $N \times 1$ 向量 B 的乘法。 B 是输入向量, 用 N 个不同波长光信号的光功率大小来表示向量 B 中的元素。这一个列向量 B 是通过 N 个外部调制或直接调制激光二极管



▲图 1 光路酉矩阵分解结构



▲图 2 微环谐振器实现矩阵运算的结构

所生成的。 N 个光信号通过一个多路复用器被多路复用到一个公共波导上, 然后通过一个 $1 \times M$ 的光分路器将其平行投影到 M 行调制器上。矩阵 A 的 a_{ij} 元素由位于矩阵第 i 行和第 j 列的 MRR 的透射率表示, 位于同一行的每个 MRR 仅对具有特定波长的光信号进行操作。随着 $M \times N$ 光脉冲通过 MRR 矩阵, 光信号就在环上进行了所有的 $M \times N$ 乘法过程。在 M 个环上进行乘法运算后, 其累加过程在公共输出波导中进行。因为不同波长的信号在公共波导中几乎不会互相干扰。结果向量 C 的元素由光检测器阵列检测到的 M 个光功率表示。由此, 利用 MRR 进行的矩阵运算便得以实现。

2 现阶段片上人工神经网络的实现

由光来进行矩阵运算是解决人工神经网络需要大量矩阵运算的思路之一。现阶段, 硅基集成的、可用于搭建人工神经网络的典型基础器件就是 MZI 和 MRR。通过这些器件的光学特性实现了 MAC 运算和脉冲神经元的模拟。借助光子数据处理方面的优势, 我们将软件和硬件进行深度匹配, 使用高效的光电计算取代微电子处理器的计算^[5]。光电子集成、数学和软件算法等领域的深度交叉是解决人工神经网络算法大量密集计算问题的路径之一, 也是人工神经网络算法片上实现的发展趋向。

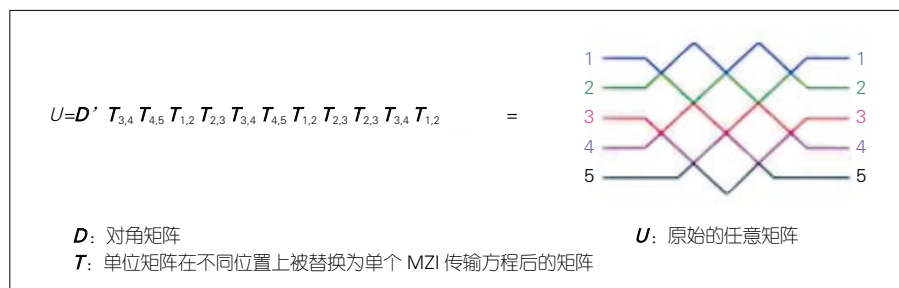
2.1 MZI 型片上人工神经网络

在基于 MZI 构建的前馈人工神经网络中, 信息从输入层单向传递到输出层。信号前向传播时, 不需要将输出再次反馈, 只需要进行加、乘, 以及比较操作即可, 这与擅长矩阵运算的光网络相匹配; 因此, 此种方法的光路硬件的实现获得了广泛的探索 and 关注。虽然 M. RECK 等发现以 MZI

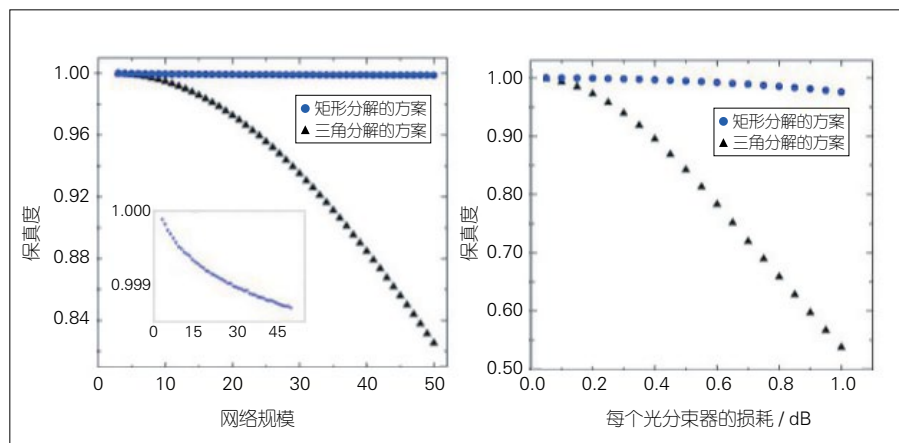
进行酉矩阵的分解方法时并未考虑集成^[3], 但是 MZI 型人工神经网络日渐向集成发展。W. R. CLEMENTS 等在 2016 年基于 M. RECK 等的三角分解法提出了矩形分解法^[6], 将 MZI 进行重新排布来实现酉矩阵运算。通过将 MZI 的排布形状从三角转化为矩形, 减少一半的光学深度, 同时也增加了计算网络的误差容忍度。酉矩阵的分解过程如图 3 所示。酉矩阵矩形分解方法比酉矩阵三角分解方法更有优势。这是因为酉矩阵三角分解方法的光路是不对称的, 从而导致了一些传输过程中的误差。矩形设计减小了线路不对称性, 并缩短了最长链路的长度, 从而减少了光传播的路径损耗和误差。在对 500 个随机生成含误差酉矩阵传输的模拟中, 随着酉矩阵规模 N 从 2 扩大到 50, 三角分解方法的准确度由 100% 下降到约 82%; 但是矩形分解方法的准确度并未发生明显下

降, 一直保持在约 100%, 具体如图 4 所示。

2017 年, SHEN Y. C. 等利用 56 个 MZI 实现了可以用于元音识别的全连接片上神经网络, 制成了光子干涉单元芯片。芯片的部分结构如图 5 所示^[7]。在这个设计中, 通过 MZI 阵列进行神经元线性部分的运算, 人工神经网络中的非线性激活函数采用电域仿真的方法得以实现, 最终可实现全连接神经网络的片上系统。该芯片搭建了 2 层、每层 4 个神经元的全连接神经网络。图 5 所示的芯片结构只有 1 个酉矩阵和 1 个对角阵, 所以应用时首先要将元音信号转为光信号, 取得结果放到电域中处理为光信号再传进来, 至此完成一层计算。将以上过程循环两遍即为 2 层神经网络。上述结构在对 4 个元音的类别的实验中, 能够从大量不同元音的语音信号中正确识别和分类元音, 准确率达到 76.7%。



▲图 3 酉矩阵的四边形分解过程



▲图 4 三角分解和矩形分解对误差的容忍度

2019年, M. Y. S. FANG等研究了两种类型的MZI神经网络, 分别为GridNet(网格网络)、FFNetNet(快速傅里叶变换网络), 其物理结构如图5所示^[8]。其中, 矩阵的分解采用的是SVD。FFT酉矩阵乘法器是非通用性的乘法器, 它由Cooley-Tukey FFT算法启发而来, 用牺牲通用性的方式来换取结构上的紧凑性。由图6可以看出, GridNet和四边形结构是同一种结构, 它和FFNetNet结构皆为 8×4 的线性矩阵运算器。这两种结构均为仿真, 在零误差的情况下, GridNet的准确率约为98%, FFNetNet的准确率约为95%, 因此零误差时的GridNet准确率比FFNetNet的准确率高; 但是在有差错的情况下, FFNetNet的容错率要比GridNet高。在综合误差从0升高到0.02时, FFNetNet的准确率由约95%降到约93%, 而GridNet的准确率由约97%降到约48%。越小的网络差错传播所带来的误差就会越小, 这导致了FFNetNet的稳定性要优于GridNet。

综上可得, 无论是在矩阵规模扩大还是误差增加的情况下, 三角结构的准确率低于矩形结构。矩形结构和GridNet是同种结构, 它和FFNetNet结构各有利弊: GridNet结构的通用性好, 但在存在误差的情况下, 准确率

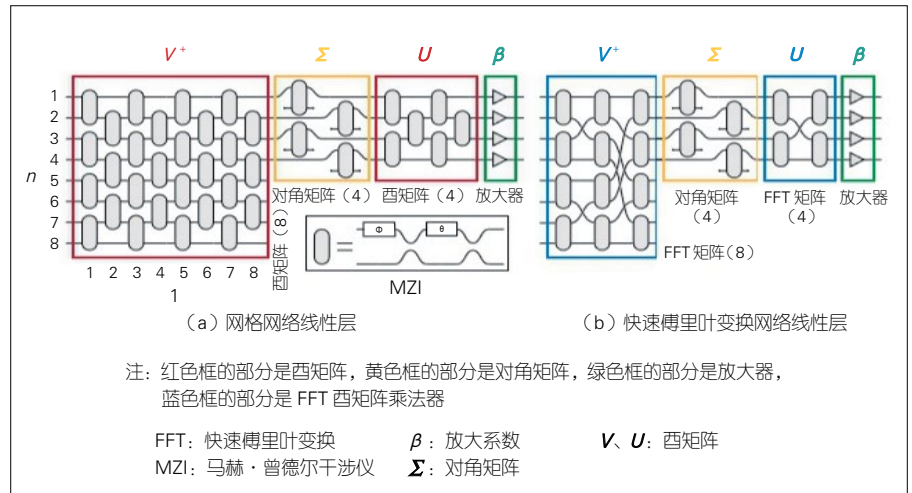
低; FFNetNet通用性差, 但在误差存在的情况下, 准确率高。

2.2 MRR型片上人工神经网络

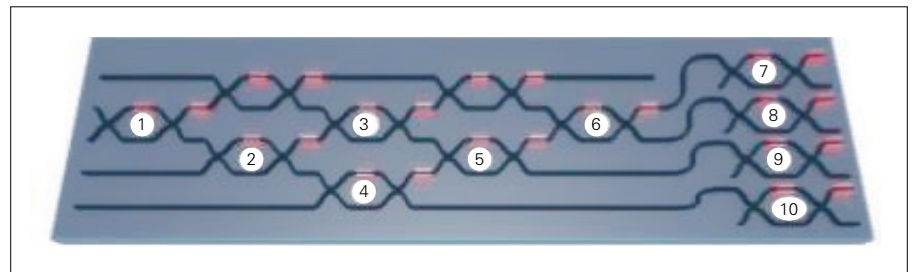
MRR神经网络主要用来实现脉冲神经网络(SNN)。这种网络考虑了时间信息, 相比于FNN和RNN更加接近于真实的人脑运作情况, 被称为

第3代神经网络^[9]。

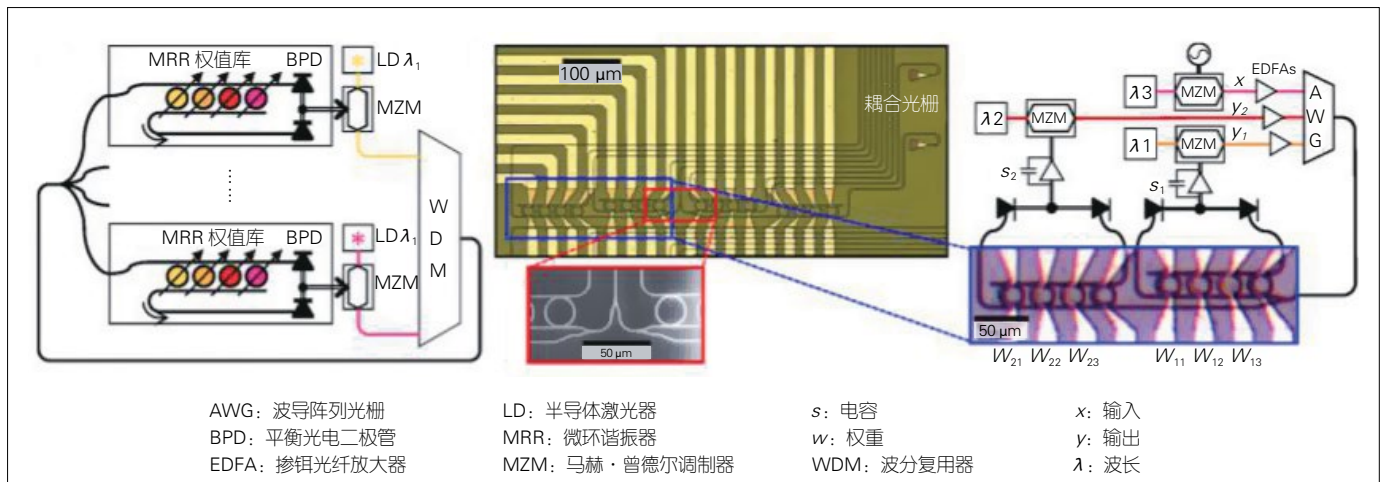
图7所示为A. N. TAIT等于2017年提出的一种广播式MRR权值库结构的神经网络^[10]。这是一种以MRR调制器作为神经元, 由MRR权值库连接而成的网络。每一个MRR都承担着一个权值, 每一横条聚集在一起的MRR叫权值库。该芯片结构包含4个节点,



▲图5 两种不同类型的模拟片上人工神经网络架构



▲图6 全连接神经网络芯片结构



▲图7 MRR权值库结构

带有 16 个 MRR。该结构证明了硅光子电路与连续神经网络模型之间存在数学同构关系。根据这种同构性,我们利用“神经编译器”对一个模拟的 24 节点硅光子神经网络进行编程,完成了微分系统仿真任务。根据推算,与传统的解决相同问题的 CPU 相比,此结构的处理速度将提高 294 倍。

2019 年, A. N. TAIT 等提出了一种神经拟态的片上结构^[11], 该结构主要由 2 个光探测器和 1 个 MRR 组成, 如图 8 (a) 所示。神经元阵列由电脉冲强度调控单元及延时单元构成, 除泵浦激光器外, 整体网络可实现片上集成。每个 MRR 只有一个波长 (λ_i)。该结构将 MRR 强度调制器和平衡的光电探测器组成电光脉冲强度和延迟调控单元, 并使用电光脉冲进行调控, 以实现复杂的脉冲神经网络。当输入为 2 ns 脉冲偶极子, 第 2 次的输入相比于第 1 次输入延迟一个波长, 进而产生 $t = 0$ 时的脉冲重合。据此测得的加强、饱和、抑制 3 种情况下的结果如图 8 (b) 所示。在图 8 (b) 中, 在

可见增强情况下, 脉冲出现过冲现象 (超过 57%); 而在饱和情况下, 脉冲只为输入脉冲和的 56%, 且单脉冲抑制不完全。虽然存在一些问题, 但是这种结构形成了全光广播权值神经网络的组件类型, 在一个集成的光子组件中包含了光到光的非线性、扇入和非确定性级联, 实现了光子神经元的网络兼容的能力。

目前 MRR 人工神经网络偏向于贴合神经网络的数学同构模型的研究, 采用比较统一的采用扇入结构、光电光转换等实现方案。

2.3 MZI 型与 MRR 型片上人工神经网络的对比

MMI 型片上人工神经网络是根据酉矩阵分解和计算来设计数学同构性的; 而 MRR 型片上人工神经网络则直接以普通矩阵的计算来设计数学同构性。两者本质上都是用光器件来表现数学计算。由于 MRR 型仍在结构探索阶段, 相比于 MZI 型, 准确性远远不足; 而 MZI 型有较为成熟的应用测试,

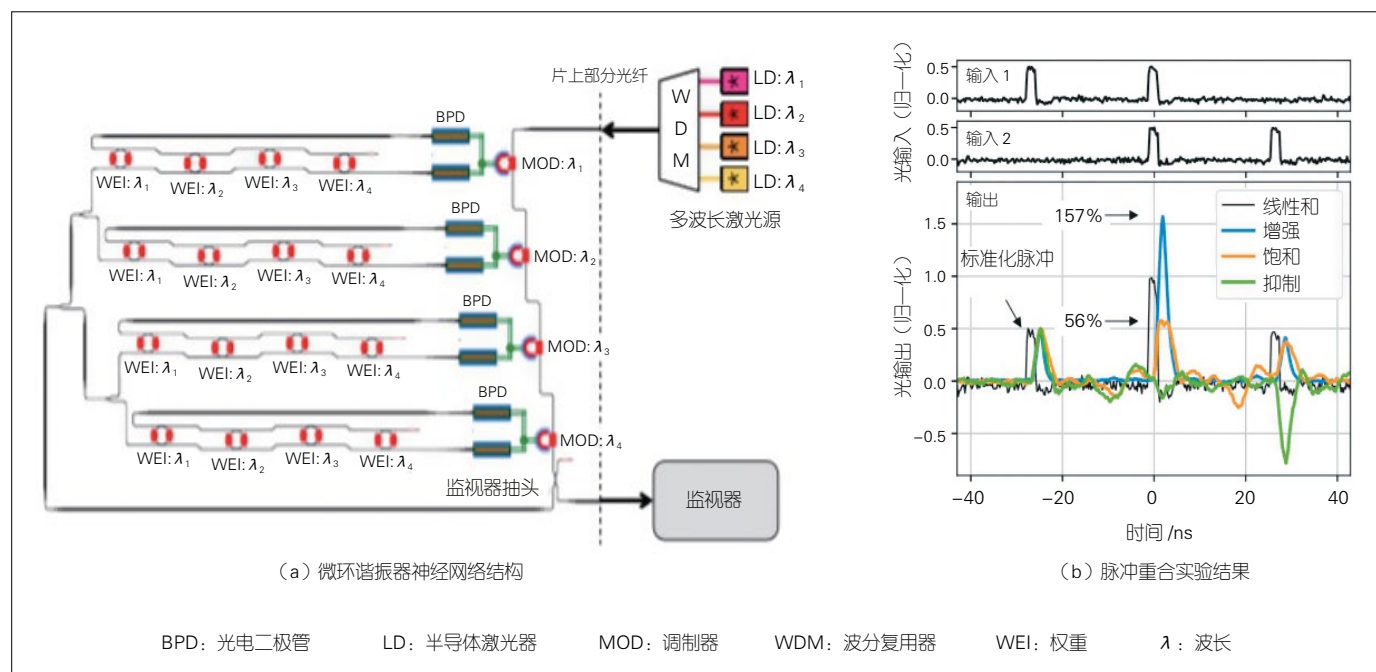
但相比于传统芯片, 准确性仍有不足。

3 片上人工神经网络实现方案面临的挑战

利用光子进行计算具有诸多优势, 但目前仍存在一些问题:

(1) 非线性激活函数是用来增加神经网络非线性的一种 S 形状的函数, 它的硬件实现起来比较困难。现在非线性函数的实现方式分为两种: 一种就是转换到电域再处理^[7]或利用电域辅助处理^[12]; 另一种就是利用特殊材料, 如可饱和吸收体和石墨烯等进行处理。一方面, 光电转换限制了数据处理速度的进一步提升; 另一方面, 大部分特殊材料的片上集成较为困难, 不能与互补金属氧化物半导体 (CMOS) 工艺兼容。

(2) MZI 的长度约为 200 μm , MRR 的长度约为 25 μm 。相比于电域的器件, 芯片集成度差, 目前工艺方面还有进一步提升的空间。虽然看起来 MRR 要比 MZI 小一些, 但是它们基本都属于一个数量级。另外, 由于



▲ 图 8 微环谐振器神经网络及其实验效果

目前硅基集成光器件的工艺仍旧不够成熟,器件的一致性、稳定性较差。

(3) 目前我们需要对光电人工智能芯片的匹配算法和外围电路进行设计^[13],并需要将各领域技术深度融合。这个结合的过程需要重新进行布局和设计,在目前没有统一标准的情况下。每一个光网络结构的出现都可能会导致外围匹配的电路和算法重新被调整和优化。

4 结束语

光电神经网络能够利用光子技术的优点并配合外围电路域进行处理,在提升计算速度的同时也可以降低运行功耗。无论是基于 FNN 的 MZI 前向人工神经网络芯片,还是基于 SNN 的 MRR 神经拟态人工神经网络芯片,都可以利用硅基光电子技术进行实现。此外,光电神经网络亦会随着硅基光电子技术的成熟而不断取得突破,如硅基片上光源、放大器、硅基单片集成、硅基新材料融合等新型硅基光电子技术,都将为光电神经网络的物理研究提供崭新的、开阔的思路。同时,随着与光电神经网络相匹配的算法演进,相信在将来的研究中,硅基光电子技术、硅基光电子芯片将为人工智能领域带来全新的技术架构和重大的产业升级。

参考文献

- [1] AL-QIZWINI M, BARJASTEH I, AL-QASSAB H, et al. Deep learning algorithm for autonomous driving using GoogLeNet [C]//2017 IEEE Intelligent Vehicles Symposium (IV). Los Angeles, USA: IEEE, 2017. DOI:10.1109/ivs.2017.7995703
- [2] ESTEVA A, KUPREL B, NOVOA R A, et al. Dermatologist-level classification of skin cancer with deep neural networks [J]. Nature, 2017, 542(7639): 115–118. DOI: 10.1038/nature21056
- [3] RECK M, ZEILINGER A, BERNSTEIN H J, et al. Experimental realization of any discrete unitary operator [J]. Physical review letters, 1994, 73(1): 58. DOI:10.1103/physrevlett.73.58 DOI:10.1103/physrevlett.73.58
- [4] YANG L, JI R Q, ZHANG L, et al. On-chip CMOS-compatible optical signal processor [J]. Optics express, 2012, 20(12): 13560. DOI:10.1364/oe.20.013560
- [5] 白冰, 赵斌, 杨钊. 一种光子神经网络芯片以及数据处理系统: CN110503196A [P]. 2019-11-26
- [6] CLEMENTS W R, HUMPHREYS P C, METCALF B J, et al. Optimal design for universal multiport interferometers [J]. Optica, 2016, 3(12): 1460. DOI:10.1364/optica.3.001460
- [7] SHEN Y C, HARRIS N C, SKIRLO S, et al. Deep learning with coherent nanophotonic circuits [C]//2017 IEEE Photonics Society Summer Topical Meeting Series (SUM). San Juan, USA: IEEE, 2017: 441–447. DOI:10.1109/phosst.2017.8012714
- [8] FANG M Y S, MANIPATRUNI S, WIERZYNSKI C, et al. Design of optical neural networks with component imprecisions [J]. Optics express, 2019, 27(10): 14009. DOI:10.1364/oe.27.014009
- [9] MAASS W. Networks of spiking neurons: the third generation of neural network models [J]. Neural networks, 1997, 10(9): 1659–1671. DOI: 10.1016/s0893-6080(97)00011-7
- [10] TAIT A N, DE LIMA T F, ZHOU E, et al. Neuromorphic photonic networks using silicon photonic weight banks [J]. Scientific reports, 2017, 7: 7430. DOI: 10.1038/s41598-017-07754-z
- [11] TAIT A N, DE FERREIRA L T, NAHMIA M A, et al. Silicon photonic modulator neuron [J]. Physical review applied, 2019, 11(6): 064043. DOI: 10.1103/physrevapplied.11.064043
- [12] WILLIAMSON I A D, HUGHES T W, MINKOV M, et al. Reprogrammable electro-optic non-linear activation functions for optical neural networks [J]. IEEE journal of selected topics in quantum electronics, 2020, 26(1): 1–12. DOI:10.1109/jstqe.2019.2930455
- [13] 白冰, 赵斌, 杨钊. 一种计算电路以及数据运算方法: CN110597756A [P]. 2019

作者简介



白冰, 北京交通大学光波技术研究所在读博士研究生; 主要从事硅光集成计算器件、光电异构计算架构和光电融合神经网络算法等领域的研究; 已申请专利 12 项。



裴丽, 北京交通大学教授、博士生导师; 主要从事全光交换、特种光纤、光电器件及基于智能光纤传感的物联网的研究; 主持科研项目 10 余项, 发表 SCI、EI 论文 200 余篇。



左晓燕, 北京交通大学光波技术研究所在读博士研究生; 主要从事神经网络、光电器件领域的研究。

《中兴通讯技术》杂志（双月刊）投稿须知

一、杂志定位

《中兴通讯技术》杂志为通信技术类学术期刊。通过介绍、探讨通信热点技术，以展现通信技术最新发展动态，并促进产学研合作，发掘和培养优秀人才，为振兴民族通信产业做贡献。

二、稿件基本要求

1. 投稿约定

- (1) 作者需登录《中兴通讯技术》投稿平台：tech.zte.com.cn/submission，并上传稿件。第一次投稿需完成新用户注册。
- (2) 编辑部将按照审稿流程聘请专家审稿，并根据审稿意见，公平、公正地录用稿件。审稿过程需要 1 个月左右。

2. 内容和格式要求

- (1) 稿件须具有创新性、学术性、规范性和可读性。
- (2) 稿件需采用 WORD 文档格式。
- (3) 稿件篇幅一般不超过 6 000 字（包括文、图），内容包括：中、英文题名，作者姓名及汉语拼音，作者中、英文单位，中文摘要、关键词（3 ~ 8 个），英文摘要、关键词，正文，参考文献，作者简介。
- (4) 中文题名一般不超过 20 个汉字，中、英文题名含义应一致。
- (5) 摘要尽量写成报道性摘要，包括研究的目的、方法、结果 / 结论，以 150 ~ 200 字为宜。摘要应具有独立性和自明性。中英文摘要应一致。
- (6) 文稿中的量和单位应符合国家标准。外文字母的正斜体、大小写等须写清楚，上下角的字母、数据和符号的位置皆应明显区别。
- (7) 图、表力求少而精（以 8 幅为上限），应随文出现，切忌与文字重复。图、表应保持自明性，图中缩略词和英文均要在图中加中文解释。表应采用三线表，表中缩略词和英文均要在表内加中文解释。
- (8) 所有文献必须在正文中引用，文献序号按其在文中出现的先后次序编排。常用参考文献的书写格式为：
 - 期刊 [序号] 作者. 题名 [J]. 刊名, 出版年, 卷号 (期号): 引文页码. 数字对象唯一标识符
 - 书籍 [序号] 作者. 书名 [M]. 出版地: 出版者, 出版年: 引文页码. 数字对象唯一标识符
 - 论文集中析出文献 [序号] 作者. 题名 [C]// 论文集编者. 论文集名 (会议名). 出版地: 出版者, 出版年 (开会年): 引文页码. 数字对象唯一标识符
 - 学位论文 [序号] 作者. 题名 [D]. 学位授予单位所在城市名: 学位授予单位, 授予年份. 数字对象唯一标识符
 - 专利 [序号] 专利所有者. 专利题名: 专利号 [P]. 出版日期. 数字对象唯一标识符
 - 国际、国家标准 [序号] 标准名称: 标准编号 [S]. 出版地: 出版者, 出版年. 数字对象唯一标识符
- (9) 作者超过 3 人时，可以感谢形式在文中提及。作者简介包括：姓名、工作单位、职务或职称、学历、毕业于何校、现从事的工作、专业特长、科研成果、已发表的论文数量等。
- (10) 提供正面、免冠、彩色标准照片一张，最好采用 JPG 格式（文件大小超过 100 kB）。
- (11) 应标注出研究课题的资助基金或资助项目名称及编号。
- (12) 提供联系方式，如：通讯地址、电话（含手机）、Email 等。

3. 其他事项

- (1) 请勿一稿多投。凡在 2 个月（自来稿之日算起）以内未接到录用通知者，可致电编辑部询问。
- (2) 为了促进信息传播，加强学术交流，在论文发表后，本刊享有文章的转摘权（包括英文版、电子版、网络版）。作者获得的稿费包括转摘酬金。如作者不同意转摘，请在投稿时说明。
- (3) 编辑部地址：安徽省合肥市金寨路 329 号凯旋大厦 1201 室，邮政编码：230061。
- (4) 联系电话：0551-65533356，联系邮箱：magazine@zte.com.cn。
- (5) 本刊只接受在线投稿，欢迎访问本刊投稿平台：tech.zte.com.cn/submission。

中兴通讯技术

(ZHONGXING TONGXUN JISHU)

办刊宗旨：

以人为本，荟萃通信技术领域精英
迎接挑战，把握世界通信技术动态
立即行动，求解通信发展疑难课题
励精图治，促进民族信息产业崛起

双月刊 1995 年创刊 总第 156 期
2021 年 2 月 第 27 卷 第 1 期

主管：安徽出版集团有限责任公司
主办：时代出版传媒股份有限公司
深圳航天广宇工业有限公司
出版：安徽科学技术出版社
编辑、发行：中兴通讯技术杂志社

总编辑：王喜瑜
主编：蒋贤骏
执行主编：黄新明
责任编辑：徐烨
编辑：杨广西、卢丹、朱莉、任溪溪
设计排版：徐莹
发行：王萍萍
外联：卢丹
编务：王坤

《中兴通讯技术》编辑部
地址：合肥市金寨路 329 号凯旋大厦 1201 室
邮编：230061
网址：tech.zte.com.cn
投稿平台：tech.zte.com.cn/submission
电子信箱：magazine@zte.com.cn
电话：(0551)65533356

传真：(0551)65850139
发行方式：自办发行
印刷：合肥添彩包装有限公司
出版日期：2021 年 2 月 10 日
中国标准连续出版物号：ISSN 1009-6868
CN 34-1228/TN
定价：每册 20.00 元