



北航学报
赠阅

ISSN 1001-5965
CODEN BHHDE8

北京航空航天大学 学报

JOURNAL OF BEIJING UNIVERSITY OF
AERONAUTICS AND ASTRONAUTICS



2020-09

Vol.46 No.9

目次

面向视频专网的网络安全态势指标体系的构建	李欣, 段詠程, 黄淑华, 樊志杰 (1625)
面向个体人员特征的跨模态目标跟踪算法	周千里, 张文靖, 赵路平, 田乃倩, 王蓉 (1635)
基于道路特征信息的车道结构化解析	罗胜, 赵丽, 王慕抽 (1643)
基于高置信局部特征的车辆重识别优化算法	窦鑫泽, 盛浩, 吕凯, 刘洋, 张洋, 吴玉彬, 柯韦 (1650)
基于 SM4 算法的白盒密码视频数据共享系统	吴震, 白健, 李大双, 李斌, 曾兵, 张正强 (1660)
基于 PF_RING 的多核视频流高性能传输模型	李欣, 樊志杰, 曹志威, 胡正梁, 陈国梁 (1670)
基于迁移学习的暴恐图像自动识别	陈猛夫 (1677)
基于局部语义特征不变性的跨域行人重识别	张晓伟, 吕明强, 李慧 (1682)
基于三元组网络的单图三维模型检索	杜雨佳, 李海生, 姚春莲, 蔡强 (1691)
基于结构加权低秩近似的泊松图像去模糊	吴庆波, 任文琦 (1701)
基于收缩场学习的 Retinex 低照度图像增强	吴庆波, 王蕊, 任文琦 (1711)
热红外视频监控下行人目标前景区域提取	张玉贵, 沈柳青, 胡海苗 (1721)
联合足迹识别与监控视频分析的智能刑侦系统	陶一宁, 苏峰, 袁培江, 王田苗, 钟涛, 郝静如 (1730)
双光谱智能体温检测与健康大数据管理系统	张洁茹, 苏峰, 袁培江, 王田苗, 陶一宁, 丁东 (1739)
基于空间注意力机制的行人再识别方法	张子昊, 周千里, 王蓉 (1747)
基于 MobileFaceNet 网络改进的人脸识别方法	张子昊, 王蓉 (1756)
一种基于区域划分的改进 ORB 算法	孙浩, 王朋 (1763)
基于面部视频分析的生命体征检测	陈辉, 郑秀娟, 倪宗军, 张昀, 杨晓梅 (1770)
一种轻量化的多目标实时检测模型	邱博, 刘翔, 石蕴玉, 尚岩峰 (1778)
基于跨尺度特征聚合网络的多尺度行人检测	曹帅, 张晓伟, 马健伟 (1786)
基于联合正则化策略的人脸表情识别方法	兰凌强, 李欣, 刘淇缘, 卢树华 (1797)
基于 U-Net 的掌纹图像增强与 ROI 提取	陆展鸿, 单鲁斌, 苏立循, 焦雨欣, 王家骅, 王海霞 (1807)

CONTENTS

Construction of network security situation indicator system for video private network
..... LI Xin, DUAN Yongcheng, HUANG Shuhua, FAN Zhijie (1625)

Cross-modal object tracking algorithm based on pedestrian attribute
..... ZHOU Qianli, ZHANG Wenjing, ZHAO Luping, TIAN Naiqian, WANG Rong (1635)

Lane semantic analysis based on road feature information
..... LUO Sheng, ZHAO Li, WANG Muchou (1643)

Vehicle re-identification optimization algorithm based on high-confidence local features
..... DOU Xinze, SHENG Hao, LYU Kai, LIU Yang, ZHANG Yang, WU Yubin, KE Wei (1650)

White-box cryptographic video data sharing system based on SM4 algorithm
..... WU Zhen, BAI Jian, LI Dashuang, LI Bin, ZENG Bing, ZHANG Zhengqiang (1660)

High-performance multi-core video stream transmission model based on PF_RING
..... LI Xin, FAN Zhijie, CAO Zhiwei, HU Zhengliang, CHEN Guoliang (1670)

Automatic recognition for terrorism related image based on transfer learning
..... CHEN Mengfu (1677)

Cross-domain person re-identification based on partial semantic feature invariance
..... ZHANG Xiaowei, LYU Mingqiang, LI hui (1682)

Monocular image based 3D model retrieval using triplet network
..... DU Yujia, LI Haisheng, YAO Chunlian, CAI Qiang (1691)

Structural weighted low-rank approximation for Poisson image deblurring
..... WU Qingbo, REN Wenqi (1701)

Learning shrinkage fields for low-light image enhancement via Retinex
..... WU Qingbo, WANG Rui, REN Wenqi (1711)

Extraction of foreground area of pedestrian objects under thermal infrared video surveillance
..... ZHANG Yugui, SHEN Liuqing, HU Haimiao (1721)

Intelligent criminal investigation system based on both footprint recognition and surveillance video analysis
..... TAO Yining, SU Feng, YUAN Peijiang, WANG Tianmiao, ZHONG Tao, HAO Jingru (1730)

Dual-spectrum intelligent temperature detection and health big data management system
..... ZHANG Jieru, SU Feng, YUAN Peijiang, WANG Tianmiao, TAO Yining, DING Dong (1739)

Pedestrian re-identification method based on spatial attention mechanism
..... ZHANG Zihao, ZHOU Qianli, WANG Rong (1747)

Improved face recognition method based on MobileFaceNet network
..... ZHANG Zihao, WANG Rong (1756)

An improved ORB algorithm based on region division
..... SUN Hao, WANG Peng (1763)

Vital signs detection via facial video analysis
..... CHEN Hui, ZHENG Xiujuan, NI Zongjun, ZHANG Yun, YANG Xiaomei (1770)

A lightweight multi-target real-time detection model
..... QIU Bo, LIU Xiang, SHI Yunyu, SHANG Yanfeng (1778)

Trans-scale feature aggregation network for multiscale pedestrian detection
..... CAO Shuai, ZHANG Xiaowei, MA Jianwei (1786)

Facial expression recognition method based on a joint normalization strategy
..... LAN Lingqiang, LI Xin, LIU Qiyuan, LU Shuhua (1797)

Palmprint enhancement and ROI extraction based on U-Net
..... LU Zhanhong, SHAN Lubin, SU Lixun, JIAO Yuxin, WANG Jiahua, WANG Haixia (1807)

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0043

面向视频专网的网络安全态势指标体系的构建



李欣^{1,2,*}, 段詠程¹, 黄淑华^{1,2}, 樊志杰³

1. 公安大学 信息技术与网络安全学院, 北京 100038; 2. 安全防范技术与风险评估公安部重点实验室, 北京 100038;

3. 公安部第三研究所 信息安全技术部, 上海 200031)

摘 要: 针对当前面向视频专网的网络安全态势评估中缺乏一套完善的指标体系问题,对构建网络安全态势指标体系进行了研究。首先,重点对视频专网面临的网络安全风险进行了研究与分析;然后,结合以往网络安全态势评估技术中的基础评价指标,针对视频专网自身特点,构建了一套适用于视频专网的网络安全态势指标体系;最后,给出了指标值的计算方法,为视频专网的网络安全态势评估奠定基础。研究成果对于完善和丰富面向视频专网的网络安全态势指标体系具有一定的参考和借鉴意义。

关 键 词: 指标体系; 视频专网; 态势要素; 态势感知; 态势评估

中图分类号: TP393.08

文献标志码: A

文章编号: 1001-5965(2020)09-1625-10

视频专网承载了道路监控、治安监控等大量公安业务,是辅助侦破案件的利器。但是在前期视频专网的建设过程中,普遍只注重视频监控系统的应用,而对于整个视频专网信息安全问题缺乏重视^[1],视频专网的安全保障工作存在一些亟待解决的问题:监控系统初始密码未修改,应急处理能力弱,特别是缺乏整体的网络安全态势指标体系。为了保障视频专网的安全,各地公安机关已经开始加强视频专网的安全管理建设,建立面向视频专网的网络安全态势评估系统。然而,建立网络安全态势评估系统首先需要一套完善的指标体系,目前视频专网的网络安全态势指标体系往往采用传统的网络安全态势指标体系^[2],应用效果不佳,适用性较差,且在此领域缺少相关研究。

为此,本文通过对视频专网的安全框架、面临的安全风险和传统的网络安全态势评估系统进行分析,采取系统、科学的方法^[3],建立科学完善的面向视频专网的网络安全态势指标体系。

1 视频专网的安全组成部分及威胁

公安视频专网是指采用专线方式或虚拟专用网方式,建设在公安信息网之外,主要用于传输公安视频图像、汇接各层级视频图像信息系统、支撑公安视频图像服务的非涉密网络。公安视频专网可以根据需要传输视频图像之外的其他相关公安业务数据。

1.1 视频专网的安全组成部分

公安视频专网各级网络内部安全域划分为:纵向边界区、横向边界区、系统应用区和前端接入区4部分。具体划分如图1所示。

纵向边界区主要实现本级公安视频传输网与上下级公安视频传输网的安全连接和访问控制。横向边界区主要实现公安视频传输网与公安信息网、互联网及其他专网之间信息的安全交互。系统应用区内的设备主要包括公安视频传输网内相关的网络设备、终端、通用服务器、存储设备、应用

收稿日期: 2020-02-21; 录用日期: 2020-03-20; 网络出版时间: 2020-04-26 08:58

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200424.2052.001.html

基金项目: 公安部科技强警基础工作专项项目(2017GABJC38); 中国人民公安大学新任教师科研启动基金(2018JKF609)

* 通信作者: E-mail: ndlixin@sina.com

引用格式: 李欣, 段詠程, 黄淑华, 等. 面向视频专网的网络安全态势指标体系的构建[J]. 北京航空航天大学学报, 2020, 46(9): 1625-1634. LI X, DUAN Y C, HUANG S H, et al. Construction of network security situation indicator system for video private network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1625-1634 (in Chinese).

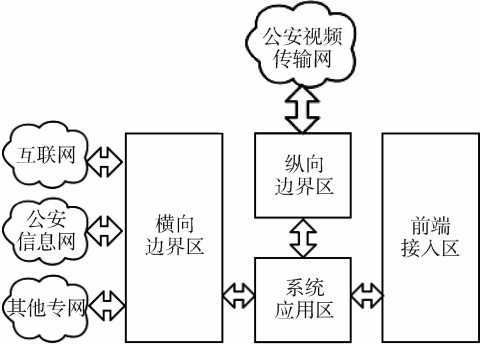


图 1 视频专网安全框架

Fig. 1 Security framework of video private network

系统/数据库服务器等设备。前端接入区主要包括前端接入的摄像机及其他设备。

1.2 视频专网的安全威胁

相对民用安防视频监控领域而言,公安视频专网由于承载着大量公安重要业务,存储有海量的敏感视频数据,面临的网络安全隐患更为突出。根据公安视频专网网络结构、覆盖范围、设备组成及业务特征等进行分析,当前主要存在以下几个方面的威胁:

1) 设备底数不清、管理不规范。由于采用以租代建模式,摄像头多数通过运营商购置、安装和更换,管理部门只有运营商上报的数据,缺乏有效技术手段进行核实。此外,终端的接入也完全没有控制,缺乏注册管理,管理部门对摄像头、电脑存在底数不清、难于监管的问题。

2) 网络非法互联风险。公安视频专网不能通过公安信息网、互联网和其他专网边界接入平台及防火墙以外的网络安全设施,与公安信息网、互联网等其他网络进行信息交互、资源共享,公安视频专网设计上与公安信息网及不同网络之间保持物理隔离。在日常工作中,公安视频专网与外部网络互联的情况已经不足为奇。通过绕过安全监管设施与外部网络进行非法通信的形式主要有网卡、网络代理、路由、NAT 边界、无线 AP 等。以上非法通信形式严重造成了视频专网信息的泄露,破坏了视频专网的物理隔离性,加快了木马病毒的传播速度,甚至使视频专网成为外部攻击的目标。因此,严格管理公安视频专网与其他网络之间的边界平台成为一项保障视频专网安全的重点工作。

3) 设备接入安全风险。由于视频专网承载了治安监控、打击犯罪等大量的业务功能,摄像头等前端设备部署分布极为广泛,大部分地区处于无人看守状态,且大部分前端摄像头为网络摄像机,极易发生网络入侵和非法访问等问题,非授权人员仅将前端摄像头替换为自己的终端就可以轻

松侵入视频专网。

4) 应用系统漏洞风险。视频设备大多以嵌入式操作系统为基础,本身存在漏洞可能,其上运行的应用也可能有漏洞;应用服务端的操作系统和应用本身也会存在漏洞,容易成为被攻击对象。

5) 计算机基础安全风险。工作人员日常管理和使用视频专网主要通过视频专网的终端计算机(服务器)。终端计算机(服务器)是系统应用区的主要组成部分,为整个视频专网提供各类应用服务。当前终端计算机(服务器)主要安全威胁包括系统漏洞、计算机病毒、弱口令和空口令、木马程序、USB 移动存储设备滥用、违规外联等。如果终端计算机(服务器)被非授权人员入侵并操控,也就意味着整个视频专网被非授权人员控制,进而可能会导致公安信息网及其他专网被入侵攻击,造成信息泄露、非法控制等威胁国家、社会和个人安全的情况。

6) 数据安全风险。视频设备拍摄的大量隐私、敏感的视频数据被传输到视频专网中,一旦视频专网发生安全事故,将对国家安全、社会安全和个人隐私造成严重的危害。当前对数据安全的保护措施主要通过审计记录的方式,工作人员每次访问视频数据将会被视频共享平台记录,但缺乏工作人员对视频数据进行其他操作的完整有效的管理与审计,如复制、下载、外发等。

7) 管理安全风险。视频专网中缺乏整体安全管理手段,大部分视频专网中没有建立一套体系化、有效联合各方工作人员协同参与的应急响应手段,一旦出现安全事件或设备故障时,管理员、运营商与工程商运维人员职责不明确,相关管理员无法从全局上把握视频专网运行安全态势。

8) 设备运维保障风险。视频设备的运行状态直接影响到工作人员对于视频监控系统的使用效果,影响办案工作效率,甚至可能会影响侦破案件的成功率。常见的设备运维保障风险有画面不清晰、监控设备像素过低、图像失真、图像偏色、硬盘异常、线路信号衰减、存在遮挡物、监控设备时钟失步、监控录像存储时间过短、监控录像重复存储等。

而对于民用安防视频监控领域,主要存在的风险则是视频设备被黑客控制、视频被窥视、盗窃和非法传播,导致隐私泄露。对目前中央社会治安综合治理委员会办公室大力推进的社会视频接入、雪亮工程等项目,如这些视频前端将来接入视频专网,则同样存在上述公安视频专网中存在的安全风险。

2 典型网络安全态势指标体系分析

典型网络安全态势指标体系采用定性与定量分析相结合的方法,分为基于风险的指标体系和基于性能的指标体系2类。

2.1 基于风险的指标体系

基于安全风险的态势评价指标^[4]较为常见,体现了网络安全态势通用评估方法。其一般步骤是:将网络按不同的层次结构进行划分,利用采集和检测工具收集各种网络攻击、日志信息和漏洞信息,通过一定的预处理和数学计算,将以上几种信息转化成网络安全态势量化数据,以此来代表网络系统当前的安全状态。该类型指标着重于对受网络攻击威胁严重的安全态势进行评估,主要涉及主机脆弱性、攻击威胁度、漏洞利用等数据。在该类型指标体系中,网络安全态势评估的目的是通过一定的数学函数和计算模型,综合计算各类型数据,从而得到态势值,并将结果进行可视化展现。

2.2 基于性能的指标体系

基于网络和主机性能的态势评价指标^[5]着重于对网络本身和网络节点性能进行态势评估,一般包括网络端口流量、主机使用率、主机内存使用率和网络负载状况等指标。该类型指标数据大多是通过现有的网络监控系统或者检测系统等安全设备收集和捕获到的网络和主机的性能数据,以及利用日志系统采集并分析得到的日志统计信息。由于在使用该类型指标数据时,需要采集大量数据,采集过程复杂且对实时性要求高,加之数据指标过多,实现起来也较为复杂。基于日志系统进行的日志审计对采集的要求高,且审计规则过于简单,对深层次的网络事件原因挖掘力度低下,常常难以满足网络安全态势对数据源实时性的要求。

3 视频专网网络安全态势指标体系

3.1 指标体系结构

本文采用指数方式对面向视频专网的网络安全态势指标体系结构进行描述,在借鉴传统网络安全态势指标体系的基础上,针对视频专网的组成和安全风险特点,将视频专网的网络安全态势指标体系按照层次化分析方法分为4个层次,分别为Ⅰ级指数层、Ⅱ级指数层、Ⅲ级指数层和指数要素层。Ⅰ级指数层用来反映视频专网在某时刻或某段时间内的整体网络安全态势情况,包含的指数对象是网络安全态势的综合指数。Ⅱ级指数

层是从视频专网的某个侧面进行刻画描述,包含的指数对象是网络安全态势综合指数的子指数集合,如系统脆弱性指数、网络安全威胁指数、前端设备运行安全指数等。Ⅲ级指数层是针对视频专网某个侧面刻画进行深入一层的分析描述,包含的指数对象是Ⅱ级指数的子指数集合,如网络安全威胁指数对应的Ⅲ级指数集合包括木马指数、网络攻击指数等。指数要素层反映的是某个Ⅲ级指数方面的深度分析与描述的量化描述,包含的对象是Ⅲ级指数的要素集合,如网络攻击指数所对应的指数要素集合包括目标资产情况、攻击严重程度、攻击强度、安全事件发生频率等统计数据。

3.2 指标体系构建

传统的网络安全评估中^[6-8],在构建网络安全态势指标时重点考虑了基础网络运行安全指数、系统脆弱性指数和网络安全威胁指数3个方面。本文参考网络安全法和网络安全等级保护2.0^[9]的相关要求,在一般IT系统的网络安全风险分析的基础上,针对视频专网的网络安全风险特点,提出了视频专网的前端设备运行安全指数和边界运行安全指数。本文中视频专网的网络安全态势指标体系结构总表如表1所示。

在指标体系构建完成后,参考文献[10]中的方法,确定和优化指标体系中各指标在每层中的权重。首先通过专家咨询法初步确定指标体系模糊权重,然后使用层次化分析法对模糊权重进行检验与优化,最后再利用专家咨询法对权重进行修正,得出最终各指标在每个层次的权重。构建指标体系的步骤如下:

步骤1 确定指标体系中的各层指标。采用层次化分析法中经典的1-9标度法,使用德尔菲调查法明确视频专网的网络安全态势评估需求,进而确定网络安全态势指标体系中各层的具体指标。

步骤2 使用三角模糊数优化指标权重。通过1-9标度法完成指标体系中指标的设计、分解、分层、描述、编码等工作后,通过专家咨询法得到指标体系中同一层各指标重要性的两两比较的比值。在此基础上,引入三角模糊数概念。假设集合 $a_{ij} = (L_{ij}, M_{ij}, U_{ij})$, $L_{ij} \leq M_{ij} \leq U_{ij}$, L_{ij} 、 M_{ij} 、 U_{ij} 分别为 a_{ij} 对应的上限值、中间值、下限值, a_{ij} 为三角模糊数。对通过专家咨询法得到的初步权重值用三角模糊数表示如下:

$$L_{ij} = \min(a_{ijk}) \quad (1)$$

$$M_{ij} = \sqrt[n]{\prod_{k=1}^n a_{ijk}} \quad (2)$$

表 1 视频专网安全态势指标体系结构

Table 1 Architecture of security situation indicator system for video private network			
I 级指标	II 级指标	III 级指标	指数要素
网络安全态势 综合指数	基础网络运行 安全指数	网络异常流量指数	平均流量、峰值流量
		资源消耗指数	带宽使用率、CPU 使用率、内存使用率
	系统脆弱性指数	漏洞状态指数	漏洞固有威胁、资产重要性
		木马指数	目标资产情况、安全事件发生频率、可清除性
	网络安全威胁指数	网络攻击指数	目标资产情况、攻击严重程度、攻击强度、安全事件发生频率、攻击防范代价
		病毒指数	目标资产情况、病毒破坏性、安全事件发生频率、可清除性
	前端设备运行 安全指数	服务状态安全指数	前端设备可用性、故障发生频率
		异常状态安全指数	前端设备存在漏洞状态、前端设备存在弱口令状态、前端设备存在违规接入状态
	边界运行安全指数	公安信息网边界安全指数	边界设备可用性、边界非法接入状态
		互联网边界安全指数	边界设备可用性、边界非法接入状态
		其他专网边界安全指数	边界设备可用性、边界非法接入状态

$U_{ij} = \max(a_{ijk})$ (3)

式中: a_{ijk} 为第 k 个专家判断得出的 a_i 和 a_j 两要素的相对重要性的数值。

步骤 3 使用去模糊化方法优化指标权重。通过构建模糊判别矩阵来完成去模糊化工作。判别矩阵的元素表示该层次各因素相对重要性的比值,模糊判别矩阵的元素为抽象的三角模糊数。判别矩阵是层次分析进行的前提,假设 B 层中 B_k 项和下一层中 $(c_1, c_2, \cdots, c_n)$ 有联系,可以得到判别矩阵 C_n 为

$$C_n = \begin{bmatrix} c_{11} & \cdots & c_{1n} \\ \vdots & & \vdots \\ c_{n1} & \cdots & c_{nn} \end{bmatrix}$$

式中: c_{ij} 表示对上层 B_k 项而言, c_i 对 c_j 的相对重要性的三角模糊数值表现形式。

三角模糊数不能直接用于层次化分析计算,采用文献[11]中的方法进行去模糊化:

$$(a_{ij}^\beta)^\mu = [\mu L_{ij}^\beta + (1 - \mu) U_{ij}^\beta] \quad (4)$$

$$(a_{ij}^\beta)^\mu = 1 / \left(\beta_{ji} \right)^\mu \quad (5)$$

式中:

$$L_{ij}^\beta = (M_{ij} - L_{ij})\beta + L_{ij} \quad (6)$$
$$U_{ij}^\beta = U_{ij} - (U_{ij} - M_{ij})\beta \quad (7)$$

其中: $0 \leq \beta \leq 1, 0 \leq \mu \leq 1, i < j; \beta$ 为偏好系数; μ 为风险容忍度,其值越小说明判别矩阵可信程度越高。

去模糊化后,判别矩阵如下:

$$C_n = \begin{bmatrix} 1 & (a_{12}^\beta)^\mu & \cdots & (a_{1n}^\beta)^\mu \\ (a_{21}^\beta)^\mu & 1 & \cdots & (a_{2n}^\beta)^\mu \\ \vdots & \vdots & & \vdots \\ (a_{n1}^\beta)^\mu & \cdots & \cdots & (a_{nn}^\beta)^\mu \end{bmatrix}$$

步骤 4 使用和积法计算优化指标权重。在去模糊化的判别矩阵基础上,采用和积法计算其最大特征值和特征向量,公式如下:

$$CI = (\mu_{\max} - n) / (n - 1) \quad (8)$$

$$CR = CI / RI \quad (9)$$

式中:CI 为一致性指标;CR 为随机一次性比率;RI 为平均随机一致性指标,RI 的取值如表 2 所示。

表 2 RI 取值

Table 2 RI value	
n	RI
1	0
2	0
3	0.58
4	0.90
5	1.12
6	1.24
7	1.32
8	1.41

求解过程中, n 为判别矩阵的阶数,当 $CR < 0.1$ 时,判别矩阵的一致性可接受,否则返回步骤 3 的模糊判别矩阵构建,对其进行重新优化。

3.2.1 网络安全态势综合指数

网络安全态势综合指数是指在某个特定时间周期内,对视频专网区域内受到的各种与网络安全相关的影响因素进行综合的科学评估和量化之后,得到的一个能够反映视频专网的整体网络安全态势的数值。

网络安全态势综合指数在本文中属于网络安全态势指标体系的 I 级指标,即网络安全态势的总指标,能够直接和直观地反映出视频专网的整体网络安全状态。网络安全态势综合指数包括基础网络运行安全指数、系统脆弱性指数、网络安全威胁指数、前端设备运行安全指数、边界运行安全

指数 5 个Ⅱ级指标,其具体概念如下:

1) 基础网络运行安全指数。指在某个特定时间周期内,对视频专网区域内的基础网络运行状态进行监控,对其受到的各种相关影响因素进行综合评估和量化之后,得到的能够反映视频专网的基础网络运行状态数值,如网络异常流量监控、资源消耗监控等。

2) 系统脆弱性指数。指在某个特定时间周期内,对视频专网区域内的系统脆弱性情况进行监控,对其受到的各种相关影响因素进行综合评估和量化之后,得到的能够反映视频专网的系统脆弱性情况的数值,如漏洞状态监控等。

3) 网络安全威胁指数。指在某个特定时间周期内,对视频专网区域内的来自内部和外部的攻击、渗透等威胁严重程度进行监控,对其受到的各种相关影响因素进行综合评估和量化之后,得到的能够反映视频专网的威胁严重程度的数值,如木马监控、网络攻击监控等。

4) 前端设备运行安全指数。指在某个特定时间周期内,对在视频专网的前端设备、网络和子系统的运行安全情况进行监控,对其受到的各种相关影响因素进行综合评估和量化之后,得到的能够反映前端设备运行的安全程度数值,如服务状态安全监控等。

5) 边界运行安全指数。指在某个特定时间周期内,对视频专网与外部网络接入通信的运行安全情况进行监控,对其受到的各种相关影响因素进行综合评估和量化之后,得到的能够反映视频专网与外部网络的接入通信的运行安全程度数值,如公安信息网边界安全监控、互联网边界安全监控等。

在得到Ⅱ级指标,即基础网络运行安全指数、系统脆弱性指数、网络安全威胁指数、前端设备运行安全指数、边界运行安全指数的量化值后,通过如下方式完成从Ⅱ指标到Ⅰ级指标网络安全态势综合指数的聚合过程:

$$Q(t) = f(q_1(t)w_1 + \cdots + q_i(t)w_i + \cdots + q_n(t)w_n)$$

(10)

式中: $Q(t)$ 为 t 时刻需要计算得到的指标状态值; $q_i(t)$ 为 t 时刻需要计算得到的指标的第 i 个子指标的状态值; w_i 为其对应的权重值; f 为标准归一化函数。

在使用式(10)计算网络安全态势综合指数时, $Q(t)$ 表示 t 时刻网络安全态势综合指数的状态值,也即态势值; $q_i(t)$ ($i=1,2,3,4,5$)表示基础网络运行安全指数、系统脆弱性指数、网络安全

威胁指数、前端设备运行安全指数、边界运行安全指数的状态值。在得到视频专网的网络安全态势综合指数数值之后,参考文献[12]的方法对其进行量化描述,量化过程如表3所示。

表 3 视频专网网络安全态势量化

Table 3 Network security situation quantification for video private network

安全级别	指数值	描述
优	0 ~ 2	网络整体运行正常,没有发现严重的网络攻击行为,没有严重威胁的安全漏洞发布,恶意代码活动保持较低水平等
良	2 ~ 4	网络整体运行受到轻微影响,发现少量较为严重的网络攻击行为,有少量且威胁级别较低的安全漏洞发布,恶意代码活动呈现一定的活跃性等
中	4 ~ 7.5	网络整体运行受到一定的影响,发现一些较为严重的网络攻击行为,有少量且威胁级别较高的安全漏洞发布,发现一些恶意代码活动等
差	7.5 ~ 9	网络整体运行受到一定的影响,局部网络严重受影响,网络发现大量且严重的攻击行为,有威胁级别很高的安全漏洞发布,发现大量恶意代码活动活跃等
危	9 ~ 10	网络整体运行受到严重影响,发现大量且严重的网络攻击行为,有大量威胁级别很高的安全漏洞发布,发现大量恶意代码活动活跃等

3.2.2 基础网络运行安全指数

基础网络运行安全指数包括网络异常流量指数、资源消耗指数 2 个Ⅲ级指标,其具体概念如下:

1) 网络异常流量指数。指当视频专网区域内的安全状态发生变化时,视频专网所接收、发送、中转的数据包流量会随着其变化而发生相应的变化,发生异常的流量能够对视频专网的网络安全状况进行刻画与描述,反映其数据流量的总体情况。

2) 资源消耗指数。对视频专网区域内的各种计算资源、运行环境资源等各种资源的可用性量化描述,反映的是视频专网中各种资源的运行消耗情况。

在得到Ⅲ级指标,即网络异常流量指数、资源消耗指数的量化值后,使用式(10)完成从Ⅲ级指标到Ⅱ级指标基础网络运行安全指数的聚合过程。本文后续的Ⅲ级指标均通过该方法完成从Ⅲ级指标到Ⅱ级指标的聚合。

1) 网络异常流量指数

在视频专网的网络安全态势评估过程中,针对其运行的业务和存在的安全风险,确定网络异常流量指数所包含的指数要素集合,具体包括平均流量、峰值流量 2 个指标要素,其具体概念如下:

① 平均流量。用来描述视频专网区域内在一个特定时间周期内接收、发送和中转的平均网络数据流量,用来描述视频专网在不同网络安全环境下一个阶段周期的平均流量状态情况。

② 峰值流量。用来描述视频专网区域内在一个特定时间周期内能够承受的最大网络数据流量,用来描述视频专网在最坏网络安全环境下的极限状态情况。

本文采用如下计算方式对平均流量、峰值流量进行量化:

$$AF_{\Delta t} = \frac{\sum_{i=t_1}^{t_2} F_i}{t_2 - t_1}$$

(11)

$$PF_{\Delta t} = \max(F_i)$$

(12)

$$A = \begin{cases} 1 & \frac{AF_{\Delta t}}{af} \leq 1 \\ 10 & \frac{AF_{\Delta t}}{af} > 1 \end{cases}$$

(13)

$$P = \begin{cases} 1 & \frac{PF_{\Delta t}}{pf} \leq 1 \\ 10 & \frac{PF_{\Delta t}}{pf} > 1 \end{cases}$$

(14)

式中: $\Delta t = t_2 - t_1$; F_i 为 i 时刻的网络流量值; $AF_{\Delta t}$ 为在时间周期 Δt 期间的平均流量值; $PF_{\Delta t}$ 为在时间周期 Δt 期间的最大流量峰值; A 为平均流量的具体量化值; af 为平均流量的阈值; P 为峰值流量的具体量化值; pf 为峰值流量的阈值。

在得到指数要素的量化值后,使用式(10)完成从指数要素到Ⅲ级指标网络异常流量指数的聚合过程。本文后续的指数要素除特殊说明外,均通过该方法完成从指数要素到Ⅲ级指标的聚合。

2) 资源消耗指数

在视频专网的网络安全态势评估过程中,针对其运行的业务和存在的安全风险,确定资源消耗指数所包含的指数要素集合,具体包括带宽使用率、CPU 使用率、内存使用率 3 个指标要素,其具体概念如下:

① 带宽使用率。用来描述视频专网区域内随时间变化的情况下,其网络带宽资源的消耗情况,用来衡量视频专网的网络带宽资源在一定时间周期内的变化情况。

② CPU 使用率。用来描述视频专网区域内随时间变化的情况下,其 CPU 计算资源的总体消耗情况,用来衡量视频专网的 CPU 计算资源在一定时间周期内的总体变化情况。

③ 内存使用率。用来描述视频专网区域内随时间变化的情况下,其内存资源的总体消耗情况,用来衡量视频专网的内存资源在一定时间周期内的总体变化情况。

本文采用如下步骤的计算方式对带宽使用率、CPU 使用率和内存使用率进行量化:

步骤 1 衡量资源消耗的严重程度。通过如下公式计算每个时间段内各个指数要素的过载度:

$$D_{ji} = \begin{cases} \frac{v_{ji}}{V_{ji}} & v_{ji} > V_{ji} \\ 1 & v_{ji} \leq V_{ji} \end{cases}$$

(15)

式中: $i = 1, 2, 3$ 表示带宽使用率、CPU 使用率、内存使用率 3 个指标数要素; V_{ji} 为属性 i 在节点 j 的阈值; v_{ji} 为属性 i 在节点 j 的真实值; D_{ji} 为属性 i 在节点 j 的过载度。

步骤 2 量化各个指数要素的过载度。通过表 4 对各个指数要素的过载度进行量化。

表 4 过载度量化

Table 4 Overload degree quantification	
D	量化数值
$D = 1$	1
$1 < D \leq 1.25$	2 ~ 5
$1.25 < D \leq 2$	6 ~ 7
$2 < D \leq 3$	8 ~ 9
$D = 4$	10

3.2.3 系统脆弱性指数

系统脆弱性指数包括漏洞状态指数一个Ⅲ级指标,其具体概念如下:

漏洞状态指数用来描述视频专网区域内存在的安全漏洞在处于网络攻击情况下,对目标资产造成了负面影响,其反映了视频专网的漏洞被利用的总体情况。

在视频专网的网络安全态势评估过程中,针对其运行的业务和存在的安全风险,确定漏洞状态指数所包含的指数要素集合,具体包括漏洞固有威胁和资产重要性 2 个指数要素,其具体概念如下:

1) 漏洞固有威胁。用来描述视频专网区域内的漏洞如果被利用后可能产生的总体危害情况。

2) 资产重要性。用来描述视频专网区域内的硬件、软件、数据等资产的重要性程度的总体量化情况。

本文采用如下计算方式对漏洞固有威胁和资

产重要性进行量化:

1) 漏洞固有威胁的量化方法是采用文献[13]中NESSUS的定义与方法进行量化评估,如表5所示。

表 5 漏洞固有威胁量化

Table 5 Vulnerability inherent threat quantification	
漏洞固有威胁等级	量化数值
注意	1~4
警告	5~8
漏洞	9~10

2) 资产重要性的量化方法是采用基于国际标准ISO/IEC 13335中的定义与方法对其进行量化,其量化值如表6所示。

表 6 资产重要性量化

Table 6 Quantification of asset importance	
资产重要性等级	量化数值
可忽略级别	1~2
低级别	3~4
中等级	5~6
高级别	7~8
严重级别	9~10

在得到漏洞固有威胁和资产重要性的量化值后,通过如下方式完成从指数要素到Ⅲ级指标漏洞状态指数的聚合过程:

$$S(t)=\lambda\frac{s(t)}{E(t)}\tag{16}$$

式中: S 为漏洞状态指数量化值; λ 为协调系数; s 为漏洞固有威胁; E 为资产重要性。

3.2.4 网络安全威胁指数

网络安全威胁指数包括木马指数、网络攻击指数、病毒指数3个Ⅲ级指标,其具体概念如下:

1) 木马指数。指在视频专网区域内,木马对视频专网中的数据的机密性造成的破坏程度。

2) 网络攻击指数。指在视频专网区域内,网络攻击对视频专网的数据的可用性造成的破坏程度。

3) 病毒指数。指在视频专网区域内,病毒对视频专网中数据的完整性和可用性造成的破坏程度。

1) 木马指数

在视频专网的网络安全态势评估过程中,针对其运行的业务和存在的安全风险,确定木马指数所包含的指数要素集合,具体包括目标资产情况、安全事件发生频率、可清除性3个指数要素,其具体概念如下:

① 目标资产情况。指安全事件对视频专网区域内的目的主机资产影响的总体情况。

② 安全事件发生频率。指视频专网区域内在一定的时间周期内发生安全事件的频率总体情况。

③ 可清除性。指视频专网区域内对存在的木马等进行清除的总体难易程度。

本文采用如下矩阵量化方式对目标资产情况、安全事件发生频率、可清除性3个指数要素进行量化,如表7所示。

表 7 木马指数的指数要素量化

Table 7 Quantification of indicator elements of Trojan index		
指数要素	选取含义	量化数值
目标资产情况	不重要/不太重要/比较重要/重要/非常重要	2/4/6/8/10
安全事件发生频率	低/中/高	5/8/10
可清除性	普通查杀工具/专用查杀工具/手工查杀	5/8/10

2) 网络攻击指数

在视频专网的网络安全态势评估过程中,针对其运行的业务和存在的安全风险,确定网络攻击指数包含的指数要素集合,具体包括目标资产情况、攻击严重程度、攻击强度、安全事件发生频率、攻击防范代价5个指数要素,其中目标资产情况和安全事件发生频率的概念在介绍木马指数时已经给出,其他指数要素的具体概念如下:

① 攻击严重程度。指当发生拒绝服务攻击后,对视频专网区域内的主机、设备等造成的总体影响程度。

② 攻击强度。指针对视频专网区域内的拒绝服务攻击发生时来自于攻击源的攻击报文包的总体数量情况。

③ 攻击防范代价。指在视频专网区域内,针对拒绝服务攻击的防范,对可采用的防护措施强度情况。

本文采用如下矩阵量化方式对攻击严重程度、攻击强度和攻击防范代价3个指数要素进行量化,如表8所示。

表 8 网络攻击指数的指数要素量化

Table 8 Quantification of indicator elements of cyber attack index		
指数要素	选取含义	量化数值
攻击严重程度	不受影响/部分服务受影响/完全拒绝服务/宕机	1/4/8/10
攻击强度	低/中/高	8/9/10
攻击防范代价	本地配置方式/部署通用设备/部署专用设备	3/8/10

3) 病毒指数

在视频专网的网络安全态势评估过程中,针对其运行的业务和存在的安全风险,确定病毒指数所包含的指数要素集合,具体包括目标资产情况、病毒破坏性、安全事件发生频率、可清除性 4 个指数要素。其中目标资产情况、安全事件发生频率和可清除性的概念在介绍木马指数时已经给出,病毒破坏性的具体概念如下:

病毒破坏性指在视频专网区域内,病毒的破坏能力对视频专网的安全运行造成的总体影响程度。

本文采用如下矩阵量化方式对病毒破坏性指数要素进行量化,如表 9 所示。

表 9 病毒指数的指数要素量化

指数要素	选取含义	量化数值
病毒破坏性	无影响/低级别影响/中级别影响/高级别影响	1/4/8/10

3.2.5 前端设备运行安全指数

前端设备运行安全指数包括服务状态安全指数和异常状态安全指数 2 个Ⅲ级指标,其具体概念如下:

1) 服务状态安全指数。对视频专网前端设备的正常安全运行进行监控,用来描述视频专网前端设备服务状态的运行安全情况。

2) 异常状态安全指数。对视频专网前端设备的异常安全运行进行监控,用来描述视频专网前端设备的脆弱性、威胁性情况。

1) 服务状态安全指数

在视频专网的网络安全态势评估过程中,针对其运行的业务和存在的安全风险^[14],确定服务状态安全指数所包含的指数要素集合,具体包括前端设备可用性、故障发生频率 2 个指数要素,服务状态安全指数的指数要素具体概念如下:

① 前端设备可用性。指视频专网的前端设备在当前时间内是否可用的情况。

② 故障发生频率。指视频专网的前端设备在一定的时间周期内发生故障的频率总体情况。

本文采用如下矩阵量化方式对服务状态安全指数指数要素进行量化,如表 10 所示。

表 10 服务状态安全指数的指数要素量化

指数要素	选取含义	量化数值
前端设备可用性	可用/不可用	1/10
故障发生频率	低/中/高	5/8/10

2) 异常状态安全指数

在视频专网的网络安全态势评估过程中,针对其运行的业务和存在的安全风险^[15],确定异常状态安全指数所包含的指数要素集合,具体包括前端设备存在漏洞状态、前端设备存在弱口令状态、前端设备存在违规接入状态 3 个指数要素,异常状态安全指数的指数要素具体概念如下:

① 前端设备存在漏洞状态。指视频专网前端设备的操作系统或应用是否存在漏洞的情况。

② 前端设备存在弱口令状态。指视频专网前端设备是否存在弱口令的情况。

③ 前端设备存在违规接入状态。指视频专网前端设备是否存在违规接入的情况。

本文采用如下矩阵量化方式对异常状态安全指数指数要素进行量化,如表 11 所示。

表 11 异常状态安全指数的指标数要素量化

指数要素	选取含义	量化数值
前端设备存在漏洞状态	不存在/存在	1/10
前端设备存在弱口令状态	不存在/存在	1/10
前端设备存在违规接入状态	不存在/存在	1/10

3.2.6 边界运行安全指数

边界运行安全指数包括公安信息网边界安全指数、互联网边界安全指数、其他专网边界安全指数 3 个Ⅲ级指标,其具体概念如下:

1) 公安信息网边界安全指数。指视频专网的主机、设备、网络和子系统与公安信息网之间的边界接入设备的总体运行安全情况。

2) 互联网边界安全指数。指视频专网的主机、设备、网络和子系统与互联网之间的边界接入设备的总体运行安全情况。

3) 其他专网边界安全指数。指视频专网的主机、设备、网络和子系统与其他专网之间的边界接入设备的总体运行安全情况。

边界运行安全指数的 3 个Ⅲ级指标在物理上反映的都是边界接入设备的运行安全状况,可以使用共性的指数要素集合从多个不同的安全维度来描述边界接入设备的总体安全情况,指数要素包括边界设备可用性、边界非法接入状态 2 个指数要素,具体概念如下:

① 边界设备可用性。指视频专网与外部网络的接入设备在当前时间内是否可用的情况。

② 边界非法接入状态。指在当前时间内视频专网与外部网络接入中是否存在违规接入的情况。

本文采用如下矩阵量化方式对边界运行安全指数指数要素进行量化,如表 12 所示。

表 12 边界运行安全指数的指数要素量化
Table 12 Quantification of indicator elements of boundary operational safety index

指数要素	选取含义	量化数值
边界设备可用性	可用/不可用	1/10
边界非法接入状态	不存在/存在	1/10

4 结 束 语

本文在传统的网络安全运行、系统脆弱性和网络安全威胁的基础上考虑了视频专网面临的前端设备安全和边界设备安全威胁,构建了一套完整的面向视频专网的网络安全态势指标体系,对指标体系中各级指标进行了详细定义,并给出了具体的量化计算方法与聚合计算方法,弥补了视频专网的网络安全态势指标体系研究的空白,对于完善和丰富面向视频专网的网络安全态势指标体系具有一定的参考和借鉴意义。

参考文献 (References)

[1] 杨健,顾涛. 公安视频专网安全管理系统的研究与设计[J]. 网络安全技术与应用,2017(3):154-155.
YANG J, GU T. Design and research of public security video private network security management system[J]. Network Security Technology & Application, 2017(3):154-155 (in Chinese).

[2] 汤光华,曾宪报. 构建指标体系的原理与方法[J]. 河北经贸大学学报,1997(4):60-62.
TANG G H, ZENG X B. Principles and methods of constructing indicator system[J]. Journal of Hebei University of Economics and Trade, 1997(4):60-62 (in Chinese).

[3] 李远远. 基于粗糙集指标体系构建及综合评价方法研究[D]. 武汉:武汉理工大学,2009.
LI Y Y. Research on construction of index system and comprehensive evaluation method based on rough set[D]. Wuhan: Wuhan University of Technology, 2009 (in Chinese).

[4] 陈秀真,郑庆华,管晓宏,等. 层次化网络安全威胁态势量化评估方法[J]. 软件学报,2006,17(4):885-897.
CHEN X Z, ZHENG Q H, GUAN X H, et al. Quantitative hierarchical threat evaluation model for network security[J]. Journal of Software, 2006, 17(4):885-897 (in Chinese).

[5] 姚婷婷,郑庆华,管晓宏,等. 一种基于主机实时流量的安全评估方法[J]. 西安交通大学学报,2006,40(4):415-419.
YAO T T, ZHENG Q H, GUAN X H, et al. Security evaluation method based on host realtime traffic of hosts[J]. Journal of Xi'an Jiaotong University, 2006, 40(4):415-419 (in Chinese).

[6] 张建锋. 网络安全态势评估若干关键技术研究[D]. 长沙:国防科学技术大学,2013.
ZHANG J F. Research on key technologies of network security assessment[D]. Changsha: National University of Defense Science and Technology, 2013 (in Chinese).

[7] 王志平. 基于指标体系的网络安全态势评估研究[D]. 长沙:国防科学技术大学,2010.
WANG Z P. Research on network security situation evaluation based on index system[D]. Changsha: National University of Defense Science and Technology, 2010 (in Chinese).

[8] 李远征. 基于指标体系的电子政务外网安全态势评估研究[D]. 长沙:国防科学技术大学,2011.
LI Y Z. Research on situation evaluation of E-government extranet based on index system[D]. Changsha: National University of Defense Science and Technology, 2011 (in Chinese).

[9] 夏冰. 网络安全法和网络安全等级保护 2.0[M]. 北京:电子工业出版社,2017.
XIA B. Cyber security law and cyber security level protection 2.0[M]. Beijing: Electronic Industry Press, 2017 (in Chinese).

[10] 李晨. 基于模糊层次分析法的舰艇编队通信网络安全管理评估指标体系[J]. 兵器装备工程学报,2017,38(11):134-137.
LI C. Fleet formation's communication network safety management evaluation targets system based on FAHP[J]. Journal of Ordnance Equipment Engineering, 2017, 38(11):134-137 (in Chinese).

[11] 费军,余丽华. 基于模糊层次分析法的计算机网络安全评价[J]. 计算机应用与软件,2011,28(10):120-123.
FEI J, YU L H. Computer network security assessment with fuzzy analytic hierarchy process[J]. Computer Applications and Software, 2011, 28(10):120-123 (in Chinese).

[12] 岳丽. 基于指标体系的网络安全态势感知技术研究[D]. 天津:天津理工大学,2016.
YUE L. Research on network security situation awareness technology based on index system[D]. Tianjin: Tianjin University of Technology, 2016 (in Chinese).

[13] ANDERSON H. Nessus, Part 3: Analysing reports[EB/OL]. (2004-02-03) [2020-02-20]. http://apachepersonal.mium.se/~janjon/oldcourse/dtab80/lab/lab2/nessus_3.pdf.

[14] 邵旭东,樊志杰. 视频安全接入的关键技术研究[J]. 信息网络安全,2012(8):191-194.
SHAO X D, FAN Z J. The research of the key technology of video security access[J]. Netinfo Security, 2012(8):191-194 (in Chinese).

[15] 杜威,邹先霞. 基于网络的城市治安视频监控系统信息安全机制设计与应用[J]. 信息网络安全,2013(10):127-129.
DU W, ZOU X X. Design and application of the city security video surveillance system's information security based on network[J]. Netinfo Security, 2013(10):127-129 (in Chinese).

作者简介:
李欣 男,博士,副教授。主要研究方向:云计算、网络安全。

Construction of network security situation indicator
system for video private network

LI Xin^{1,2,*}, DUAN Yongcheng¹, HUANG Shuhua^{1,2}, FAN Zhijie³

(1. College of Information Technology and Network Security, People’s Public Security University of China, Beijing 100038, China;

2. Key Laboratory of Security Prevention Technology and Risk Assessment of the Ministry of Public Security, Beijing 100038, China;

3. Ministry of Public Security Third Institute of Information Security Technology, Shanghai 200031, China)

Abstract: In view of the lack of a complete set of index system in the current network security situation assessment for video private networks, the article studies the construction of network security situation index system. Firstly, the paper focuses on the research and analysis of the network security risks faced by the video private network; then, combined with the basic evaluation indicators of previous network security situation assessment technology, a network security situation indicator system suitable for video private network is constructed according to the characteristics of video private network; finally, the calculation method of indicator values is given, which lays a foundation for network security situation assessment of video private network. It has certain reference significance for perfecting and enriching the network security situation indicator system for video private network.

Keywords: indicator system; video private network; situational element; situational awareness; situation assessment

Received: 2020-02-21; **Accepted:** 2020-03-20; **Published online:** 2020-04-26 08:58

URL: kns.cnki.net/kcms/detail/11.2625.V.20200424.2052.001.html

Foundation items: Special Project of the Ministry of Public Security to Strengthen the Basic Work of Science and Technology Police (2017GABJC38); Start-up Fund for Research of New Teachers of the People’s Public Security University of China (2018JKF609)

* **Corresponding author.** E-mail: ndlixin@sina.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0042

面向个体人员特征的跨模态目标跟踪算法



周千里^{1,2}, 张文靖¹, 赵路平¹, 田乃倩¹, 王蓉^{1,3,*}

(1. 中国人民公安大学 警务信息工程与网络安全学院, 北京 100038; 2. 北京市公安局, 北京 100740;
3. 安全防范技术与风险评估公安部重点实验室, 北京 100038)

摘 要: 针对类内干扰影响基于个体人员特征目标跟踪算法的精确性和鲁棒性问题,分析当前跟踪算法在个体人员跟踪方面存在的不足,提出了利用语言先验知识引导辅助跟踪器的方法。在视觉跟踪器的基础上增加语言引导分支,对跟踪目标产生注意力,从而减少对类内干扰的影响。利用位置置信度进行回归目标框定位的方法解决基于孪生网络目标跟踪算法中利用分类置信度定位候选目标框的局限性,实现跨模态信息融合提升特定目标跟踪的精度。为提升所提模型对特定人员目标跟踪的针对性,构建了跨模态的人员目标跟踪数据集用于训练和验证。实验表明:所提模型应用于个体人员跟踪时表现更佳,其有效性得到了证明。

关 键 词: 目标跟踪; 孪生网络; 跨模态; 数据集; 语言先验

中图分类号: TP183; TP389.1

文献标志码: A **文章编号:** 1001-5965(2020)09-1635-08

个体人员的跨模态目标跟踪在视频智能分析、无人驾驶、自主系统导航等各个领域都有着广泛应用,特别是在公安视频监控信息深度挖掘分析方面。传统目标跟踪对象非常多样,包括动物、人、物品、车辆、飞机等一切可以移动的对象,而在实际应用或某一领域内,重点关注的往往是某一类别的对象,如交通领域关注的主要是机动车、物流领域关注的是物品等,针对特定类别对象目标跟踪开展研究能够为具体应用提供更具针对性的模型和方法,有很强的实用价值。

本文将跟踪目标聚焦在个体人员上,主要有如下考虑:①在复杂场景下对个体人员开展跟踪具有较强的技术挑战性,特别是在自然环境下多个人员同时出现时,类内干扰是制约跟踪性能的主要问题,主要原因在于当前跟踪器主干网络的预训练基于图片分类任务,对于对象类别不敏感,单纯依靠视觉特征进行区分难度较大,如何利用

多维度信息解决这类问题具有学术研究价值。②个体人员的跟踪问题在多个领域有着广泛的应用,如公共安全智能视频监控、无人驾驶对道路行人检测、机器人导航避障、个体生物特征步态识别等,都需要对人员目标进行跟踪,而目前针对具体对象的跟踪算法研究比较缺乏,对这些问题开展研究具有很强的实用价值和应用前景。③将语言先验知识与视觉任务进行结合的研究方法具有很强的研究意义。单纯以数据驱动的深度学习模型在可解释性、鲁棒性、稳定性上都存在一定的局限性,实际应用中用户都有任务意图,如何将这种意图反映到算法中实现真正的任务驱动具有非常重要的现实意义,而利用先验知识协同视觉模型完成跟踪任务实现了知识驱动与数据驱动的结合,将能够在多个方面提升算法能力,且能够应用在多种视觉任务上。

综上,本文提出了跨模态的个体人员目标跟踪

收稿日期: 2020-02-21; 录用日期: 2020-03-15; 网络出版时间: 2020-03-27 14:33
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200326.2132.004.html
基金项目: 国家重点研发计划(A19808); 中国人民公安大学基本科研业务费年度重大项目(2019JKF111)
* 通信作者. E-mail: dbdxwangrong@163.com

引用格式: 周千里, 张文靖, 赵路平, 等. 面向个体人员特征的跨模态目标跟踪算法[J]. 北京航空航天大学学报, 2020, 46(9): 1635-1642. ZHOU Q L, ZHANG W J, ZHAO L P, et al. Cross-modal object tracking algorithm based on pedestrian attribute[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1635-1642 (in Chinese).

框架,该框架结合了自然语言特征和视觉特征,增强了个体人员跟踪的鲁棒性,有效降低了类内干扰给跟踪性能带来的影响。为保证训练和检验本文算法的有效性,构建了基于个体人员特征的专用目标跟踪数据集,并进行了仿真验证,本文算法在专用数据集上的表现也超过了现有主流算法模型。

1 关键技术

1.1 视觉单目标跟踪

目前,从目标跟踪算法上来说,单目标跟踪算法可以分为基于检测和基于模板匹配两大类。基于检测的目标跟踪算法以相关滤波跟踪为主,将待搜索图片输入相关滤波器,计算与目标区域的相关响应,根据相关性计算得分来区分前景和背景,此类算法在跟踪速度和准确性上都取得了很好的效果。随着深度神经网络的不断演进发展,基于模板匹配的目标跟踪算法开始引入相关模块,孪生匹配网络首先被应用于目标跟踪任务中^[1],在此基础上利用区域推荐网络^[2]、注意力机制^[3]、更深更宽的主干网络^[4-5]、数据增强训练^[6]、强化学习^[7]、模板更新^[8]等方法,能够进一步提升性能。但是目前性能最好的跟踪器,在对外表特征相似的同类对象区分上也存在不足,这种情况在针对个体人员的跟踪上尤为突出。

1.2 视觉引用表达

近年来,将自然语言处理应用于计算机视觉中的研究得到了广泛关注^[9]。本文重点关注的视觉引用表达(Referring Expression 或 Visual Grounding),其是指在接收一张图片和一个语言指令或描述之后,定位出图片当中与语言描述相关的目标。文献[10]利用卷积神经网络(CNN)和长短时记忆单元网络(LSTM)分别从图片和语言描述中抽取特征拼接后训练,完成图像的分割任务,如图1所示,左图为原图,中图为分割标准,右图为预测结果,语言描述为:man in blue shirt in the right side。文献[11]提出了利用语言进行目标跟踪的3种模型。MAttNet^[12]将语言描述划分为主语、位置 and 关系3个部分,并用不同模块与图像进行联合处理训练获得图像分割。近期的一些研究着眼于将最新的BERT架构引入视觉与自然



图1 视觉引用表达图像分割示意

Fig.1 Image segmentation of visual referring expression

语言融合处理的领域,在视觉应用表达上也取得了不错的效果^[13-14]。

1.3 目标跟踪数据集

数据集是用来对模型算法进行训练、评估和测试的基础工具,目前目标跟踪领域的数据集非常丰富。OTB(Object Tracking Benchmark)^[15-16]在2013年和2015年分别推出了包含50个和100个视频的数据和标注;2013年推出的VOT(Visual Object Tracking)包含16个短视频序列,而后每年进行更新;2017年发布的NFS(Need for Speed)^[17]包含100个高帧率视频;TrackingNet^[18]包含30 132个视频;GOT-10k^[19]包含10 000段视频;LaSOT^[20]包含了70类1 400段高密度标注的视频。但是还没有专门针对个体人员跟踪的数据集,要想训练出适应于复杂场景下跨模态的跟踪模型,需要有定制的专用数据集。

2 跨模态目标跟踪模型

本文提出利用语言先验知识对个体人员目标跟踪进行监督和引导,充分利用非视觉信息,辅助视频信息的分析处理,提高特定人员目标跟踪的精确性,提升目标的发现和追踪能力,为视频数据与其他异构数据和知识进行关联融合提供一种新的思路和方法。

2.1 整体框架

该框架主要分为3个模块(见图2):基于Siamese孪生架构的视觉跟踪模块、基于递归神经网络(RNN)的语言引导模块、基于IoU交并比优化的融合判别模块。模型从数据集中第1帧图像的标注(Ground Truth)获得跟踪目标特征,与要跟踪的后续视频帧一起作为输入送入视觉跟踪分支进行视觉判别;语言引导分支将语言描述和被搜索视频帧作为输入,通过语言模型处理在图片中描述确定要跟踪目标的位置和范围,形成注意力蒙版;将视觉分支和语言分支输出的结果分别送入融合判别模块,根据视觉跟踪模块输出的分类置信度与基于交并比IoU的位置置信度选出最佳目标框,再通过位置框优化,最终输出结果。

2.2 基于Siamese孪生架构的视觉跟踪模块

视觉跟踪模块基于SiamMask模型^[21],将ResNet-50^[22]作为视觉特征抽取主干网络,输出256维的目标跟踪对象和搜索帧的特征图;再对特征图进行深度交叉相关运算,得到多通道响应图,该响应图是对目标跟踪对象和搜索帧视觉相似度的编码;基于该编码通过区域推荐网络(RPN)^[2]

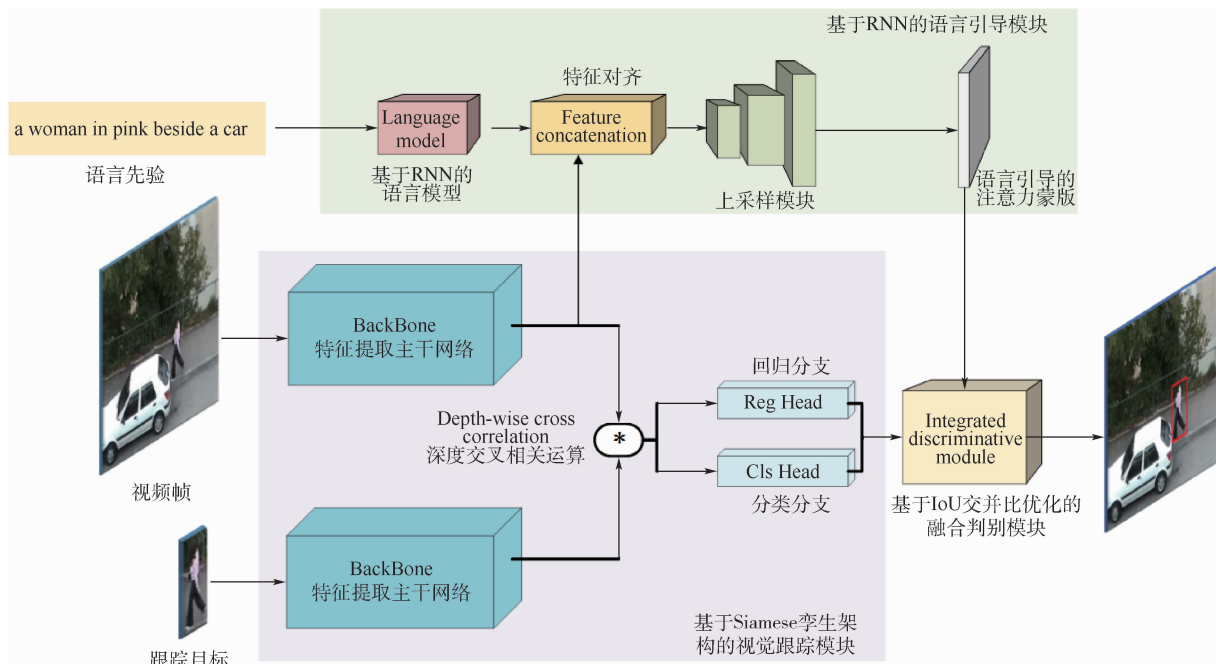


图 2 跨模态目标跟踪整体框架

Fig. 2 Cross-modal object tracking framework

生成目标候选区域(分类置信度),并通过回归优化更准确地定位目标区域(位置置信度)。SiamMask 采用最高分类得分值选出最终目标框,本文将采用位置置信度和分类置信度相结合的方式获得更加准确的结果。

2.3 基于 RNN 的语言引导模块

语言引导模块借鉴 DMN^[23] 的网络结构,分别使用 CNN 和 RNN 来提取每帧图片和描述语言的特征。利用 LSTM 的升级版 SRU 模块处理词嵌入 e_i 后的句子隐含状态 h_i ,再将其与词嵌入进行拼接得到状态语言特征 r_i 。

$$r_i = [e_i, h_i] \quad (1)$$

借鉴文献[11]的思想,利用动态滤波器处理 r_i 后得到 $f_{k,i}$,与视觉特征 I_N 进行卷积获得 F_i ,将语言特征、视觉特征和位置特征进行拼接后用 1×1 卷积得到特征响应图 M_i 。

$$M_i = \text{Conv}_{1 \times 1}([I_N, F_i, r_i]) \quad (2)$$

用双插值进行上采样后,获得与原始图片尺寸一致的蒙版响应图,用最小封闭矩形框(MER)计算得到基于位置的目标框。为了确保该模块对人员特征的提取有效性,利用专用数据集进行了针对性训练,使该模型对类内干扰、人员移动、遮挡、模糊等问题具有更强的判定识别能力。

2.4 基于 IoU 交并比优化的融合判别模块

在获得视觉特征推荐和语言引导的定位后,该融合判别模块负责利用位置置信度和分类置信度获得最佳目标框。首先,将视觉跟踪模块推荐

分类得分最高的 50 个候选框提出,再计算它们与语言引导模块所得目标框的重叠率 IoU,根据 IoU 得分的高低来确定最佳目标框。受 IoU guided NMS^[24] 启发,临近的候选框对于精确定位仍有价值,因此基于位置得分最高的候选框,找出分数排名前 20 的其他候选框,用 IoU 值作为权重,计算得到需要微调的位置坐标。然后,对位置得分最高的候选框进行微调处理,获得最终回归的目标框。融合判别模块有 2 个功能:①基于位置置信度和分类置信度选择最佳的候选框;②基于相邻众多候选框,对最佳候选框进行微调优化,获得最终目标框结果,如图 3 所示,粉色框为 SiamMask 分类最高分的框,蓝色框为语言引导模块生成的框,绿色框是利用位置置信度和分类置信度获得的结果。语言描述为:man walking along the street。

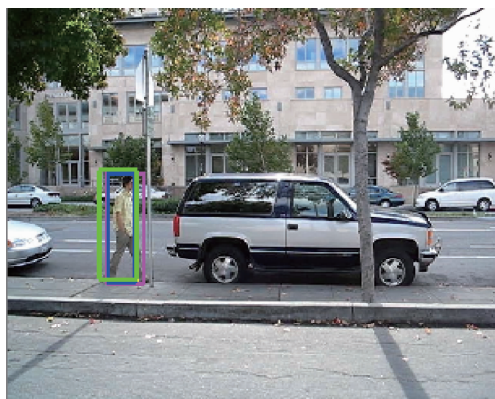


图 3 多模块回归预测结果

Fig. 3 Results of multiple modules predicted regression

3 专用数据集构建及技术实现

3.1 专用数据集构建

本文旨在研究个体人员特征的目标跟踪任务,现有数据集无法满足需求,需要构建针对个体人员专用数据集进行训练。①目前通用数据集中没有专门针对人员个体跟踪的数据集,通用数据集都有部分个体的视频数据片段和标注,需要将这类数据进行汇集整合,供模型训练、评估和测试用;②目前的单目标数据集中除 LaSOT 和 lingual OTB99^[11]外,都没有提供自然语言的描述标注,因此在引入语言引导模块时,没有语言描述数据可用,需要构建专用数据集支持这类模型。基于此,本文构建了 PerTrack 专用数据集,从多个数据集中收集了以人员为跟踪目标的视频数据,按照固定格式对数据标注和语言描述进行整理,对没有语言描述的数据进行补充。从 GOT-10k 获取了 58 个视频,从 LaSOT 获取了 20 个视频,从 OTB 获取了 25 个视频,视频片段都与人员相关,类别包含跳舞、短跑、路人、歌手、滑冰、马拉松等与人相关的场景。另外,按照 7:1:2 的比例,对人员目标跟踪数据进行训练集(Train)、评估集(Train-val)和测试集(Test)的分类。每个具体的视频文件夹下,除了每个视频帧的图片外,提供了标注 Ground Truth(GT);还提供了自然语言的描述文本和根据 Ground Truth 生成的双值目标分割蒙版,用于为 DMN 模型提供训练。经过前期工作,目前从不同渠道收集到针对人员的视频数据 103 段,共 90 926 帧;73 段用于训练,共 63 818 帧;10 段用于评估,共 8 292 帧;20 段用于测试,共 18 816 帧。

3.2 技术实现

为使视觉跟踪模块和语言引导模块分支保持非相关性,进行了独立的训练。前者采用 Siam-Mask 的预训练模型,后者使用专门数据集进行训练和优化,最后在决策层进行融合。

在数据预处理阶段,将专用数据集中的标注 Ground Truth 转化为二值蒙版供训练使用,由于跟踪数据集标注的 Ground Truth 提供了左上角坐标 c_x 、 c_y 和目标框宽度 w 、高度 h ,需要将其转化为二值蒙版供语言引导模块使用。将目标框内的值设为 255,目标框外的值设为 0,就形成训练数据。

在语言引导模块的参数优化和微调阶段,将 DMN 模型的预训练权重作为参考。DMN 的预训练模型参数是基于 ReferIt^[25]进行训练的,ReferIt 针对具有 19 894 张图片和 130 525 个语言描述开展训练,但适应跟踪的模型参数有所不同。针对

人员特征的目标跟踪任务,需要用到本文构建的训练数据集进行训练。因此在本文实验中,分别将模型在低分辨率和高分辨率下训练 7 个和 5 个 epoch,获得最终的模型参数。

在测试阶段,将 DMN 网络输出的蒙版利用最小封闭矩形框进行变换后输出结果,如图 4 所示,右图为通过语言引导模块输出的结果,通过最小封闭矩形框获得目标框,左图为在视频某帧上的对应展示,语言描述为:girl in yellow shirt with purple pants,与视觉跟踪模块输出进行融合。



图 4 最小封闭矩形框生成图

Fig. 4 Illustration of minimum enclosing rectangle

4 实验与结果分析

利用自建的数据集对模型进行训练,得到实验需要用到的参数。主要完成 3 组实验:①对语言引导模型在执行跟踪数据训练前后的效果进行对比评估;②选用部分主流跟踪算法在自建数据集上测试评估,与本文算法的性能作对比;③在不给出第 1 帧跟踪目标的情况下,直接用语言引导模型检测被跟踪对象测试结果的评估。

4.1 语言引导模块评估

所有参数训练都是基于 DMN 在 ReferIt 数据集预训练的基础上开展的,针对不同的训练数据生成不同参数,对比测试评估不同数据集训练后的模型参数对人员个体目标检测任务的效果。测试效果用平均交并比 mIoU 来评估,其是衡量图像分割精度的重要指标。计算平均真实值和预测值的交集和并集之比,即在每个类别上计算 IoU 值,再在所有类别上求平均即得 mIoU。表 1 中,参数 0 表示只在 ReferIt 数据集训练的结果,参数 1 表示在参数 0 的基础上,在 lingual OTB99 上做微调的结果,参数 2 表示在参数 0 的基础上,在 lingual OTB99 和 LaSOT 上做微调的结果,优化参

表 1 语言引导模块评估结果对比

Table 1 Comparison results of language guided module

模型参数类型	平均交并比
参数 0	0.241 3
参数 1	0.359 8
参数 2	0.349 8
优化参数	0.465 0

数表示在参数 0 的基础上,在专用数据集 Per-Track 的训练集上的结果。

从表 1 分析对比看出,不同数据集对语言引导模块效果有着直接影响。lingual OTB 和 LaSOT 都是不分类别的目标跟踪数据集,而本文采用的是针对人员类别的数据集,因此在针对人员跟踪检测的评估集上的测试结果上来看,效果有了大幅度提升。

另外,利用语言引导模型还对 lingual OTB99 的评测结果进行了对比实验,目的是评估针对特定跟踪目标类别训练后的参数对通用数据集的影响。

从图 5 结果进行分析,对比的跟踪器为主流孪生网络跟踪器,包括 SiamMask^[21]、SiamRPN^[2]和 SiamFC^[1],其中 LangTrack 为用通用数据集训练的模型参数,取得了最优效果,PerTrack 为用行人专用数据集训练的结果,效果比利用纯视觉信息的主流跟踪器的效果差。结果说明,在通用目标跟踪任务中采用专用数据集训练,会使模型具有对专用目标的偏向性,导致整体效果欠佳,说明模型是具有类别敏感性的。

的单目标跟踪算法进行了测试,结果如表 2 和图 6 所示。相比于原有纯视觉模型,增加语言引导模块有效提升了算法的准确性和鲁棒性,同时也超越了现有主流跟踪算法的性能,但跟踪速度只能达到 4 帧/s,低于现有主流跟踪算法,原因在于语言引导模型的运行影响了跟踪处理速度。算法性能的提升主要得益于专用数据集的训练和语言先验知识的监督。本文采用一遍成功率(One-Pass Evaluation,OPE)作为评估指标,即运行一次跟踪算法,获取每一帧跟踪目标的位置和大小,由平均精度和成功率对其进行评分。

本文还对不同主流跟踪器效果进行了可视化。如图 7 所示,其中红色框为 SiamRPN++,绿色

表 2 本文模型与主流跟踪算法评估结果对比

Table 2 Comparative results between proposed model and mainstream tracking algorithms

算法	平均精度	成功率
SiamRPN ^[2]	0.493	0.566
SiamRPN++ ^[5]	0.508	0.612
SiamMask ^[21]	0.708	0.808
ECO ^[26]	0.647	0.797
ATOM ^[27]	0.732	0.848
DIMP ^[28]	0.787	0.841
本文模型	0.930	0.978

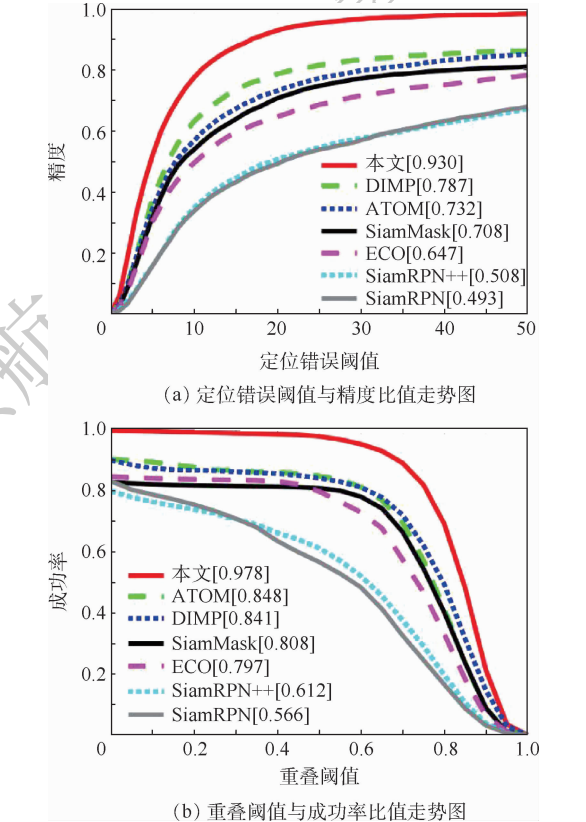


图 6 本文模型与主流跟踪算法的 OPE 评估结果
Fig. 6 OPE evaluation results between proposed model and mainstream tracking algorithms

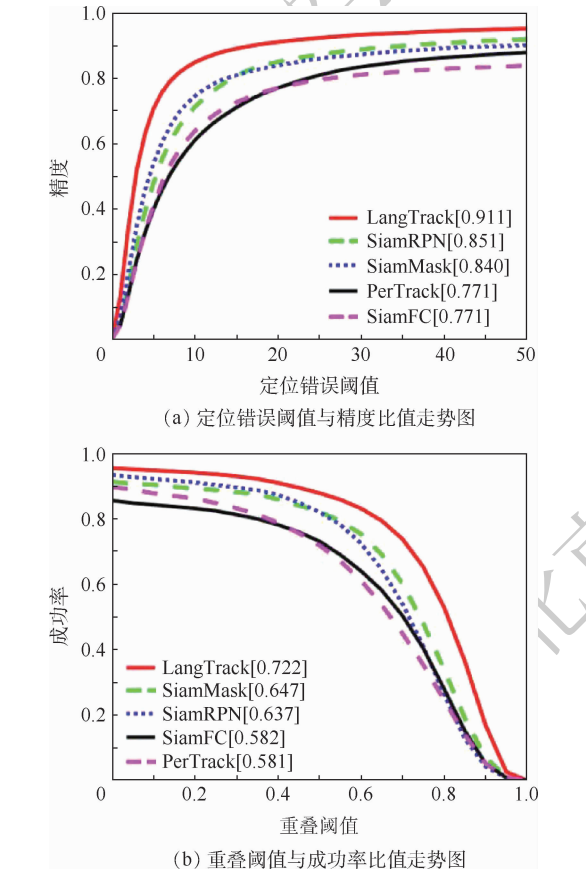


图 5 主流跟踪器的结果比较
Fig. 5 Comparative results among mainstream trackers

4.2 跟踪任务评估

为了将本文设计的模型与主流跟踪算法在行人跟踪专用数据集上的表现做比较,对当前最新

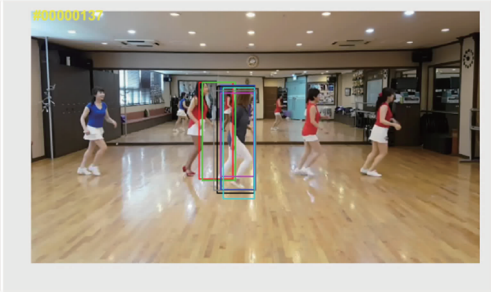


图7 不同跟踪器效果可视化

Fig. 7 Results visualization of different trackers

色框为 SiamRPN, 蓝色框为 ATOM, 黑色框为 DIMP, 粉色框为 ECO, 青色框为 SiamMask, 灰色框为本文模型。

4.3 语言检测跟踪评估

传统跟踪任务是给定第 1 帧对象后开展视觉跟踪,但在实际公安视频应用中,大部分场景能够获取对目标对象的语言描述,因此,利用模型完成先检测目标后跟踪的任务更具有实用价值。如图 8 所示,pertrack_DMN 表示原模型的结果,pertrack_GTfree 表示没有给定目标框,仅靠语言检测后进行跟踪的结果,pertrack_SiamMask 表示 SiamMask 跟踪模型的结果,数据集用 PerTrack 的测试集。实验表明,在不给定第 1 帧目标框情况下,仅

用语言先验检测出目标再进行跟踪的结果也优于纯视觉跟踪,但由于没有视觉信息的辅助,整体性能略低于原模型。

4.4 结果分析

通过上述 3 组实验,得出如下结论:

- 1) 将语言先验引入面向人员特征的目标跟踪任务能有效提升跟踪精度和鲁棒性,特别是在多人干扰的场景下,语言描述能够很好地定位被跟踪人员位置,提升算法跟踪抗类内干扰的能力。
- 2) 语言引导模型具有目标类别敏感性,用于针对行人跟踪的模型不能用于通用目标跟踪,否则对纯视觉模型造成干扰,因此,在应用时需要针对某类对象进行针对性训练和微调。
- 3) 语言描述不适用所有跟踪场景。不是所有场景都能够用语言来描述目标对象,而且某一帧适用的描述随着场景变化,在后面的帧中不一定适用,因此,部分复杂场景应用中可以考虑用语言进行目标检测后,用纯视觉进行跟踪。

5 结束语

针对个体人员特征目标跟踪中存在的类内干扰问题,提出了一种引入语言先验知识引导的人员目标跟踪算法,并设计了由视觉跟踪模块、语言引导模块和融合判别模块组成的跨模态目标跟踪框架,同时为保证训练和测试模型的有效性,构建了专用的跨模态个人目标跟踪数据集。实验结果表明本文提出的模型与现有主流跟踪模型相比具有更好的精度和鲁棒性。

本文将数据驱动的视觉深度学习模型与语言先验知识相结合,为公安领域视频监控信息的深度挖掘和关联等提供了一种新的思路,下一步的研究可以将语言先验知识的引导拓展到多目标跟踪、行人再识别、图像分类等其他视觉任务中。另外,利用本文提供的专用数据集,可以提升跨模态人员跟踪的精度,如果能够进一步丰富该数据集的内容,将更多真实公安场景下的数据和标注补充到数据集中,将能够训练出更多适应不同场景可复制、可重用的算法模型,或者为不同机构提供的算法进行测评,这些工作将有利于进一步提升公安视频监控的应用效能。

参考文献 (References)

[1] BERTINETTO L, VALMADRE J, HENRIQUE J F, et al. Fully-convolutional siamese networks for object tracking[C] // European Conference on Computer Vision. Berlin: Springer, 2016: 850-865.

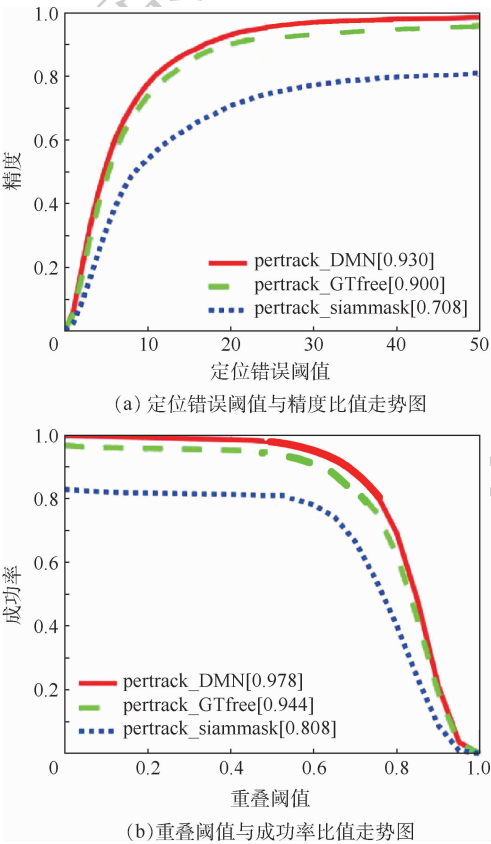


图8 语言检测跟踪评估

Fig. 8 Tracking assessment of language detection

- [2] LI B, YAN J, WU W, et al. High performance visual tracking with siamese region proposal network[C] // Proceedings of the IEEE Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8971-8980.
- [3] KOSIOREK A R, BEWLEY A, POSNER I, et al. Hierarchical attentive recurrent tracking[C] // Neural Information Processing Systems, 2017, 36: 3053-3061.
- [4] ZHANG Z, PENG H. Deeper and wider siamese networks for real-time visual tracking[C] // Proceedings of the IEEE Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4591-4600.
- [5] LI B, WU W, WANG Q, et al. Evolution of siamese visual tracking with very deep networks[J]. IEEE Computer Vision and Pattern Recognition, 2019, 35(9): 4282-4291.
- [6] ZHU Z, WANG Q, LI B, et al. Distractor-aware siamese networks for visual object tracking[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 103-119.
- [7] REN L, YUAN X, LU J, et al. Deep reinforcement learning with iterative shift for visual tracking[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 684-700.
- [8] ZHANG L, GONZALEZGARCIA A, DE WEIJER J V, et al. Learning the model update for siamese trackers[C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 4010-4019.
- [9] MOGADALA A, KALIMUTHU M, KLAKOWI D, et al. Trends in integration of vision and language research: A survey of tasks, datasets, and methods[J]. IEEE Computer Vision and Pattern Recognition, 2019, 30(19): 1183-1986.
- [10] HU R, ROHRBACH M, DARRELL T, et al. Segmentation from natural language expressions[C] // European Conference on Computer Vision. Berlin: Springer, 2016: 108-124.
- [11] LI Z, TAO R, GAVVES E, et al. Tracking by natural language specification[C] // Proceedings of the IEEE Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 7350-7358.
- [12] YU L, LIN Z, SHEN X, et al. Modular attention network for referring expression comprehension[C] // Proceedings of the IEEE Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 1307-1315.
- [13] SUN C, MYERS A, VONDRICK C, et al. A joint model for video and language representation learning[C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 7464-7473.
- [14] SU W, ZHU X, CAO Y, et al. Pre-training of generic visual-linguistic representations[C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 13-23.
- [15] WU Y, LIM J, YANG M, et al. Online object tracking: A benchmark[C] // Proceedings of the IEEE Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2013: 2411-2418.
- [16] WU Y, LIM J, YANG M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848.
- [17] GALOOGAHI H K, FAGG A, HUANG C, et al. A benchmark for higher frame rate object tracking[C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 1134-1143.
- [18] MULLER M, BIBI A, GIANCOLA S, et al. A large-scale dataset and benchmark for object tracking in the wild[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 310-327.
- [19] HUANG L, ZHAO X, HUANG K, et al. A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 45(21): 1374-1391.
- [20] FAN H, LIN L, YANG F, et al. A high-quality benchmark for large-scale single object tracking[C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2018: 5374-5383.
- [21] WANG Q, ZHANG L, BERTINETTO L, et al. Fast online object tracking and segmentation: A unifying approach[C] // Proceedings of the IEEE Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 1328-1338.
- [22] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C] // Proceedings of the IEEE Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
- [23] MARGFOYTUAY E A, PEREZ J C, BOTERO E, et al. Dynamic multimodal instance segmentation guided by natural language queries[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 656-672.
- [24] JIANG B, LUO R, MAO J, et al. Acquisition of localization confidence for accurate object detection[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 816-832.
- [25] KAZEMZADE S, ORDONEZ V, MATTENV M, et al. Referring to objects in photographs of natural scene[C] // Empirical Methods in Natural Language Processing, 2014, 28: 787-789.
- [26] DANELLJIAN M, BHAT G, KHAN F S, et al. Efficient convolution operators for tracking[C] // Proceedings of the IEEE Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6931-6939.
- [27] DANELLJIAN M, BHAT G, KHAN F S, et al. Accurate tracking by overlap maximization[C] // Proceedings of the IEEE Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 4660-4669.
- [28] BHAT G, DANELLJIAN M, VAN GOOL L, et al. Learning discriminative model prediction for tracking[C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 6182-6191.

作者简介:

周千里 男, 博士研究生。主要研究方向: 计算机视觉。

张文靖 男, 博士研究生。主要研究方向: 计算机视觉。

赵路平 女, 硕士研究生。主要研究方向: 单目标跟踪。

田乃倩 女, 硕士研究生。主要研究方向: 单目标跟踪。

王蓉 女, 博士, 教授, 博士生导师。主要研究方向: 计算机视觉。

Cross-modal object tracking algorithm based on pedestrian attribute

ZHOU Qianli^{1,2}, ZHANG Wenjing¹, ZHAO Luping¹, TIAN Naiqian¹, WANG Rong^{1,3,*}

- (1. Police Information Engineering and Network Security College, People's Public Security University of China, Beijing 100038, China;
2. Beijing Municipal Public Security Bureau, Beijing 100740, China;
3. Key Laboratory of Security Technology and Risk Assessment, Beijing 100038, China)

Abstract: The accuracy and robustness of the object tracking algorithm have been influenced by the intra-class interference when tracking pedestrian. In this paper, we analyze the drawbacks of current tracking algorithms and propose a model to combine the visual feature and language priori to improve the performance of the tracker. The language guided branch is added to supervise the visual tracking branch by generating the attention, so the intra-class interference can be alleviated. We also propose a method to improve the accuracy of the cross-modal object tracking based on the location confidence instead of classification confidence for siamese trackers. To validate our method, we customize the dataset specialized for pedestrian tracking. The experiment shows the effectiveness of this model.

Keywords: target tracking; siamese network; cross-modal; dataset; language priori

Received: 2020-02-21; Accepted: 2020-03-15; Published online: 2020-03-27 14:33

URL: kns.cnki.net/kcms/detail/11.2625.V.20200326.2132.004.html

Foundation items: National Key R & D Program of China (A19808); The Operating Expenses of Basic Scientific Research Project of the People's Public Security University of China (2019JKF111)

* **Corresponding author.** E-mail: dbdxwangrong@163.com

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0079

基于道路特征信息的车道结构化解析

罗胜, 赵丽*, 王慕抽

(温州大学 计算机与人工智能学院, 温州 325000)



摘 要: 高速道路动态执法要求车道检测算法能够结构化解析道路,但是基于传统手工设计特征的车道检测算法准确率和召回率不足,而基于深度学习的算法又对计算资源要求太高,因此提出基于道路特征信息的车道结构化解析算法。利用边缘点的梯度统计信息筛选 Hough 空间的候选点,用动态规划的方法在剩余的 Hough 空间候选点中寻找最合理的车道线组合,能够在较少计算资源的平台上准确地检测到道路上的全部车道。在自有数据的检测实验中,所提算法能够准确定位结构化和非结构化道路;在对比实验中,所提算法在准确率、召回率和计算速度上均比同类算法有所提高。

关 键 词: 车道检测;道路解析;动态执法;Hough 变换;动态规划

中图分类号: TP391.4

文献标志码: A

文章编号: 1001-5965(2020)09-1643-07

目前,道路执法的视觉监控系统中,从静态摄像头中抽取语义信息的自动化程度较高^[1-6],能自动识别并记录违规违法行为;但是架设在执法车上的动态摄像头,其信息处理和自动化程度较低,无法自动识别违规违法行为,甚至无法准确解析道路结构。解析道路结构是自动化动态执法的基础和关键。尤其在高速道路上,识别出快车道和慢车道、路肩,随机动态执法,可以提高高速道路的通行效率,消除安全隐患。而且,道路结构化解析也是无人驾驶、辅助行驶、自主导航等智能系统至关重要的基础技术。

近几十年来,各国研究者提出了自顶向下^[7]、自底向上^[6,8-13]的人工设计算法,抽取道路特征解析道路,但是难于普适多种道路,而且往往侧重于当前车辆两侧的两条车道线。其中,最为经典的方法是基于 Hough 变换检测车道^[8],但其计算速度低。因此,Satzoda 等^[9]提出利用多层次 Hough 变换来提高计算速度,Katru^[10]采用分支、合并的并行方法提高计算速度。另外,Wang

等^[11]提出基于 B-snake 的车道模型以描述更广泛的车道结构,Jung 和 Kelber^[12]提出近场采用线性函数、远场采用二次函数的抛物线车道模型,Bertozzi 和 Broggi^[13]采用附加非线性项的扩展双曲线处理直线和曲线之间的过渡。近十年来,深度学习迅猛发展,迅速出现了 LaneNet^[14]、3D-LaneNet^[15]、LaneNet + H-Net^[16]、LineNet^[17]、SC-NN^[18]、UPSNet^[19] 等准确自动化解析道路的网络,能够结构化解析道路上的全部车道,但是计算设备复杂、昂贵,并不适合架设在车上。

传统的车道检测算法把图像空间的边缘线经 Hough 变换到 $\rho-\theta$ 空间,根据车道线总是长线的特点定位到累加值较大的候选点,再回到图像空间筛选候选线段。这种思路具有以下缺点:①需要把图像空间边缘上的所有点都投影到离散的 $\rho-\theta$ 空间,然后对 $\rho-\theta$ 空间的每个点进行筛选,并对每条候选线段再次筛选,流程长,算法复杂,时间和空间复杂度都很高;②得到的 $\rho-\theta$ 空间候选点并不与道路线一一对应,而是零散、重复甚至包含

收稿日期: 2020-03-03; 录用日期: 2020-03-20; 网络出版时间: 2020-03-27 13:27

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200326.2017.003.html

基金项目: 浙江省自然科学基金(LQ19F020005); 浙江省公益项目(2016C31117); 温州市科技计划(2018ZG021)

* 通信作者。E-mail: lizhao@wzu.edu.cn

引用格式: 罗胜, 赵丽, 王慕抽. 基于道路特征信息的车道结构化解析[J]. 北京航空航天大学学报, 2020, 46(9): 1643-1649.
LUO S, ZHAO L, WANG M C. Lane semantic analysis based on road feature information[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1643-1649 (in Chinese).

不相关的线段,导致图像空间筛选候选线段的处理比较复杂,检测结果不准确。因此,Ding 等^[20]采用增强 RANSAC 进行线段筛选。事实上,图像空间边缘点上就有梯度信息,对此梯度计算直方图,长线总是峰值,而且此直方图比 Hough 变换后的 ρ - θ 空间更直观。但是,此直方图综合了整个图像各个位置相同角度线段的梯度,并不能独立筛选出车道线。而 Hough 变换后的 ρ - θ 空间中有距离 ρ 变量,因此能够将各个直线区分开来。如果结合这 2 种信息,就可以简化 ρ - θ 空间和直线的筛选步骤,简化算法,提升检测效果。

基于以上思路,本文提出了梯度直方图和 Hough 变换互补的算法,具有以下优点:①用边缘点的梯度直方图筛选 ρ - θ 空间的候选点;②用动态规划的方法在剩余的 ρ - θ 空间候选点中寻找最合理的车道线组合。这种算法能够大幅降低时间和空间复杂度,提高检测效率,并且能够定位道路上的所有车道,为动态执法提供结构化的道路信息。

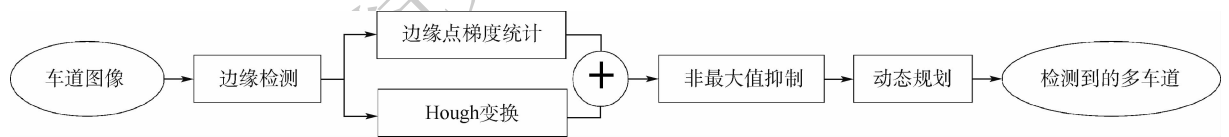


图 1 本文算法流程
Fig. 1 Flowchart of proposed algorithm

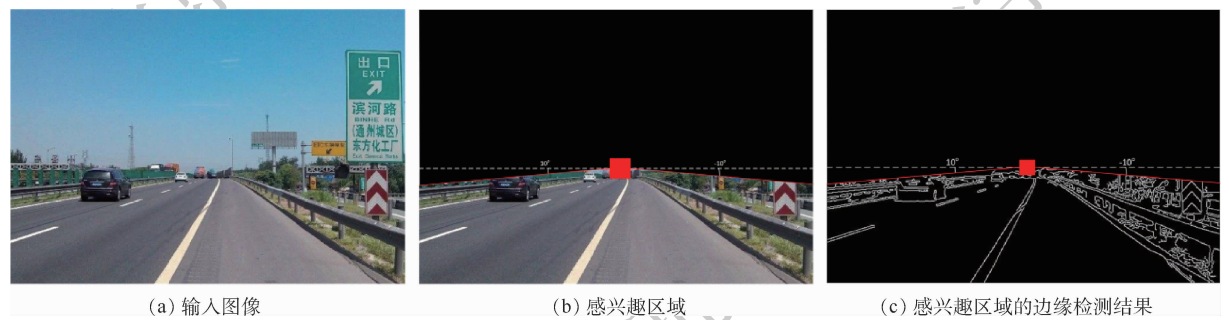


图 2 输入图像及感兴趣区域部分的边缘
Fig. 2 Original image and edges of ROI

1.2 边缘点梯度统计直方图

对感兴趣区域中的所有点,在原图像的灰度图中计算梯度。为提高梯度的准确性,扩大了邻域范围,采用 5×5 的滤波器,权重如图 3 所示。统计每个梯度角度上的边缘点数目,得到边缘点梯度统计直方图,如图 4 所示。因为是对整个感兴趣区域进行统计,同一倾角但不同线段上的边缘点会被合计成一个值。

1.3 Hough 变换

Hough 变换将图像空间的参数离散化成累加器^[2]。Hough 变换如下:

$$\rho = x \cos \theta + y \sin \theta \tag{1}$$

1 车道结构化解析算法

本文算法流程如图 1 所示,分为边缘检测、边缘点梯度统计、Hough 变换、非最大值抑制和动态规划等步骤,最后输出检测到的多车道数据。

1.1 边缘检测

采用常规的边缘检测方法,如 canny 算子,对输入的图像抽取边缘。为提高计算速度,采取了以下 2 个措施:①由于定位车道计算出车道线在图像中的相对位置,把分辨率较高的输入车道图像降采样为只有 960×800 的低分辨率图像;②由于车道主要位于图像的下半部分,并且车道线的消失点通常位于图像中间的一个矩形框内,假设矩形框高度和宽度为整个图像的 $1/10$,同时车道线的角度在 $-80^\circ \sim 80^\circ$ 范围内,将图像的下半部分、 $-80^\circ \sim 80^\circ$ 的范围设置为感兴趣区域。预处理效果如图 2 所示。

-2	-4	-4	-4	-2
-4	0	8	0	-4
-4	8	24	8	-4
-4	0	8	0	-4
-2	-4	-4	-4	-2

图 3 梯度滤波器
Fig. 3 Gradient filter

将图像中的点坐标 (x, y) 转换到参数空间 (ρ, θ) ,每个点对应的累加器值加 1。通过改变 ρ 和 θ 值,能获得大量累加器,如图 5 中灰点所示。由于参数空间中的累加器会出现局部最大值,通

过检测该局部最大值,能够寻找到覆盖多数点 (x,y) 的直线参数方程 (ρ,θ) 。

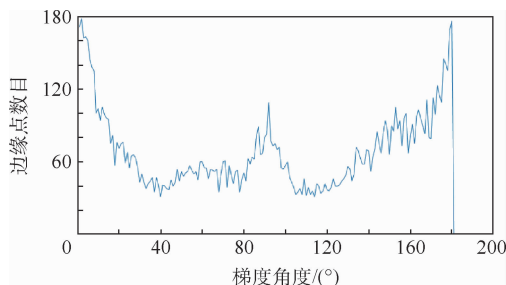


图4 边缘点梯度统计直方图

Fig.4 Gradient statistic histogram of edge points

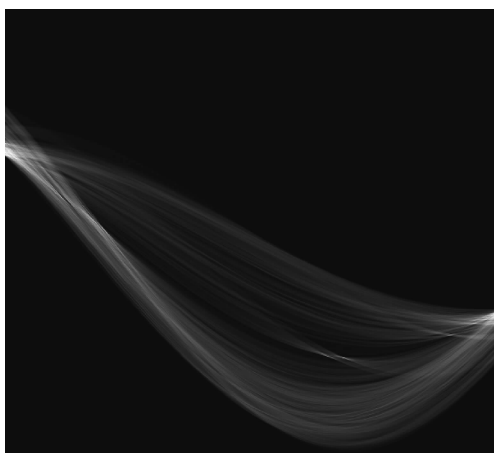


图5 边缘的 Hough 空间图像

Fig.5 Image of Hough space of all edges

1.4 非最大值抑制

车道线往往是图像下半部分中最长的线,因此取边缘点梯度统计直方图中的峰值 m_i 。Hough变换进行了离散化处理,每个峰值 m_i 向左右2个方向拓展 Δm ,因此在Hough变换的 ρ - θ 空间中,保留 $[m_i - \Delta m, m_i + \Delta m]$ 间的候选点,去除非其余最大值角度中的候选点,如图6所示。图6中绿色的区间就是结合梯度的方向和Hough变换的位置信息筛选出的候选区域。

车道线往往是各个角度最长的线,因此只保留每列的最大值,如图6所示。图6绿色区间的白色点就是每列的最大候选点。相邻车道线间存在夹角,因此把每个候选区域中的候选点 $[c_i^1, c_i^2, \dots, c_i^j, \dots]$ 合并成一个候选点 c_i , c_i 的位置参数 (ρ_i, θ_i) 为此段候选点 $[c_i^1, c_i^2, \dots, c_i^j, \dots]$ 中累加值最大的候选点的位置参数,累加值 A_i 为候选点 $[c_i^1, c_i^2, \dots, c_i^j, \dots]$ 中累加值 $[A_i^1, A_i^2, \dots, A_i^j, \dots]$ 的平均值:

$$A_i = \text{mean}(A_i^1, A_i^2, \dots, A_i^j, \dots) \quad (2)$$

筛选后的候选点,相比由Hough变换得到的初始候选点,数量大为减少,仅有几十个。如图6

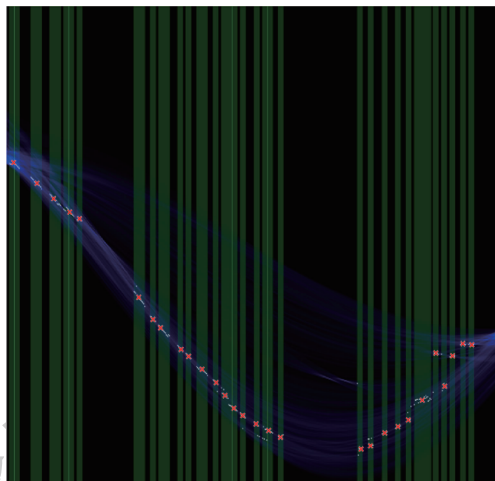


图6 Hough空间候选点的三次筛选

Fig.6 Candidate points in Hough space filtered for 3 times

中红色叉点所示,候选点的准确性提高。

1.5 动态规划优选 Hough 候选点

将候选点升序排列,分别做为概率矩阵 K 的行和列。概率矩阵 K 中, (θ_l, θ_r) 位置的元素为 p_{lr} ,表示左车道线的角度为 θ_l 、右车道线的角度为 θ_r 时车道成立的概率。在非最大值抑制步骤已经确定了 θ_l 的距离为 ρ_l 、 θ_r 的距离为 ρ_r ,因此 (ρ_l, θ_l) 代表了左车道线、 (ρ_r, θ_r) 代表了右车道线。概率 p_{lr} 为

$$p_{lr} = p_l + p_r + p_w + p_s \quad (3)$$

式中: p_l 为左车道线成立的概率; p_r 为右车道线成立的概率; p_w 为左右车道线间宽度成立的概率; p_s 为左右车道线长度成立的概率。

$$\begin{cases} p_l = \frac{A_l}{A_m} \\ p_r = \frac{A_r}{A_m} \\ p_w = \frac{d_{lr}}{d_0} \\ p_s = \begin{cases} 1 - \frac{1 - r_0}{1 + r_0} & 1 \geq r_0 > 0.67 \\ 1 - \left| \frac{r_0 - 0.33}{r_0 + 0.33} \right| & 0.67 \geq r_0 > 0 \end{cases} \\ d_{lr} = \frac{H}{2}(\cot \theta_l - \cot \theta_r) \\ r_0 = \frac{\min(A_l, A_r)}{\max(A_l, A_r)} \end{cases} \quad (4)$$

式中: p_l 为左车道线 (ρ_l, θ_l) Hough变换后累加值 A_l 与候选点中最大累加值 A_m 的比值; p_r 为右车道线 (ρ_r, θ_r) Hough变换后累加值 A_r 与 A_m 的比值; p_w 为左右车道线间宽度 d_{lr} 与车道经验宽度 d_0

的比率; H 为图像的高度; p_s 为左右车道线累加比值接近1或接近0.33的程度; r_0 为左右车道线累加比值。

按照高速道路的设计规则,最左边或者最右边的车道线是实线,其余车道线是虚线,虚线实线长2 m,间隔长4 m,占空比为0.33。如果左右车道线累加比值为1或者0.33,左右车道线成为相邻车道线的概率最大,其余比值下概率就小。

因此,概率矩阵 $\mathbf{K}$ 中的元素包括左车道线 (ρ_l, θ_l) 、右车道线 (ρ_r, θ_r) 、宽度 d_{lr} 和概率 p_{lr} 等4个属性。概率矩阵中的元素是一条车道的成立概率,因此车道检测就是在概率矩阵中从左上角到右下角的路径寻找最优的车道组合方案。采用动态规划选择路径内各车道概率均值最大、宽度方差最小的方案为最优的组合。目标函数 s 定义为

$$\begin{cases} \min(s) \\ p_{lr} \\ s = \lambda_1 \cdot \text{mean}(p_{lr}^i) + \lambda_2 \cdot \text{var}(d_{lr}^i) \end{cases} \quad (5)$$

式中: p_{lr}^i 和 d_{lr}^i 分别为路径中车道的概率和宽度; λ_1 和 λ_2 分别为车道概率均值和宽度方差的加权系数。

动态规划步骤如下:

步骤1 在概率矩阵中找到最大的 k 个值,构成 k 个初始种子组合。

步骤2 从步骤1的种子方案向右边生长,如第 i 个方案为 $p_i = [p_i^1, p_i^2, \dots, p_i^j, \dots]$,其最后车道 p_i^j 在矩阵中的位置为 (l_i, r_i) 。在 r_i 行找到宽度相近、概率最大的 m 个车道,其概率值为 $r_i^t (t \in [0, m])$,放入方案中作为新的、最右边的成员 $p_i^{(j+1)}$,计算此方案的概率均值和宽度方差。

步骤3 从所有种子方案中挑选出概率均值和宽度方差加权和最大的前 k 个组合。

步骤4 迭代步骤2、步骤3,直到右边没有候选点。

步骤5 将种子方案向左边生长,按步骤2、步骤3迭代直到左边没有候选点。

步骤6 选择概率均值和宽度方差加权和最大的方案作为结果。

按高速道路设计原则,道路最左和最右的车道线是实线,最右车道线的右边是路肩。因此,动态规划寻优得到的方案中,最左线段累加值与此方案中各线段最大累加值的比值应接近1;如果最左的多条相邻线段的比值相近且都接近1,那么选择该多条相邻线段中最右的线段为方案的最左车道线,其余作为围栏等干扰,从方案中去除;如果最右的多条相邻线段的比值相近且都接近1,那么选择该多条相邻线段中最左的线段为方案的最右车道线,最左的线段右边邻接线段作为路肩,其余作为围栏等干扰,也从方案中去除。

2 实验

2.1 自有高速数据的检测实验

从笔者拍摄的高速行驶视频中抽取200张评测本文算法,结果如图7所示。图7中红线是最左、最右车道线,紫线是路肩的车道线,绿线是淘汰的干扰线,中间的红点为消失点。可以看出,本文算法能够准确地检测出实车道线、虚车道线、路肩,结构化解析高速道路,可以用于高速道路的动态执法。



图7 高速道路的车道检测结果

Fig.7 Lane detection results of expressways

2.2 自有非高速数据的检测实验

本文算法主要针对高速道路的结构化解析。为了验证本文算法对普通道路的效果,也从笔者拍摄的行驶视频中抽取 50 张进行评测。与高速道路不同,普通道路最右实车道线的右边没有路肩,因此对动态规划的规则进行了相应修改,结果如图 8 所示。与图 7 相同,图中红线是最左、最右车道线,绿线是淘汰的干扰线,中间的红点为消失点。图 8 中间的例子,由于停车的遮挡,左边的实车道线被误检为虚线。可以看出,本文算法能够较准确地检测出实车道线、虚车道线,结构化解析普通道路,也可以用于普通道路的动态执法。

2.3 对比实验

为了评比本文算法的性能,将其与传统的 Hough 变换车道检测^[8]、层次化 Hough 变换车道检测^[9]、扩展双曲线模型车道检测^[13]及深度学习

的 LaneNet^[14]、UPSNet^[19]在 i7-6700K 处理器(无 GPU)、MATLAB2017 的平台上进行了对比。从 KITTI^[21]的高速道路中选择 100 张样本作为测试集。如果检测到的车道线段有 30% 或者以上落在 ground truth 中,就认为是正确检测到车道;否则失败。定义准确率为检测到的车道线是真实车道线的数量比例,即

Accuracy = TP/(TP + TN) (6)

召回率为检测正确的车道线占 ground truth 中全部车道线的比率,即

Recall = TP/(TP + FP) (7)

式中:TP(True Positive)为正确检测到的真实车道线数量;TN(True Negative)为将非车道线检测为真实车道线的数量;FP(False Negative)为未检测到的真实车道线数量。

实验结果如表 1 所示。

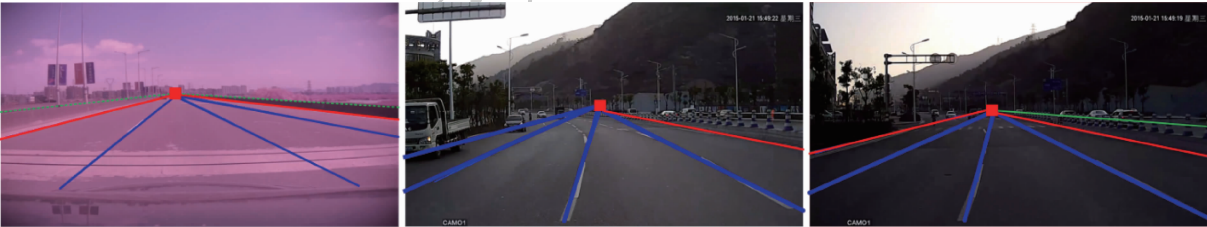


图 8 普通道路的车道检测结果
Fig. 8 Lane detection results of urban roads

表 1 不同算法的性能对比
Table 1 Comparison of performance between different algorithms

算法	准确率/%	召回率/%	速度/FPS
文献[8]	63.47	59.04	8.2
文献[9]	76.29	73.34	32.7
文献[13]	79.57	87.13	23.6
LaneNet ^[14]	92.81	93.19	18.4
UPSNet ^[19]	89.96	91.37	13.6
本文算法	91.23	94.18	36.5

注:FPS为帧/s。

基于传统的 Hough 变换车道检测^[8]的准确率和召回率只有 63.47% 和 59.04%,速度只有 8.2 FPS;基于层次化 Hough 变换的车道检测^[9]的准确率和召回率均有较大提高,达到了 76.29% 和 73.34%,速度提高更快,达到了 32.7 FPS;基于扩展双曲线模型车道检测^[13]的准确率和召回率一步提高到 79.57% 和 87.13%,速度为 23.6 FPS;LaneNet^[14]和 UPSNet^[19]都是基于深度学习,没有 GPU 的支持,速度只有 18.4 FPS 和 13.6 FPS,但是准确率分别达到 92.81% (最大值)和 89.96%,召回率也分别达到 93.19% 和 91.37%;本文算法筛选掉了大部分伪车道线,能

够去掉阴影和近似印迹,但是容易受车辆遮挡,准确率为 91.23%,不如 LaneNet 算法的 92.81%,但是召回率达到了最大值 94.18%,而且检测速度也达到了最大值 36.5 FPS。

3 结 论

1) 相比于传统的基于 Hough 变换车道检测的算法,本文算法利用边缘梯度减少了 Hough 变换后候选点的数量,基本上排除了不含车道线信息的 Hough 空间候选点,保留了包含车道线信息的候选点,并采用动态规划筛选出真实车道线,定位到道路上的全部车道,结构化解析高速道路,能够用于道路的动态执法。

2) 算法简单,计算资源耗费少,能够用于实时的道路分析检测。

参考文献 (References)

[1] ZHANG X Q, HU W M, XIE N H, et al. A robust tracking system for low frame rate video[J]. International Journal of Computer Vision, 2015, 115(3) :279-304.
[2] ZHANG X Q, HU W M, CHEN S Y, et al. Graph-embedding-based learning for robust object tracking[J]. IEEE Transactions

on Industrial Electronics,2014,61(2):1072-1084.

[3] ZHANG X Q,WANG D,ZHOU Z Y,et al. Robust low-rank tensor recovery with rectification and alignment[J/OL]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019 (2019-07-16) [2020-03-01]. DOI: 10. 1109/TPAMI. 2019. 2929043.

[4] LINGWORTH J,KITTLER J. A survey of the Hough transforms [J]. Computer Vision, Graphics, and Image Processing, 1988, 43(2):280-289.

[5] ZHANG X Q,HU W M,QU W,et al. Multiple object tracking via species-based particle swarm optimization[J]. IEEE Transactions on Circuits and Systems for Video Technology,2010, 20 (11):1590-1602.

[6] ZHANG X Q,LI C C,HU W M,et al. Human pose estimation and tracking via parsing a tree structure based human model [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems,2014,44(5):580-592.

[7] ZHANG X Q,HU W M,BAO H J,et al. Robust head tracking based on multiple cues fusion in the kernel-Bayesian framework [J]. IEEE Transactions on Circuits and Systems for Video Technology,2013,23(7):1197-1208.

[8] SHIN B S,TAO J,KLETTE R. A superparticle filter for lane detection[J]. Pattern Recognition,2015,48(11):3333-3345.

[9] SATZODA R K,SATHYANARAYANA S,SRIKANTHAN T,et al. Hierarchical additive Hough transform for lane detection [J]. IEEE Embedded Systems Letters,2010,2(2):23-26.

[10] KATRU A. Improved parallel lane detection using modified additive Hough transform [J]. International Journal of Image, Graphics and Signal Processing,2016,8(11):10-17.

[11] WANG Y,TEOH E K,SHEN D. Lane detection and tracking using B-snake[J]. Image and Vision Computing,2004,22(4):269-280.

[12] JUNG C R,KELBER C R. Lane following and lane departure using a linear parabolic model[J]. Image and Vision Computing,2005,23(13):1192-1202.

[13] BERTOZZI M,BROGGI A. Gold:A parallel real-time stereo vision system for generic obstacle and lane detection[J]. IEEE Transactions on Image Processing,1998,7(1):62-81.

[14] WANG Z,REN W Q,QIU Q. LaneNet:Real-time lane detection networks for autonomous driving [EB/OL]. (2018-07-04) [2020-03-01]. <https://arxiv.org/abs/1807.01726>.

[15] GARNETT N,COHEN R,PE'ER T,et al. 3D-LaneNet:End-to-end 3D multiple lane detection [EB/OL]. (2018-11-26) [2020-03-01]. <https://arxiv.org/abs/1811.10203>.

[16] NEVEN D,BRABANDERE B D,GEORGULIS S,et al. Towards end-to-end lane detection: An instance segmentation approach[C]//IEEE Intelligent Vehicles Symposium (IV). Piscataway:IEEE Press,2018:286-291.

[17] LIANG D,GUO Y C,ZHANG S K,et al. LineNet:A zoomable CNN for crowdsourced high definition maps modeling in urban environments[EB/OL]. (2018-07-16) [2020-03-01]. <https://arxiv.org/abs/1807.05696v1>.

[18] PAN X,SHI J,LUO P,et al. Spatial as deep:Spatial CNN for traffic scene understanding[C]//The 32nd AAAI Conference on Artificial Intelligence(AAAI-18). Palo Alto:AAAI,2019.

[19] XIONG Y W,LIAO R J,ZHAO H S,et al. UPSNet:A unified panoptic segmentation network [C]//IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway:IEEE Press,2019.

[20] DING Y,XU Z,ZHANG Y B,et al. Fast lane detection based on bird's eye view and improved random sample consensus algorithm [J]. Multimedia Tools and Applications,2017, 76 (21):22979-22998.

[21] GEIGER A,LENZ P,STILLER C,et al. Vision meets robotics: The KITTI dataset [J]. International Journal of Robotics Research,2013,32(11):1231-1237.

作者简介:
罗胜 男,博士,副教授。主要研究方向:机器视觉、机器学习和机器人。

赵丽 女,硕士,助理研究员。主要研究方向:检测、跟踪、视频分析和机器学习。

王慕抽 男,硕士,助理研究员。主要研究方向:图像处理、计算机视觉和机器学习。

Lane semantic analysis based on road feature information

LUO Sheng, ZHAO Li*, WANG Muchou

(College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou 325000, China)

Abstract: Law enforcement on express roads in moving car requires to semantically analyze the road by lane detection algorithm, but the accuracy and recall rate of the algorithms based on human-crafted features are not good enough, and the algorithms based on deep learning require too much computing resource. Therefore, this paper proposes a semantical analysis algorithm based on road feature information. The proposed algorithm makes use of the gradient statistical information of edge points to filter out the candidate points in Hough space, and dynamic programming to find the most reasonable solution of lane line combination among the remaining candidate points. Thus it can accurately find all lane markings on roads with less computing resource. The experiment with self acquisition of data shows that the proposed method can structurally find all lanes on structured and unstructured roads. In a comparative experiment, contrasted with some other traditional lane detection methods and some deep learning networks, the proposed algorithm demonstrates its improvement in accuracy, recall rate and computing speed.

Keywords: lane detection; road analysis; dynamic law enforcement; Hough transformation; dynamic planning

Received: 2020-03-03; **Accepted:** 2020-03-20; **Published online:** 2020-03-27 13:27

URL: kns.cnki.net/kcms/detail/11.2625.V.20200326.2017.003.html

Foundation items: Zhejiang Provincial Natural Science Foundation of China (LQ19F020005); Zhejiang Province Public Welfare Technology Project (2016C31117); Wenzhou Science and Technology Project (2018ZG021)

* **Corresponding author.** E-mail: lizhao@wzu.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0067

基于高置信局部特征的车辆重识别优化算法



鑫泽^{1,2}, 盛浩^{1,2,*}, 吕凯^{1,2}, 刘洋^{1,2}, 张洋^{1,2}, 吴玉彬^{1,2}, 柯韦³

(1. 北京航空航天大学 计算机学院 软件开发环境国家重点实验室, 北京 100083;

2. 北京航空航天大学 大数据与脑机智能高精尖创新中心, 北京 100083;

3. 澳门理工学院 应用科学高等学校, 澳门 999078)

摘 要: 根据车辆重识别中区域置信度不同, 提出了基于高置信局部特征的车辆重识别优化算法。首先, 利用车辆关键点检测获得对应的多个关键点坐标信息, 分割出车标扩散区域和其他重要的局部区域。根据车标扩散区域的高区分度特性, 提升局部区域的置信度。使用多层卷积神经网络对输入图片进行处理, 根据局部区域分割信息, 对卷积得到的特征张量进行空间维度上的切割, 获得代表全局信息和关键局部信息的特征张量。然后, 通过全连接层特征张量转化为表示车辆个体的一维向量, 计算损失函数。最后, 在测试阶段使用全局特征, 并利用训练好的车标扩散区域提取分支获得高置信局部特征, 缩短局部识别一致的车辆目标距离。在典型车辆重识别数据集 VehicleID 上进行测试, 验证了所提算法的有效性。

关 键 词: 车辆重识别; 高置信局部特征; 特征优化; 区域检测; 神经网络

中图分类号: TP399

文献标志码: A

文章编号: 1001-5965(2020)09-1650-10

随着监控系统智能化需求^[1-2]的发展, 重识别问题获得了越来越多的关注^[3]。车辆重识别算法能够应用在车辆识别、追踪和定位等众多难点问题上, 进一步提升公安监控系统的安全性和可靠性。车辆重识别所针对的问题是: 如何在一个图片集合中快速、准确地找到与给定车辆图片中完全相同的车辆。车辆的角度、光照、环境、遮挡物等许多因素给车辆重识别带来了巨大的挑战。

车辆重识别在国内视频监控研究领域受到了较大关注。基于车辆的定位和照片的时间戳等信息的方法^[4]能够利用额外的信息来优化车辆的重识别结果。基于重要局部特征的方法^[5]能够为不同的局部设计权重, 来影响最后拼接的特征

向量。关注车辆方向差异的方法^[6]能够在车辆图片计算距离时, 根据具体的方向选择对应的全局特征。基于部件融合特征的方法^[7]能够检测车窗和车脸区域, 并将提取出的特征进行融合, 利用融合后的特征计算距离, 用来进行车辆的分类识别。基于特征融合和度量学习的方法^[8]从降低车辆所处环境影响的角度出发, 对图片进行特征表示, 通过计算马氏距离来重识别车辆。基于特征融合的方法^[9]可以对融合后的特征矩阵进行奇异值分解, 提取特征值, 并利用自适应调整算法来优化神经网络, 提升算法鲁棒性。

在国外, 针对车辆重识别问题也有很多有意义的研究工作。基于无监督的注意力方法^[10]能够在重识别的过程中自动获得车辆的重要局部区

收稿日期: 2020-03-02; 录用日期: 2020-03-20; 网络出版时间: 2020-04-24 08:57

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200423.1757.002.html

基金项目: 国家重点研发计划(2018YFB2100500); 国家自然科学基金(61861166002, 61872025, 61635002); 澳门特别行政区科学技术发展基金(0001/2018/AFJ); 软件开发环境国家重点实验室开放基金(SKLSDE2019ZX-04)

*通信作者: E-mail: shenghao@buaa.edu.cn

引用格式: 鑫泽, 盛浩, 吕凯, 等. 基于高置信局部特征的车辆重识别优化算法[J]. 北京航空航天大学学报, 2020, 46(9): 1650-1659. DOU X Z, SHENG H, LYU K, et al. Vehicle re-identification optimization algorithm based on high-confidence local features [J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1650-1659 (in Chinese).

域。在多相机系统中,以车辆三维检测结果为基础的方法^[11]能够准确检测出车辆形状、颜色和方向等信息,用于改善车辆的特征。类似地,根据车辆不同的方向来确定关键点的位置^[12],提取车辆的细节特征,并结合全局特征,能够提升模型的识别准确率。此外,多任务同步学习算法^[13]能够在提取车辆特征的同时关注车辆车型和方向等信息,借助学习到的额外信息来改善识别结果。

车辆重识别问题的主要难点在于:在交通监控视频中,汽车数量庞大,而车辆种类有限,必然会有很多外形极为相似的车辆,区分难度较高。然而,在人类视觉中,高置信局部特征可以为相似车辆分类提供重要线索。参考类人视觉的原理,本文选取车标扩散区域作为提取特征的高置信局部区域,车标及其周边区域是车辆重要的高置信特征,几乎所有出现在交通监控视频中的车辆都可以捕获到车标信息。因此,通过提取车标扩散区域特征,可以有效增强不同车辆之间的区分度。

本文利用车辆不同区域置信度不同,提出了基于高置信局部特征的车辆重识别优化(High Confidence Local Features for Vehicle Re-Identification Optimization, HCLO)算法。首先,利用卷积神经网络提取车辆的高置信局部特征,通过车辆关键点选取带有具体局部标记的车辆图片检测车标扩散区域,并训练好特征提取模型,对具体的高置信局部特征进行准确分割。然后,在局部区域检测阶段,与车标扩散区域一同检测的还有车窗和车灯区域,这些局部区域作为辅助局部特征,利用在特征图上截取特征张量的方式,辅助训练全局特征。通过高置信局部特征对计算得到的全局特征向量距离进行优化。最后,设计验证集,确定预测结果的最优权重参数,根据局部特征预测结果来增强目标车辆间的区分度,通过这种方式优化全局特征向量之间的距离,提升识别准确率。

本文的主要贡献如下:

- 1) 以车辆关键点检测算法为基础,设计了车辆高置信局部区域提取算法,能够精确提取车辆的高置信局部特征。
- 2) 训练阶段,利用车辆关键点检测算法分割出车辆关键的局部区域,辅助训练全局特征。
- 3) 测试阶段,利用高置信局部特征的识别结果,对图片全局特征的距离进行改善,提升车辆重识别的准确率。

1 车辆高置信局部特征

1.1 高置信局部特征定义

高置信局部特征是车辆重识别的重要线索,能够作为依据精准地区分外形相似的车辆,能够作为高置信特征的局部区域需要满足 2 个条件:特征提取的可行性和具有区分车辆的判别能力。本文选取车标扩散区域作为高置信局部特征提取区域。

车标是车标扩散区域的核心特征,作为汽车品牌标志几乎存在于所有的交通车辆上,并且位置相对固定,一般是车头和车尾相对靠下的中央区域。在设计车标扩散区域时,为了突出品牌独特性,使消费者能够容易地与其他汽车品牌进行区分,往往会设计风格迥异的车标。此外,车标扩散区域的组成比较简单,大部分由简洁的银白色线条和图案组成,少部分使用了鲜明颜色。图 1 展示了部分常见的车标扩散区域。车辆品牌不同,设计风格具有明显的差异,因此,车标扩散区域成为了车辆重识别一个重要的局部特征。上述先决条件使得卷积神经网络算法能够比较容易地提取区域特征,为基于高置信局部特征区域提升区分度提供了可行性的依据。

虽然汽车制造商的设计理念差异很大,但设计的具体汽车却有很多极为相似的车型,甚至有很多贴牌生产的车型。此外,城市监控摄像头所拍摄的图片角度比较固定,甚至无法清晰地捕捉车辆各个角度的细节。本文提取车标扩散区域特征作为高置信局部特征,提升相似车辆区分度,提高识别准确率。图 2 展示了测试数据集中车标扩散区域不同,外形却很相似的白色轿车。



图 1 常见的车标扩散区域

Fig. 1 Popular vehicle brand extension regions



图 2 不同车标扩散区域的相似车型

Fig. 2 Similar vehicle models with different vehicle brand extension regions

的拍摄角度完全一致,外形也极其相似,但根据矩形框中的车标扩散区域却可以进行区分。

1.2 高置信局部特征区域检测

通过提取高置信局部特征优化识别结果,前提条件是要能够准确定位高置信局部特征区域,即车标扩散区域的位置信息。车辆关键点检测算法能够检测到车辆的关键点,如前后车灯、4 个车轮、左右倒车镜、车窗顶点等重要位置,利用关键点坐标,可以准确定位车标扩散区域。本文利用 Stacked Hourglass 网络^[14] 获取车辆关键点,进而检测并提取高置信局部特征区域。

Stacked Hourglass 网络最初被用来识别行人的姿势,为了准确地捕捉局部信息,利用漏斗形的网络在每一层都提取特征。通过这种漏斗形状结构的堆积来实现关键点的获取。本文通过改进 Stacked Hourglass 网络,用于车辆关键点检测。图 3 展示了车辆关键点的检测结果,一共检测 36 个关键点,对称地分布在车身两侧,用线条将相关的关键点连接,能够准确分割车辆的表面区域,进而获取车标扩散区域。其中,红色线条表示由车体左侧关键点连接所形成的车体轮廓,绿色线条表示由车体右侧关键点连接所形成的车体轮廓,蓝色线条由对称的 2 个关键点连接形成。

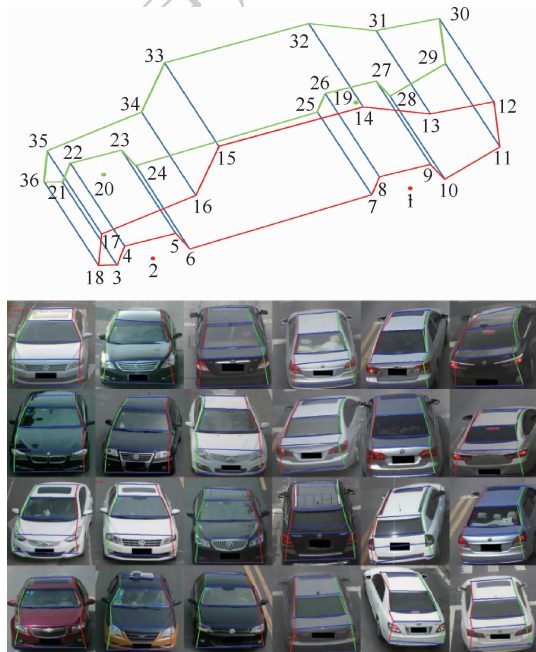


图 3 车辆关键点检测结果

Fig. 3 Vehicle key points detection results

1.3 高置信局部特征提取

完成一致化后,利用卷积神经网络提取矩形车标扩散区域的高置信局部特征。在本文所采用的数据集中,车辆图片的车标扩散区域几乎包括了常见车辆的所有车标,对其从 0 开始标记序号,

高置信局部特征最后会转化成具体的预测数值,便于用来计算局部特征识别的准确率。

卷积神经网络在计算机视觉领域得到了广泛应用,在图片分类、目标检测、目标跟踪等问题上具有显著超出传统算法的表现。高置信局部特征提取的网络模型以 ResNet50^[15] 为基础进行设计改进,在网络之后设计了额外的操作。通过 ResNet50 网络对图片进行处理,每张图片可以得到一个三维的图片张量,维度分别为图片的高度、宽度和通道数。通过车标扩散区域检测算法能够得到车标所在矩形区域的 4 个顶点坐标。因为卷积后的图片高与宽在分辨率上与原图片并不匹配,所以按照现有高和宽的数值对坐标数据进行近似,使得能够在三维特征张量上切割出表示高置信局部特征的张量。将获取到的张量进行全局平均池化(Global Average Pooling, GAP)操作,该操作在通道的每一维上计算高度和宽度 2 个维度内的平均值,作为池化后的数值。全局平均池化操作后,张量从三维变到一维,长度与通道数相等。对向量进行 1×1 卷积操作,其目的是缩短特征向量的维度,减少计算过程中的误差, 1×1 卷积核的数量等于维度缩减后希望的向量长度。在本文提出的算法中,设置为 256 维。

通过上述计算步骤,对每一张图片都得到了一个 256 维的表示向量。此向量并不能表示具体的高置信局部特征提取结果,需添加一层全连接层,将向量的长度转化成训练集中局部特征序号的数量。再做 softmax 操作,使得向量数值之和等于 1。

$$x'_i = \frac{e^{x_i}}{\sum_{j=1}^{\text{len}} e^{x_j}} \quad (1)$$

式中: x 为变化前的向量; x' 为 softmax 操作变化后的向量;len 为向量的长度。

变化后的向量表示的含义是输入车辆图片高置信局部特征在各个维度上的概率。选取概率最大数值所在的维度序号作为模型所预测的局部特征。训练时,借助局部特征的真实值使用交叉熵损失函数。

$$\text{loss} = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^M y_{ij} \ln P_{ij} \quad (2)$$

式中: N 为所训练的图片数量; M 为局部特征的总数; y_{ij} 表示第 i 张车辆图片的局部特征是否为 j ; P_{ij} 为模型预测第 i 张图片的局部特征为 j 的概率。

在应用于车辆重识别之前,先做单独的局部特征识别准确率测试,用来验证高置信局部特征

的准确性。高置信局部特征提取算法的主要计算过程可表示为

$$L = E(Wf(x, \theta) + b, y) \quad (3)$$

式中: E 为交叉熵损失函数; W 和 b 分别为最后全连接层的权重和偏置参数; 函数 f 表示提取高置信局部特征的多层卷积神经网络, 其参数为 θ ; x 为原始的图片; y 为局部特征的真实值。

2 基于高置信局部特征的车辆重识别优化网络

2.1 网络结构设计

图 4 为基于高置信局部特征的车辆重识别优化算法的网络结构。网络结构整体上可以分为 2 个分支: 上方的高置信局部特征分支和下方的主分支。输入图片上标注的 4 个红色虚线矩形框表示根据车辆关键点所截取出的车标、车窗和车

灯等关键局部信息, 2 个分支都使用 ResNet50 作为基础的特征提取网络, 但是并不共享权重, 因为主分支关注全局特征, 而局部特征分支更关注车标扩散区域特征。高置信局部特征分支负责提取高置信的车标扩散区域特征, ResNet50 得到的特征图大小是 $16 \times 16 \times 2048$, 根据关键区域信息对其进行切割操作, 得到的车标扩散区域特征图大小为 $5 \times 5 \times 2048$, 再进行全局平均池化操作和 1×1 卷积操作, 高置信局部特征的维度变化为 $1 \times 1 \times 256$; 最后经过全连接层, 计算损失函数。主分支负责提取车辆图片的全局特征, 用来计算目标车辆间的距离, 将 ResNet50 得到的特征图分割成 4 个部分, 在图 4 中, 使用 4 个颜色不同的长方体表示, 分别表示全局、车标、车灯、车窗; 每个部分的计算流程与高置信局部特征分支处理车标扩散区域的步骤一致, 最后计算 4 个损失函数。

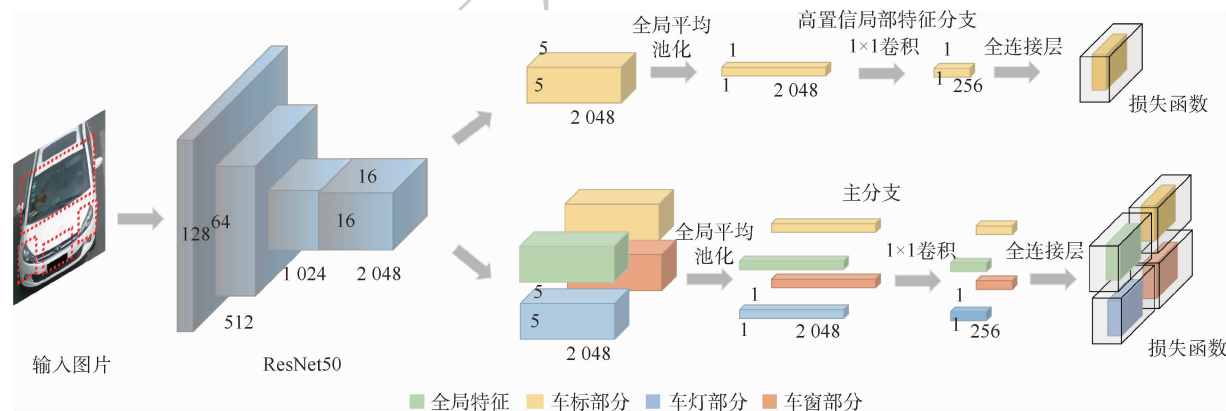


图 4 基于高置信局部特征的车辆重识别优化算法网络结构

Fig. 4 Network structure of vehicle re-identification optimization algorithm based on high-confidence local features

受 Part-Regularized 网络模型^[16]启发, 本文算法利用提取局部信息作为辅助局部特征的方式来补充全局特征, 使向量的表征能力更强。车辆关键点检测算法除了获取车标扩散区域以外, 还获取了车灯区域和车窗区域。在主分支中, 车标扩散区域信息被重复利用。与高置信局部特征分支中抽取局部张量类似, 提取 ResNet50 的输出张量能够得到 3 个局部信息, 其中车灯的局部有 2 个, 提取后取平均值。此时, 局部特征连同全局特征共得到 4 组特征张量。对每一组特征都进行全局平均池化操作和 1×1 卷积操作, 将三维的张量转化成一维向量, 计算损失函数。

2.2 损失函数

除高置信局部特征分支外, 主分支里共需要计算 4 个损失函数, 包括全局特征和辅助局部特征, 具体分别为全图特征、车标辅助特征、车窗辅助特征和车灯辅助特征的损失函数。4 组特征的真实值都设置为车辆真实 ID。在测试过程中, 并

不将 4 组特征进行拼接, 而是只使用全局特征作为计算距离的向量。所以, 在训练过程中, 另外 3 组局部特征只是起到辅助全局特征进行训练的作用。主分支的损失函数与高置信局部特征分支相同, 都使用交叉熵损失函数。

$$L_r = E(W_r f_r(x, \theta) + b_r, y_{\text{id}}) \quad (4)$$

$$r \in \{\text{global, window, light, brand}\} \quad (5)$$

式中: r 表示 4 个特征, 即全局、车窗、车灯和车标; W_r 和 b_r 分别为对应的全连接层的权重和偏置参数; f_r 为提取关键区域特征的函数; y_{id} 表示图片对应的真实车辆 ID。

所以, 主分支的总损失函数为

$$L = \sum_r L_r \quad (6)$$

式中: r 根据式 (5) 选取数值; L_r 为对应特征损失函数。总损失函数为各个损失函数的加总。

2.3 高置信局部特征优化

车辆重识别问题, 是在一个图片集合中找到

与给定车辆图片完全相同的车辆。给定的图片命名为 query 集合,要在其中搜索的图片集合命名为 gallery 集合。对于每张图片,本文算法模型都会得到表征向量和高置信局部特征预测值。为了在 gallery 图片集合中找到与 query 图片最相似的图片,相似度的计算方式采用计算全局特征向量之间的欧氏距离:

$$\text{dist} = \sqrt{\sum_{i=1}^{\text{len}} (a_i - b_i)^2}$$
 (7)

式中:dist 为计算得到的距离;a 和 b 分别为需要计算距离的 2 个长度相等的向量。

按照 gallery 集合中每张图片与 query 图片的距离从小到大,对 gallery 数据集中的图片按照相似度从高到低进行排序,获得高置信局部特征的相似度度量结果。该结果被用来改善排序后的 gallery 序列。

在车辆重识别中,车辆完全相同的 2 张图片,局部特征必然一致。高置信局部特征不同的图片,则表示确定的不相同车辆。根据这一点,可以根据局部特征的相似度度量结果来改善图片间距离,通过添加权重参数的方式使得局部特征预测 2 张相同图片的距离适当缩短。

$$w_{ij} = \begin{cases} \lambda & \text{pred}(i) = \text{pred}(j) \\ 1 & \text{pred}(i) \neq \text{pred}(j) \end{cases}$$
 (8)

式中:w_{ij}为施加在第 i 张 query 图片和第 j 张 gallery 图片距离上的权重参数;λ 为一个小于 1 的正实数常量;pred 函数表示预测高置信局部特征的神经网络,其输出是一个具体的局部特征相似度数值。

获得的权重参数与原始的向量距离相乘,可以获得更新后的向量距离:

$$\text{dist}_{ij}^* = w_{ij} \cdot \text{dist}_{ij}$$
 (9)

式中:dist_{ij}为第 i 张 query 图片和第 j 张 gallery 图片向量的原始欧氏距离;dist_{ij}^{*}为更新后的距离。

如果对于 2 张图片算法预测的高置信局部特征相同,因为参数 λ 是小于 1 的,所有向量距离会缩短,图片的相似度提高,在排序后的 gallery 序列中更靠前。若预测的局部特征不同,则距离保持不变。通过这种方式,利用高置信局部特征实现了图片的区分度优化。此外,为了设置一个合适的 λ 值,需要额外设计验证数据集以获得一个合适的数值。

2.4 算法流程

在主分支和高置信局部特征分支的基础上,本文算法的流程如图 5 所示,步骤如下:

步骤 1 划分高置信局部特征训练和测试的

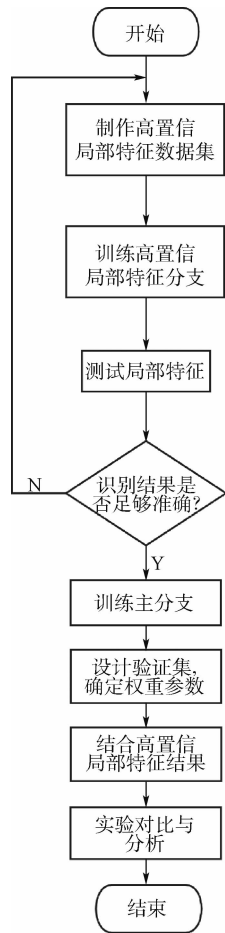


图 5 本文算法流程

Fig. 5 Proposed algorithm flowchart

数据集,对局部特征进行排序,标记训练集中所有车辆的局部区域。

步骤 2 训练模型中的高置信局部特征分支,对于每张输入图片都能得到预测的具体局部特征,并将训练好的模型在单独的数据集上做测试,使其能获得较好的识别表现,否则重新训练。

步骤 3 划分车辆重识别训练和测试的数据集,检测所有车辆图片的关键点坐标,划分需要的局部区域,将得到的区域坐标信息以图片名称为索引全部保存到一个文件中。

步骤 4 读入车辆图片和相关的局部区域坐标,训练车辆重识别的主分支。

步骤 5 引入高置信局部特征结果,在公开数据集中,寻找没有参与模型训练和测试的图片,设计验证集。在验证集上以不同的参数 λ 不断地进行测试,记录每个数值的识别表现结果,选取最优的 λ 数值。

步骤 6 在车辆重识别测试集上评估本文算法模型的表现,计算性能量化评估指标。整理其他车辆重识别算法的结果,最后进行结果的对比和分析。

3 实 验

3.1 实验数据集

实验所采用的数据集是 VehicleID 数据集^[17]。VehicleID 数据集是使用最频繁、关注度最高的车辆重识别数据集之一,并且有着丰富的车辆信息标注。训练集共有图片 110 178 张,车辆 ID 的数量为 13 164。因为测试的图片很多,所以数据集作者设计了 3 个大小的测试集,分别命名为 small、medium 和 large,其中所具有的车辆 ID 数量分别为 800 个、1 600 个和 2 400 个。在 3 个测试集中,不对 query 集合和 gallery 集合进行划分,在测试时,需要在每个 ID 的几张图片中,随机抽取一张图片加入 gallery 集合当中,剩余图片自动并入到 query 集合中。所以,各个尺寸的测试集中,gallery 集合大小与车辆 ID 数量相等,对于每一张 query 图片,gallery 中只有一张相同车辆图片。VehicleID 数据集的具体信息可参见表 1。

此外,VehicleID 数据集还有一个重要特点:所有车辆都是正面或背面图像。尽管一小部分车辆图片存在轻微歪斜,但正面车头车窗或者背面车尾信息并无丢失。基于高置信局部特征优化距离的方法所需车标扩散区域都存在于车辆正面或背面,因此,VehicleID 数据集这一特点满足本文算法模型所需条件。在训练全局特征向量时,需使用的车灯、车窗等辅助局部信息也可以比较容易地获得。

表 1 VehicleID 数据集信息

Table 1 VehicleID dataset information

数据集	图片数量	车辆 ID 数量
训练集	110 178	13 164
small 测试集	7 332	800
medium 测试集	12 995	1 600
large 测试集	20 038	2 400

3.2 评估指标

在排序后的 gallery 图片序列中,只有一张图片是真实值。实验使用 top-K 作为重识别结果的评估指标,表示在排序后的 gallery 序列里前 K 张图片里包括真实值图片的概率,K 常见的取值有 1、5 和 10。具体的计算公式为

$$\text{top-K} = \frac{1}{Q} \sum_{i=1}^Q \sum_{j=1}^K g_{ij} \quad (10)$$

式中:Q 为 query 数据集的图片数量; g_{ij} 为第 i 张 query 图片和第 j 张图片所代表的车辆是否相同,若相同,其值为 1,否则,其值为 0。

3.3 实验参数设置

车辆关键点检测算法以 Stacked Hourglass 网

络为基础,共标记 36 个关键点。每个 batch 的大小设置为 4,共有 200 个 epoch,每个 epoch 设置了 1 000 次迭代,初始的学习率设置为 2.5×10^{-4} 。训练过程中,当准确率连续 5 次处于平稳状态时,下降学习率,当前的学习率变为 0.96 倍。

多层卷积神经网络的训练条件十分繁杂,相同的网络结构、不同的训练条件,模型会有截然不同的表现。在本文实验中,主分支的训练使用梯度下降法修改权重参数的值,迭代次数为 240 次。训练过程中,参数的学习率对模型表现有很大影响,本文实验中,学习率初始值设置为 0.1,在 150 次迭代时学习率下降到 0.01,在 220 次迭代时学习率再下降到 0.001。当学习率不变时,训练的损失下降到一定程度后不再继续下降,在一定范围内振荡。采用学习率衰减的策略进行训练可以避免这一情况的发生,让模型训练更充分。batch 大小设置为 64。输入图片全部将大小一致化为 256×256 像素,以增强数据。

在本文模型使用的 ResNet50 网络中,最后一层降采样过程里,原始步长(stride)设置为 2。降采样的目的是为了缩小特征图的空间尺寸,便于后续操作。但图片分辨率越高,容纳信息也就越多。为了提高特征图的表示能力,本文算法将 ResNet50 最后降采样的步长设置为 1,通过传递更多信息的方式来提升模型表现。

此外,随机消除(Random Erasing)^[18]也是一种数据增强策略。为了模拟图片中存在遮挡物的现象,对于部分图片可以随机选择一个矩形区域消除掉所有的像素值。对于每一张图片,设使用随机消除的概率为 p_{re} ,本文实验中设置为 0.5。随机选取一个点的坐标 (p, q) ,再随机选取一个矩形的长 length 和宽 width;如果 $(p + \text{length}, q + \text{width})$ 所表示的点不在图片之中,则重新随机;直到 $(p + \text{length}, q + \text{width})$ 在图片之中为止。这样就得到了一个矩形对角线的 2 个顶点,在该随机过程中,将矩形的纵横比限制在一定范围内,矩形面积与整张图片面积的比例设置为 0.02 ~ 0.3 之间。所选矩形内的像素并不全部清除,而是设置为各个像素点的平均值。通过这种方式,每张输入图片都有 0.5 的概率会有一个矩形区域覆盖掉,模拟了实际中存在遮挡物的情形,增强了模型的泛化能力。

3.4 高置信局部特征模型的训练分析

高置信局部特征分支需要在主分支训练和测试之前进行训练,才能将局部预测结果应用到主分支当中。VehicleID 数据集中标记了一部分图

片的车型,车型总数共计 250 个,对这些车型按照是否具有相同车标这一标准进行聚类,可以得到 46 种不同的车标扩散区域。对于高置信局部特征模型的训练,一共选取 50 000 张图片,其中 40 000 张图片用来训练,10 000 张图片用来测试车标扩散区域识别的准确率。计算识别准确率的公式为

$$Accuracy = \frac{1}{T} \sum_{i=1}^T ans_i \tag{11}$$

式中: T 为车标扩散区域识别测试集中的图片数量; ans_i 为第 i 张图片的局部特征预测结果,若预测正确,其值为 1,预测错误,其值为 0。

在实际实验中发现,虽然标注了车型的图片 and 用于重识别问题的图片都来自于同一数据集,但是二者本身存在较大差异:前者的图片一般比较清晰,分辨率较高,细节特征很突出,可以借助这些信息标注车辆的具体车型;而后者大部分图片不够清晰,分辨率较低,存在遮挡,光照条件也不尽相同。为了解决这一问题,尽可能地消除或降低 2 类图片的差异,本文实验在高置信局部特征提取的训练集中,加入了重识别训练集中标注车标扩散区域的部分图片,使得新的高置信局部特征训练集的车标扩散区域总数增加到了 93 个,图片总数为 44 024 张。车标扩散区域种类的具体分布可参见图 6,因为部分车标非常常见,对于这样的图片只选择 2 000 张加入训练集中,确保每种车标的扩散区域至少有一张图片。

训练过程中的每次迭代只训练 93 张图片,即每次迭代中每种车标都会被训练,迭代次数为 6 000 次,学习率初始值设置为 0.1,3 000 次迭代后学习率下降为 0.01,5 000 次迭代后学习率下降到 0.001。虽然迭代次数很多,但每次迭代训练的图片很少,所以全部的训练过程只需要 5~6 h 就可以完成。在测试集上的高置信局部特征准确率为 98.4%,这充分说明了局部特征识别结果的可信程度,为后续局部特征识别结果的应用创造了前提条件。

为了将车标扩散区域的识别结果应用到车辆重识别问题中,需要确定权重参数 λ 。因为 VehicleID 数据集有大量可用于测试的车辆图片,标准测试集只是抽取了一部分,所以可以选择一部分没有用到的图片作为验证集,来确定参数最优值。本文实验选取了 400 个车辆 ID,共计 3 136 张图片做验证集,并划分验证集的 query 集合和 gallery 集合。在验证集上的测试,权重参数 λ 从 1 开始,以 0.01 为跨度逐渐减小。实验结果如图 7 所示,横坐标为权重参数 λ ,纵坐标为验证集

在该条件下 top1 的识别准确率。
当权重参数 λ 刚开始下降时,top1 的识别准确率逐渐上升,在权重参数 λ 为 0.91 时达到峰值。此后,随着权重参数 λ 的下降,top1 的识别准确率略有下降。因此,本文实验中,权重参数 λ 的最优值取 0.91。在后续车辆重识别的实验中,也将权重参数 λ 设置为 0.91。

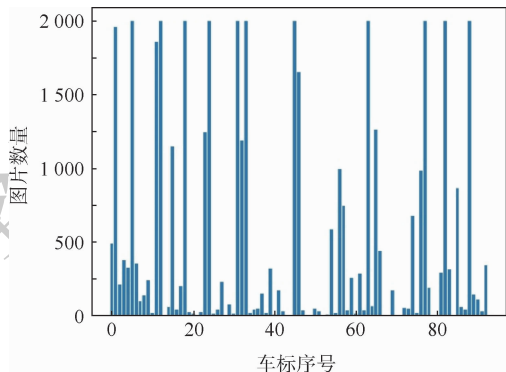


图 6 车标图片分布

Fig. 6 Image distribution of vehicle brands

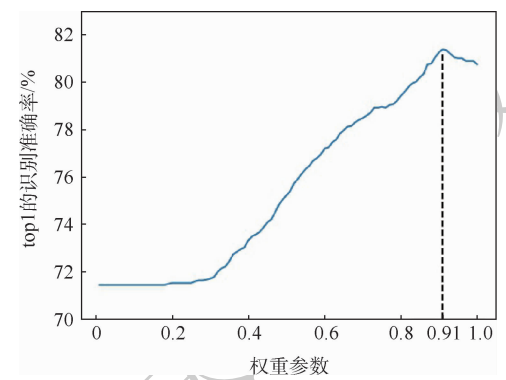


图 7 验证集上的实验结果

Fig. 7 Experimental results on validation set

3.5 实验结果与分析

本文算法与各算法在 VehicleID 数据集上的对比可参见表 2。OIFE 模型针对不同方向的车辆图片,能够将提取得到的特征转化为一个方向无关的向量,以尽可能消除或降低车辆方向差异带来的影响。NuFACT 模型能够根据车辆的不同方向来提取特征,同时为了保证模型运行的效率,还对特征向量进行一定程度的降维。VAMI 模型也重点关注图片中车辆的不同方向问题。C2F 模型设计了 coarse-to-fine 损失函数,将相同车辆的图片尽可能拉近,将不同车辆的图片拉远。RAM 模型从局部区域抽取特征,重要的局部区域会包含可以分辨相似车辆的重要细节。根据表 2 的对比结果显示,本文算法在 VehicleID 数据集的 top1 识别准确率上均好于其他算法。

本文算法模型以 Part-Regularized 网络模型为基础进行改进,使用车辆关键点检测算法来获取车

灯、车窗和车标等关键的局部区域。同时检测车标扩散区域,在重识别的测试阶段,将高置信局部特征预测的结果添加到向量距离计算的过程中,使得局部特征识别相同的车辆的向量距离更近。实验结果表明,在 VehicleID 的 3 个测试集中,top1 的识别准确率均有明显提升,分别提升了 0.8%、2.1% 和 1.0%,证明了本文算法模型的有效性。

单独使用一个分支的车辆重识别准确率结果可参见表 3。仅使用主分支,VehicleID 的 3 个测试集的重识别准确率分别为 78.4%、75.9% 和 74.4%。仅使用高置信局部特征分支,准确率分别为 24.1%、18.9% 和 15.5%。2 个分支都使用本文算法。结果显示,2 个分支均使用的情况能够获得最高的重识别准确率,并且比只使用一个分支均有明显提升。

表 2 不同算法 top1 的识别准确率对比结果

Table 2 Top1 recognition accuracy results of different methods

算法	top1 的识别准确率/%		
	small	medium	large
	测试集	测试集	测试集
VGG + CCL ^[17]	43.6	37.0	32.9
OIFE ^[19]			67.0
NuFACT ^[20]	48.9	43.6	38.6
VAMI ^[21]	63.1	52.9	47.3
C2F ^[22]	61.1	56.2	51.4
RAM ^[23]	75.2	72.3	67.7
Part-Regularized ^[16]	78.4	75.0	74.2
本文	79.2	77.1	75.2

表 3 top1 的识别准确率消融对比结果

Table 3 Ablation comparison results of top1 recognition accuracy

算法	top1 的识别准确率/%		
	small	medium	large
	测试集	测试集	测试集
仅主分支	78.4	75.9	74.4
仅高置信局部特征分支	24.1	18.9	15.5
两个分支(本文算法)	79.2	77.1	75.2

4 结 论

本文提出了一种基于高置信局部特征的车辆重识别优化算法,具有以下特点:

1) 具有提取高置信局部特征的分支,能够准确检测并识别车标扩散区域,识别准确率高达 98.4%。

2) 重点关注车辆的辅助局部信息,利用车辆关键点检测算法,检测车辆的车灯、车标和车窗区

域,以 Part-Regularized 网络为基础,利用局部信息来辅助训练全局特征。

3) 将高置信局部特征引入到图片间距离计算的过程中,通过设置验证集确定具体的权重参数。实验表明,本文算法的重识别 top1 的识别准确率有明显提升。

参考文献 (References)

[1] XUE G, ZHU H, HU Z, et al. Pothole in the dark: Perceiving pothole profiles with participatory urban vehicles [J]. IEEE Transactions on Mobile Computing, 2017, 16(5): 1408-1419.

[2] LYU F, FANG L, XUE G, et al. Large-scale full WiFi coverage: Deployment and management strategy based on user spatio-temporal association analytics [J]. IEEE Internet of Things Journal, 2019, 6(6): 9386-9398.

[3] LIU Y, SHENG H, ZHENG Y, et al. GDMN: Group decision-making network for person re-identification [J]. IEEE Access, 2018, 6: 64169-64181.

[4] LV K, DENG W, HOU Y, et al. Vehicle reidentification with the location and time stamp [C] // Proceedings of the IEEE International Conference on Computer Vision Workshops. Piscataway: IEEE Press, 2019: 399-406.

[5] ZHANG X, ZHANG R, CAO J, et al. Part-guided attention learning for vehicle re-identification [EB/OL]. (2020-02-17) [2020-03-01]. <https://arxiv.org/abs/1909.06023>.

[6] CHU R, SUN Y, LI Y, et al. Vehicle re-identification with viewpoint-aware metric learning [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 8282-8291.

[7] 李熙莹,周智豪,邱铭凯. 基于部件融合特征的车辆重识别算法 [J]. 计算机工程, 2019, 45(6): 12-20.

LI X Y, ZHOU Z H, QIU M K. Vehicle re-identification algorithm based on component fusion feature [J]. Computer Engineering, 2019, 45(6): 12-20 (in Chinese).

[8] 王盼盼,李玉惠,李福卫. 基于特征融合和度量学习的车辆重识别 [J]. 电子科技, 2018, 31(9): 29-31.

WANG P P, LI Y H, LI F W. Vehicle re-identification based on multi-features and metric learning [J]. Electronic Science and Technology, 2018, 31(9): 29-31 (in Chinese).

[9] 王盼盼,李玉惠. 基于特征融合和 L-M 算法的车辆重识别方法 [J]. 电子科技, 2018, 31(4): 12-15.

WANG P P, LI Y H. Vehicle re-identification method based on feature fusion and L-M algorithm [J]. Electronic Science and Technology, 2018, 31(4): 12-15 (in Chinese).

[10] SUPREM A, PU C. Looking GLAMORous: Vehicle Re-Id in heterogeneous cameras networks with global and local attention [EB/OL]. (2020-02-06) [2020-03-01]. <https://arxiv.org/abs/2002.02256>.

[11] ZAPLETAL D, HEROUT A. Vehicle re-identification for automatic video traffic surveillance [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2016: 25-31.

[12] KHORRAMSHAHI P, KUMAR A, PERI N, et al. A dual-path

model with adaptive attention for vehicle re-identification [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 6132-6141.

[13] KANACI A, LI M, GONG S, et al. Multi-task mutual learning for vehicle re-identification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE Press, 2019: 62-70.

[14] NEWELL A, YANG K, DENG J. Stacked Hourglass networks for human pose estimation [C] // European Conference on Computer Vision. Berlin: Springer, 2016: 483-499.

[15] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.

[16] HE B, LI J, ZHAO Y, et al. Part-Regularized near-duplicate vehicle re-identification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 3997-4005.

[17] LIU H, TIAN Y, YANG Y, et al. Deep relative distance learning: Tell the difference between similar vehicles [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2167-2175.

[18] ZHONG Z, ZHENG L, KANG G, et al. Random erasing data augmentation [EB/OL]. (2017-11-16) [2020-03-01]. <https://arxiv.org/abs/1708.04896>.

[19] WANG Z, TANG L, LIU X, et al. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 379-387.

[20] LIU X, LIU W, MEI T, et al. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance [J]. IEEE Transactions on Multimedia, 2017, 20 (3) : 645-658.

[21] ZHOU Y, SHAO L. Aware attentive multi-view inference for vehicle re-identification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6489-6498.

[22] GUO H, ZHAO C, LIU Z, et al. Learning coarse-to-fine structured feature embedding for vehicle re-identification [C] // Thirty-Second AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2018: 6853-6860.

[23] LIU X, ZHANG S, HUANG Q, et al. RAM: A region-aware deep model for vehicle re-identification [C] // 2018 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2018: 1-6.

作者简介:

窦鑫泽 男, 硕士研究生。主要研究方向: 计算机视觉。

盛浩 男, 博士, 副教授, 博士生导师。主要研究方向: 计算机视觉、模式识别和机器学习。

吕凯 男, 博士研究生。主要研究方向: 计算机视觉。

刘洋 男, 博士研究生。主要研究方向: 计算机视觉。

张洋 男, 博士研究生。主要研究方向: 计算机视觉。

吴玉彬 男, 博士研究生。主要研究方向: 计算机视觉。

柯韦 男, 博士, 副教授。主要研究方向: 模式识别。

Vehicle re-identification optimization algorithm based on high-confidence local features

DOU Xinze^{1,2}, SHENG Hao^{1,2,*}, LYU Kai^{1,2}, LIU Yang^{1,2}, ZHANG Yang^{1,2}, WU Yubin^{1,2}, KE Wei³

(1. State Key Laboratory of Software Development Environment, School of Computer Science and Engineering, Beihang University, Beijing 100083, China; 2. Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing 100083, China; 3. School of Applied Sciences, Macao Polytechnic Institute, Macao SAR 999078, China)

Abstract: In solving vehicle re-identification problems, different vehicle regions have different recognition degree of confidence. Based on this observation, we propose a vehicle re-identification optimization algorithm that takes advantage of the high-confidence local features. First, the vehicle key point detection algorithm is utilized to obtain the corresponding multiple key points' coordinate information of the vehicles, and to divide the vehicle brand extension regions and other prominent local regions. As the brand extension region is the most salient region, we propose to improve the degree of confidence of the local region in the testing phase. We also utilize a multi-layer convolutional neural network for processing the input images, cutting the convolutional features into several parts based on the obtained local regions, and acquiring feature tensors representing global and key regional information. Then, a fully connected layer is applied to combine the above features and output a one-dimensional vector for loss function calculating. In the testing phase, to reduce the target distances of vehicles with the same local identification, we propose to utilize the global features together with the high-confidence local features obtained by trained brand extension region extraction branch. Experiments on the widely used vehicle re-identification VehicleID dataset show that the proposed algorithm is effective.

Keywords: vehicle re-identification; high-confidence local features; feature optimization; region detection; neural network

Received: 2020-03-02; **Accepted:** 2020-03-20; **Published online:** 2020-04-24 08:57

URL: kns.cnki.net/kcms/detail/11.2625.V.20200423.1757.002.html

Foundation items: National Key R & D Program of China (2018YFB2100500); National Natural Science Foundation of China (61861166002, 61872025, 61635002); Science and Technology Development Fund, Macao SAR (0001/2018/AFJ); Open Fund of the State Key Laboratory of Software Development Environment (SKLSDE2019ZX-04)

* **Corresponding author.** E-mail: shenghao@buaa.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0080

基于 SM4 算法的白盒密码视频数据共享系统



吴震^{1,2,*}, 白健¹, 李大双³, 李斌^{1,2}, 曾兵^{1,2}, 张正强^{1,2}

1. 中国电子科技网络信息安全有限公司, 成都 610041; 2. 成都三零凯天通信实业有限公司, 成都 610041;
3. 中国电子科技第三十研究所, 成都 610041)

摘 要: 基于 SM4 算法的白盒密码视频数据共享系统是一种保障监控视频数据跨级跨域共享安全的系统,提出了一种基于国密 SM4 算法的白盒密码实现方式,并分析了算法的安全性,解决了 SM4 算法在非信任硬件环境中的安全运行问题。研制了基于后台权限控制机制的视频数据安全共享软件系统,包括共享数据上传/下载、共享审核、数据白盒加密处理、访问控制、基于白盒密码算法的共享视频解密播放器,实现了视频数据共享全过程的安全管控。搭建了实验环境,对所提系统进行了功能性能实验。实验结果表明,所提系统功能性能满足设计要求。

关 键 词: 白盒密码; SM4 算法; 加密; 视频数据; 安全共享

中图分类号: TN918.91

文献标志码: A **文章编号:** 1001-5965(2020)09-1660-10

在公共视频监控领域的纵向多级网络中,存储有海量视频数据,根据数据的敏感程度,一部分数据需要加密存储,一部分数据是明文存储,但均存在跨级或跨域共享需求。对于加密数据,由于受限于密钥管理系统的跨域访问限制及解密密钥的安全传输问题,需要先解密成明文后,才能共享。为保证原有加密或非密数据的共享安全,需要采用密码算法对跨级跨域共享数据进行完整性、机密性保护。传统密码学中,密码算法的设计都是建立在黑盒模型基础上的:假设密码算法的运行环境是安全的,攻击者只能访问算法的输入和输出,不能观察代码执行及动态加解密的过程,保障数据安全的关键是保护好密钥。

密码算法的实现主要有硬件和软件 2 种方式。采用硬件方式需使用嵌入了密码芯片的专用存储介质,密钥经保护后存储在芯片中,安全性高,在《公共安全视频监控联网信息安全技术要

求》(GB 35114—2017)^[1]中,对加密视频的导出及共享推荐使用这种方式,但该方法存在使用成本高、通用性较差等问题。采用软件方式共享方便快捷,但在软件加解密过程中,密钥将会出现在算法运行平台的内存中,由于不能确保密码算法软件运行环境的安全可信,攻击者可能使用静态分析工具、调试器,直接读取内存中的数据等方法找出密钥^[2],数据安全性得不到保障,也无法抵御近年来提出的时间分析、功耗分析、插入错误分析等旁路攻击^[3-6]。

对于这种不可靠的密码算法运行环境,2002 年,Chow 等^[7]将其命名为白盒攻击环境(White-Box Attack Context, WBAC),其描述如下:①在同一主机中,可以同时运行加密软件与拥有特殊权限的攻击软件,并且对密码算法在主机中的运行过程拥有完全的控制权;②主机中程序的动态执行过程可以被监视;③密码算法的实现原理及执

收稿日期: 2020-03-03; 录用日期: 2020-03-27; 网络出版时间: 2020-05-07 16:24
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200507.1149.003.html
基金项目: 国家重点研发计划(2016YFC0801004)
* 通信作者. E-mail: wuzhen@30kt.com

引用格式: 吴震, 白健, 李大双, 等. 基于 SM4 算法的白盒密码视频数据共享系统[J]. 北京航空航天大学学报, 2020, 46(9): 1660-1669. WU Z, BAI J, LI D S, et al. White-box cryptographic video data sharing system based on SM4 algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1660-1669 (in Chinese).

行过程是可见的,并且可被任意修改。

白盒攻击环境也可理解为白盒攻击模型,与黑盒攻击模型相比,白盒攻击模型可以认为是最坏的模型,然而其也是现今信息技术迅速发展很适用的模型,用于分析在不可信环境下运行的密码算法。为了保证密码算法在不可信的终端运行而不受到威胁,科研人员设计了在白盒攻击环境下能够确保安全性的密码算法,该密码算法被称为白盒密码算法,其目的就是为了应对白盒攻击环境,将密钥信息隐藏在密码算法中,防止密钥被攻击者获取。因此,研究安全高效的白盒密码算法具有非常重要的意义。

目前,对白盒密码算法的研究分为 2 种技术路线:①基于已有白盒密码算法,利用混淆、扰乱等白盒密码技术进行设计,在不改变原算法功能的前提下,保障算法在白盒攻击环境下的安全性;②研究一种全新的白盒密码算法,直接可以抵抗白盒攻击^[8]。Chow 等^[7]在提出白盒攻击环境的同时,通过将若干步骤组合起来所构成的函数用查找表来表示(即遍历函数的输入,存储函数的输出值构成查找表,从而可隐藏该函数的内部信息),将密钥嵌入到查找表并使用随机双射来保护查找表^[9],开创性地构造了首个白盒 AES 方案及首个白盒 DES 方案^[10]。2009 年, Xiao 和 Lai^[11]提出了 SMS4 白盒实现算法,使用查找表和仿射变换相结合的方式,从白盒多样性和白盒含混度的角度来看,该方案达到了应用所需的安全性,也针对 Billet 等^[12]提出的攻击方法说明了该方案的安全性。

本文基于国家密码管理局颁布的 SM4 算法设计白盒密码算法。SM4 (原 SMS4) 密码算法于 2006 年由国家密码管理办公室正式发布,是一种对称加密算法,2012 年成为密码行业标准 GM/T 0002—2012^[13],并于 2016 年成为国家标准 GB/T 32907—2016^[14],在 2017 年颁布的《公共安全视频监控联网信息安全技术要求》(GB 35114—2017)^[1]中,也将其采纳为用于视频数据加密保护的算法标准。

本文在进行算法设计时,参考了 Xiao 和 Lai^[11]提出的白盒算法实现思路,运用仿射变换与查找表来保护算法运行的内部信息,重新设计了编码解码流程,并引入播放口令、设备硬件因子等混淆参数增加算法分析难度。以此为基础进行工程化,研制开发视频数据跨级安全共享软件,并根据系统运行实际效率,对白盒密码算法及共享软件进行多次优化,实现跨级共享数据加解密速

率大于 1 Gbit/s 的目标,减少视频数据共享的加密成本,实现使用软件加密方式的视频数据安全快捷跨域共享。

1 基于 SM4 算法的白盒密码算法实现

SM4 算法为分组密码算法,算法的分组长度和密钥长度均为 128 bit。加密后的密文分组也是 128 bit,加密过程为 32 轮迭代操作及最后 1 轮输出的反序变换,每一轮的迭代操作称为轮函数^[15]。其解密过程类似于加密过程,但是轮密钥使用了与加密轮密钥相反的顺序。SM4 算法的整体结构如图 1 所示。图中:MK 为加密密钥,FK 为系统参数,CK 为固定参数,dk 为由 MK 扩展得到的轮密钥, F 为轮函数, F' 为密钥扩展函数。

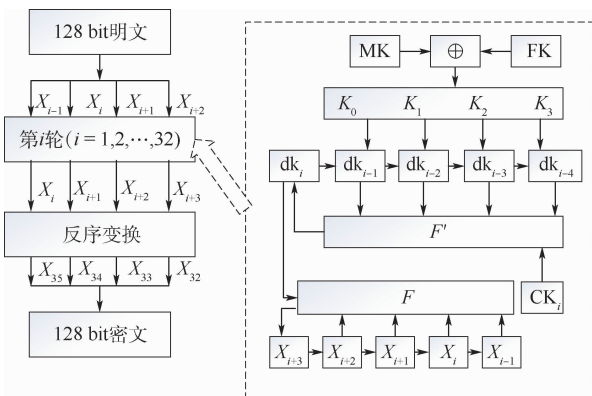


图 1 SM4 算法整体结构

Fig. 1 Overall structure of SM4 algorithm

设计基于 SM4 算法的白盒密码算法主要需要突破以下要点:

1) 可逆 S 盒变换。S 盒变换指按字节进行的代替变换,是作用在状态中每个字节上的一种非线性字节变换。S 盒变换又分为可逆 S 盒变换和不可逆 S 盒变换 2 种,本文设计的白盒密码算法的基本部件主要采用可逆 S 盒变换,以实现不同轮中间查表输出数据的隐藏。

2) 仿射变换。仿射变换是平面(空间)的一个可逆变化,由一个非奇异的线性变换和一个平移变换构成。在有限维的情况下,每个仿射变换可以由一个可逆矩阵 D 和一个向量 b 给出,对向量 x 的仿射变换为: $y = Dx + b$ 。

在本文的 SM4 白盒密码算法设计中,使用该方法结合可逆 S 盒变换实现相邻轮运算过程中的中间数据隐藏,防止攻击者直接获取加解密过程中的中间参数,继而分析破解获得相应的轮密钥,计算获得最终密钥。

根据 SM4 算法及加解密流程,本文基于查表

运算设计了 SM4 白盒密码算法。该白盒密码算法将 SM4 算法的每一轮运算分成 3 个部分,通过将第 2 部分和第 3 部分的运算转换为查表运算,并引入混淆参数,从而将原始的轮密钥隐藏进表中,即使攻击者在掌握整个软件之后,使用白盒攻击方法,也无法获得视频数据的加密密钥,有效保护了密钥在算法执行过程中的安全。

算法主要由 2 个部分组成:白盒密钥表生成模块和基于白盒密钥表的数据加解密模块。

1.1 白盒密钥表生成模块

白盒密钥表生成模块由绑定信息生成算法、随机数生成算法、白盒密钥表构造算法 3 部分组成,如图 2 所示。

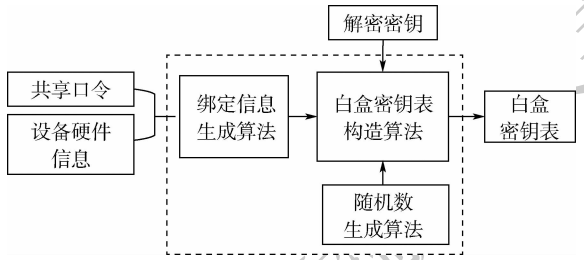


图 2 白盒密钥表生成模块构成

Fig.2 White-box key table generation module composition

绑定信息生成算法用于生成白盒密钥表构造算法的初始运行因子,该因子源于视频共享口令及播放设备硬件信息,目的是抵抗共享输出数据被非法浏览及拷贝风险,限制共享视频数据只能在指定播放设备上被知晓共享口令的人解密查看。输入因子必须为 256 bit 的数据,因此使用绑定信息生成算法将视频共享口令和指定视频播放设备的硬件信息进行融合。设共享口令为 Password,设备硬件信息为 Hi,使用 SM3 杂凑算法^[16]作为绑定信息生成算法,算法的输入为 Password 和 Hi,算法的输出结果为绑定信息即 Password||Hi 的哈希值。

随机数生成算法主要是生成白盒密钥表构造算法所需要的随机参数信息。需要共享的视频数据使用 SM4 白盒密码算法进行加密,而共享输出系统需要对视频密文进行解密,因此设计的白盒密码算法为 SM4 白盒解密算法,对应的白盒密钥表为 SM4 白盒解密表。随机数生成算法产生 SM4 白盒解密表所需要的有限域 $\mathbf{Z}_2$ 上的 420 个 8×8 的可逆矩阵。该算法的输入是随机数种子,输出为 $\mathbf{Z}_2$ 上的 8×8 的可逆矩阵 $\mathbf{M}_i, i=1,2,\dots,420$ 。

白盒密钥表构造算法根据实际的数据加密密钥生成解密所需的白盒密钥表,主要分为 3 个部分:①负责通过置换矩阵对输入数据进行混淆;

②通过遍历输入,结合轮密钥、置换矩阵、S 盒置换进行计算,将轮密钥嵌入表中;③对第②部分输出的结果进行解混淆,如果是最后一轮,则直接输出,否则加入新的混淆数据。

具体步骤如下:

步骤 1 定义 4 类矩阵,分别为 $\mathbf{P}, \mathbf{E}_i, \mathbf{Q}_j, \mathbf{R}_k$ 。令: $\mathbf{P} = \text{diag}(\mathbf{P}_1, \mathbf{P}_2, \mathbf{P}_3, \mathbf{P}_4)$, $\mathbf{E}_i = \text{diag}(\mathbf{E}_{i,1}, \mathbf{E}_{i,2}, \mathbf{E}_{i,3}, \mathbf{E}_{i,4})$, $\mathbf{Q}_j = \text{diag}(\mathbf{Q}_{j,1}, \mathbf{Q}_{j,2}, \mathbf{Q}_{j,3}, \mathbf{Q}_{j,4})$, $\mathbf{R}_k = \text{diag}(\mathbf{R}_{k,1}, \mathbf{R}_{k,2}, \mathbf{R}_{k,3}, \mathbf{R}_{k,4})$ 。

步骤 2 产生复合线性变化 $\mathbf{A}_{i,j}, i=1,2,\dots,36, j=1,2,3$ 。 $\mathbf{A}_{1,1} = \mathbf{E}_1 \mathbf{P}, \mathbf{A}_{1,2} = \mathbf{E}_1 \mathbf{P}, \mathbf{A}_{1,3} = \mathbf{E}_1 \mathbf{P}; \mathbf{A}_{2,1} = \mathbf{E}_2 \mathbf{P}, \mathbf{A}_{2,2} = \mathbf{E}_2 \mathbf{P}, \mathbf{A}_{2,3} = \mathbf{E}_2 \mathbf{R}_1^{-1}; \mathbf{A}_{3,1} = \mathbf{E}_3 \mathbf{P}, \mathbf{A}_{3,2} = \mathbf{E}_3 \mathbf{R}_1^{-1}, \mathbf{A}_{3,3} = \mathbf{E}_3 \mathbf{R}_2^{-1}; \mathbf{A}_{4,1} = \mathbf{E}_4 \mathbf{R}_1^{-1}, \mathbf{A}_{4,2} = \mathbf{E}_4 \mathbf{R}_2^{-1}, \mathbf{A}_{4,3} = \mathbf{E}_4 \mathbf{R}_3^{-1}; \mathbf{A}_{5,1} = \mathbf{E}_5 \mathbf{R}_2^{-1}, \mathbf{A}_{5,2} = \mathbf{E}_5 \mathbf{R}_3^{-1}, \mathbf{A}_{5,3} = \mathbf{E}_5 \mathbf{R}_4^{-1}; \dots; \mathbf{A}_{i+1,1} = \mathbf{E}_{i+1} \mathbf{R}_i^{-1}, \mathbf{A}_{i+1,2} = \mathbf{E}_{i+1} \mathbf{R}_{i-1}^{-1}, \mathbf{A}_{i+1,3} = \mathbf{E}_{i+1} \mathbf{R}_i^{-1}; \dots; \mathbf{A}_{33,1} = \mathbf{E}_{33} \mathbf{R}_{26}^{-1}, \mathbf{A}_{33,2} = \mathbf{E}_{33} \mathbf{R}_{27}^{-1}, \mathbf{A}_{33,3} = \mathbf{E}_{33} \mathbf{R}_{28}^{-1}; \mathbf{A}_{34,1} = \mathbf{E}_{34} \mathbf{R}_{27}^{-1}, \mathbf{A}_{34,2} = \mathbf{E}_{34} \mathbf{R}_{28}^{-1}, \mathbf{A}_{34,3} = \mathbf{E}_{34} \mathbf{R}_{29}^{-1}; \mathbf{A}_{35,1} = \mathbf{E}_{35} \mathbf{R}_{28}^{-1}, \mathbf{A}_{35,2} = \mathbf{E}_{35} \mathbf{R}_{29}^{-1}, \mathbf{A}_{35,3} = \mathbf{E}_{35} \mathbf{R}_{30}^{-1}; \mathbf{A}_{36,1} = \mathbf{E}_{36} \mathbf{R}_{29}^{-1}, \mathbf{A}_{36,2} = \mathbf{E}_{36} \mathbf{R}_{30}^{-1}, \mathbf{A}_{36,3} = \mathbf{E}_{36} \mathbf{R}_{31}^{-1}。$

步骤 3 产生第 1 类查找表 SM4Dec-Table¹ = {SM4Dec-Table¹_{i,j} | $i=1,2,\dots,36; j=1,2,3,4$ }。具体方法为:设绑定信息 $\mathbf{BI} = (\mathbf{BI}_1, \mathbf{BI}_2, \mathbf{BI}_3, \mathbf{BI}_4) \in (\mathbf{Z}_2^{32})^4$, $\mathbf{BI}_j = (\mathbf{BI}_{j,1}, \mathbf{BI}_{j,2}, \mathbf{BI}_{j,3}, \mathbf{BI}_{j,4}) \in (\mathbf{Z}_2^8)^4, j=1,2,3,4, \mathbf{BI}_i = 0, i>4$, 混淆参数 $\mathbf{r} = (\mathbf{r}_0, \mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3) \in (\mathbf{Z}_2^{32})^4$ 。设在第 i 轮,第 1 类查找表 SM4Dec-Table¹_i = {SM4Dec-Table¹_{i,1}, SM4Dec-Table¹_{i,2}, SM4Dec-Table¹_{i,3}, SM4Dec-Table¹_{i,4}} 的输入为 $\mathbf{y}_i = (\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, \mathbf{y}_{i,3}, \mathbf{y}_{i,4})^T$ 和 $\mathbf{BI}_i$, SM4Dec-Table¹_{i,j} 的构建方法如下:}}}}

① $\mathbf{y}'_i = \mathbf{y}_i \oplus \mathbf{BI}_i = (\mathbf{y}_{i,1} \oplus \mathbf{BI}_{i,1}, \mathbf{y}_{i,2} \oplus \mathbf{BI}_{i,2}, \mathbf{y}_{i,3} \oplus \mathbf{BI}_{i,3}, \mathbf{y}_{i,4} \oplus \mathbf{BI}_{i,4})^T$ 。

② $\mathbf{y}'_i$ 经过线性变换 $\mathbf{P}^{-1} \mathbf{E}_i^{-1}$ 作用,输出为

$$\mathbf{P}^{-1} \mathbf{E}_i^{-1}(\mathbf{y}'_i) = \begin{pmatrix} \mathbf{P}_1^{-1} \mathbf{E}_{i,1}^{-1} & & & \\ & \ddots & & \\ & & \mathbf{P}_4^{-1} \mathbf{E}_{i,4}^{-1} & \\ & & & \end{pmatrix} * \begin{pmatrix} \mathbf{y}'_{i,1} \\ \vdots \\ \mathbf{y}'_{i,4} \end{pmatrix} = \begin{pmatrix} \mathbf{P}_1^{-1} \mathbf{E}_{i,1}^{-1} \mathbf{y}'_{i,1} \\ \vdots \\ \mathbf{P}_4^{-1} \mathbf{E}_{i,4}^{-1} \mathbf{y}'_{i,4} \end{pmatrix}$$

③ 将轮密钥 $\mathbf{rk}_{36-i}$ 隐藏在 S 盒中,设 $\mathbf{P}^{-1} \mathbf{E}_i^{-1} \cdot (\mathbf{y}'_i)$ 经过 S 盒作用后的输出为

$$(\mathbf{z}_{i,1}, \mathbf{z}_{i,2}, \mathbf{z}_{i,3}, \mathbf{z}_{i,4}) = (\mathbf{S}(\mathbf{P}_1^{-1} \mathbf{E}_{i,1}^{-1} \mathbf{y}'_{i,1} \oplus \mathbf{rk}_{36-i,1}), \mathbf{S}(\mathbf{P}_2^{-1} \mathbf{E}_{i,2}^{-1} \mathbf{y}'_{i,2} \oplus \mathbf{rk}_{36-i,2}), \mathbf{S}(\mathbf{P}_3^{-1} \mathbf{E}_{i,3}^{-1} \mathbf{y}'_{i,3} \oplus \mathbf{rk}_{36-i,3}), \mathbf{S}(\mathbf{P}_4^{-1} \mathbf{E}_{i,4}^{-1} \mathbf{y}'_{i,4} \oplus \mathbf{rk}_{36-i,4}))$$

④ 计算 $\mathbf{Q}_i \cdot \mathbf{P} \cdot \mathbf{L} \cdot (\mathbf{z}_{i,1}, \mathbf{z}_{i,2}, \mathbf{z}_{i,3}, \mathbf{z}_{i,4})^T$ 。

记 $H_i = Q_i P L = (H_{i,1} \ H_{i,2} \ H_{i,3} \ H_{i,4})$, $H_{i,j}$ 为 32×8 的矩阵, 则

$$Q_i \cdot P \cdot L \cdot \begin{pmatrix} z_{i,1} \\ \vdots \\ z_{i,4} \end{pmatrix} = (H_{i,1} \ H_{i,2} \ H_{i,3} \ H_{i,4}) \begin{pmatrix} z_{i,1} \\ \vdots \\ z_{i,4} \end{pmatrix} =$$

$$(H_{i,1} \cdot z_{i,1}) \oplus (H_{i,2} \cdot z_{i,2}) \oplus (H_{i,3} \cdot z_{i,3}) \oplus (H_{i,4} \cdot z_{i,4}) = v_{i,1} \oplus v_{i,2} \oplus v_{i,3} \oplus v_{i,4}$$

据上, 可获得 4 个 8×32 bit 的查找表 SM4Dec-Table $^1_{i,j}$: $y_{i,j} (\rightarrow z_{i,j}) \rightarrow v_{i,j}$, $i = 1, 2, \dots, 36$, $j = 1, 2, 3, 4$ 。该查找表主要用于隐藏解密密钥和绑定信息。

步骤4 产生第2类查找表 SM4Dec-Table2 = $\{\text{SM4Dec-Table}^2_{i,j} | i = 1, 2, \dots, 36; j = 1, 2, 3, 4\}$ 。具体方法为: 设在第 i 轮中, 第2类查找表 SM4Dec-Table $^2_i = \{\text{SM4Dec-Table}^2_{i,1}, \text{SM4Dec-Table}^2_{i,2}, \text{SM4Dec-Table}^2_{i,3}, \text{SM4Dec-Table}^2_{i,4}\}$ 的输入为 $s_i = (s_{i,1}, s_{i,2}, s_{i,3}, s_{i,4})$ 和 $t_i = (t_{i,1}, t_{i,2}, t_{i,3}, t_{i,4})$ 。

对于第1轮~第4轮, 第2类查找表的构建方法如下:

① 计算 $R_i P(s_i)$ 和 $R_i Q_i^{-1}(t_i)$ 。

$$R_i P(s_i) = (R_{i,1} P_1 s_{i,1}, R_{i,2} P_2 s_{i,2}, R_{i,3} P_3 s_{i,3}, R_{i,4} P_4 s_{i,4})$$

$$R_i Q_i^{-1}(t_i) = (R_{i,1} Q_{i,1}^{-1} t_{i,1}, R_{i,2} Q_{i,2}^{-1} t_{i,2}, R_{i,3} Q_{i,3}^{-1} t_{i,3}, R_{i,4} Q_{i,4}^{-1} t_{i,4})$$

② 计算 $R_i P(s_i) \oplus R_i Q_i^{-1}(t_i) \oplus r_{i-1}$ 。

$$\begin{aligned} R_i P(s_i) \oplus R_i Q_i^{-1}(t_i) \oplus r_{i-1} = & (u_{i,1}, u_{i,2}, u_{i,3}, u_{i,4}) = \\ & (R_{i,1} P_1 s_{i,1} \oplus R_{i,1} Q_{i,1}^{-1} t_{i,1} \oplus r_{i-1,1}, \\ & R_{i,2} P_2 s_{i,2} \oplus R_{i,2} Q_{i,2}^{-1} t_{i,2} \oplus r_{i-1,2}, \\ & R_{i,3} P_3 s_{i,3} \oplus R_{i,3} Q_{i,3}^{-1} t_{i,3} \oplus r_{i-1,3}, \\ & R_{i,4} P_4 s_{i,4} \oplus R_{i,4} Q_{i,4}^{-1} t_{i,4} \oplus r_{i-1,4}) \end{aligned}$$

对于第5轮~第32轮, 第2类查找表的构建方法如下:

③ 计算 $R_i R_{i-4}^{-1}(s_i)$ 和 $R_i Q_i^{-1}(t_i)$ 。

$$R_i R_{i-4}^{-1}(s_i) = (R_{i,1} R_{i-4,1}^{-1} s_{i,1}, R_{i,2} R_{i-4,2}^{-1} s_{i,2}, R_{i,3} R_{i-4,3}^{-1} s_{i,3}, R_{i,4} R_{i-4,4}^{-1} s_{i,4})$$

$$R_i Q_i^{-1}(t_i) = (R_{i,1} Q_{i,1}^{-1} t_{i,1}, R_{i,2} Q_{i,2}^{-1} t_{i,2}, R_{i,3} Q_{i,3}^{-1} t_{i,3}, R_{i,4} Q_{i,4}^{-1} t_{i,4})$$

④ 计算 $R_i R_{i-4}^{-1}(s_i) \oplus R_i Q_i^{-1}(t_i)$ 。

$$\begin{aligned} R_i R_{i-4}^{-1}(s_i) \oplus R_i Q_i^{-1}(t_i) = & (R_{i,1} R_{i-4,1}^{-1} s_{i,1} \oplus R_{i,1} Q_{i,1}^{-1} t_{i,1}, \\ & R_{i,2} R_{i-4,2}^{-1} s_{i,2} \oplus R_{i,2} Q_{i,2}^{-1} t_{i,2}, R_{i,3} R_{i-4,3}^{-1} s_{i,3} \oplus \\ & R_{i,3} Q_{i,3}^{-1} t_{i,3}, R_{i,4} R_{i-4,4}^{-1} s_{i,4} \oplus R_{i,4} Q_{i,4}^{-1} t_{i,4}) = \\ & (u_{i,1}, u_{i,2}, u_{i,3}, u_{i,4}) \end{aligned}$$

对于第33轮~第36轮, 第2类查找表的构建方法如下:

$$\text{SM4Dec-Table}^2_{i,j} : (s_{i,j}, t_{i,j}) \rightarrow u_{i,j} = P_j^{-1} R_{i-8,j}^{-1} s_{i,j} \oplus P_j^{-1} Q_{i,j}^{-1} t_{i,j} \oplus r_{4,j} \quad i = 33, j = 1, 2, 3, 4$$

$$\text{SM4Dec-Table}^2_{i,j} : (s_{i,j}, t_{i,j}) \rightarrow u_{i,j} = P_j^{-1} R_{i-8,j}^{-1} s_{i,j} \oplus P_j^{-1} Q_{i,j}^{-1} t_{i,j} \oplus r_{3,j} \quad i = 34, j = 1, 2, 3, 4$$

$$\text{SM4Dec-Table}^2_{i,j} : (s_{i,j}, t_{i,j}) \rightarrow u_{i,j} = P_j^{-1} R_{i-8,j}^{-1} s_{i,j} \oplus P_j^{-1} Q_{i,j}^{-1} t_{i,j} \oplus r_{2,j} \quad i = 35, j = 1, 2, 3, 4$$

$$\text{SM4Dec-Table}^2_{i,j} : (s_{i,j}, t_{i,j}) \rightarrow u_{i,j} = P_j^{-1} R_{i-8,j}^{-1} s_{i,j} \oplus P_j^{-1} Q_{i,j}^{-1} t_{i,j} \oplus r_{1,j} \quad i = 36, j = 1, 2, 3, 4$$

该查找表主要用于隐藏随机产生信息。

1.2 基于白盒密钥表的数据加解密模块

基于白盒密钥表的数据加解密模块中的绑定信息生成算法与白盒密钥表构造算法中的绑定信息生成算法相同。共享视频接收者输入共享口令后, 绑定信息生成算法结合播放设备的硬件信息产生绑定信息 **BI**, **BI** 将作为 SM4 白盒解密算法的输入。白盒加解密算法框架如图3所示。

SM4 白盒解密算法的输入为使用 SM4 算法

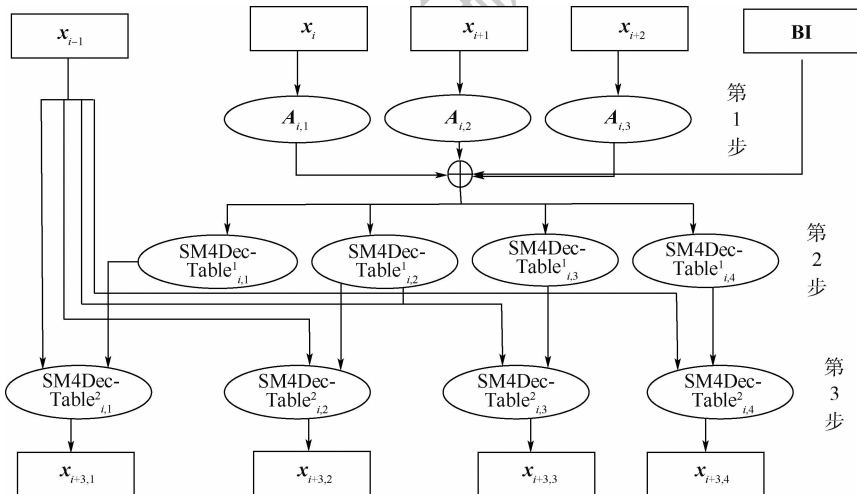


图3 白盒加解密算法框架

Fig.3 White-box encryption and decryption algorithm framework

加密的密文视频、白盒密钥表和绑定信息 $\mathbf{BI}$, 输出为明文视频, 其是一个 36 轮的分组迭代算法, 每一轮的具体步骤如下:

设第 i 轮的输入为 $(\mathbf{x}_{i-1}, \mathbf{x}_i, \mathbf{x}_{i+1}, \mathbf{x}_{i+2})$, 其中, $\mathbf{x}_{i-1}, \mathbf{x}_i, \mathbf{x}_{i+1}, \mathbf{x}_{i+2}$ 都为 32 bit。

1) 利用复合线性变换 $\mathbf{A}_{i,1}, \mathbf{A}_{i,2}, \mathbf{A}_{i,3}$ 分别左乘 $\mathbf{x}_i, \mathbf{x}_{i+1}, \mathbf{x}_{i+2}$, 然后将结果进行异或, 即

$$\mathbf{y}_i = (\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, \mathbf{y}_{i,3}, \mathbf{y}_{i,4})^T = \mathbf{A}_{i,1}\mathbf{x}_i \oplus \mathbf{A}_{i,2}\mathbf{x}_{i+1} \oplus \mathbf{A}_{i,3}\mathbf{x}_{i+2}$$

2) 计算 $\mathbf{y}_i = (\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, \mathbf{y}_{i,3}, \mathbf{y}_{i,4})^T$ 经过第 1 类查找表 $\text{SM4Dec-Table}_{i,j}^1$ 变换后的结果, 即

$$\mathbf{t}_{i,1} = \text{SM4Dec-Table}_{i,1}^1(\mathbf{y}_{i,1})$$

$$\mathbf{t}_{i,2} = \text{SM4Dec-Table}_{i,2}^1(\mathbf{y}_{i,2})$$

$$\mathbf{t}_{i,3} = \text{SM4Dec-Table}_{i,3}^1(\mathbf{y}_{i,3})$$

$$\mathbf{t}_{i,4} = \text{SM4Dec-Table}_{i,4}^1(\mathbf{y}_{i,4})$$

3) 计算 $\mathbf{s}_i = \mathbf{x}_{i-1} = (\mathbf{s}_{i,1}, \mathbf{s}_{i,2}, \mathbf{s}_{i,3}, \mathbf{s}_{i,4})^T$ 和 $\mathbf{t}_i = (\mathbf{t}_{i,1}, \mathbf{t}_{i,2}, \mathbf{t}_{i,3}, \mathbf{t}_{i,4})^T$ 经过第 2 类查找表 $\text{SM4Dec-Table}_{i,j}^2$ 变换后的结果, 即

$$\mathbf{x}_{i+3,1} = \text{SM4Dec-Table}_{i,1}^2(\mathbf{s}_{i,1}, \mathbf{t}_{i,1})$$

$$\mathbf{x}_{i+3,2} = \text{SM4Dec-Table}_{i,2}^2(\mathbf{s}_{i,2}, \mathbf{t}_{i,2})$$

$$\mathbf{x}_{i+3,3} = \text{SM4Dec-Table}_{i,3}^2(\mathbf{s}_{i,3}, \mathbf{t}_{i,3})$$

$$\mathbf{x}_{i+3,4} = \text{SM4Dec-Table}_{i,4}^2(\mathbf{s}_{i,4}, \mathbf{t}_{i,4})$$

经过上述 3 步变换后, 第 i 轮的输出为

$$\mathbf{x}_{i+3} = (\mathbf{x}_{i+3,1}, \mathbf{x}_{i+3,2}, \mathbf{x}_{i+3,3}, \mathbf{x}_{i+3,4})$$

1.3 SM4 白盒密码算法安全性分析

本节主要分析 SM4 白盒密码算法面对 BGE 攻击的安全性。BGE 是针对 Chow 等^[7] 设计的 AES 白盒算法提出的一种攻击方法。Chow 等^[7] 设计的 AES 白盒算法每一轮的实现方式为: $\text{Out}^r \cdot \text{AES}' \cdot \text{In}^r$, AES' 表示 AES 第 r 轮的变换, In^r 表示第 r 轮的输入编码, Out^r 表示第 r 轮的输出编码, 且 $\text{Out}^r = (\text{In}^{r+1})^{-1}$ 。

BGE 攻击的主要思想是: 先根据 AES 白盒算法的查找表把输入编码和输出编码恢复为仿射变换, 然后计算出仿射变换的输出编码。每一轮的输出编码和下一轮的输入编码是互逆的, 即

$\text{Out}^r = (\text{In}^{r+1})^{-1}$, 因此除了第 1 轮, 其他输入编码也可以计算出来。此时, 攻击者就可以得到 AES' , 从而得到隐藏着的密钥。Michiels 等^[17] 将其改进为一种通用攻击方法, 可以对类似算法的白盒实现进行攻击。

在本文设计的 SM4 白盒解密算法中, 如果将查找表与下一步的线性变换结合起来考虑, 将得到如下所示的合成变换, 其中 $\mathbf{Q}_i$ 与其逆正好相抵消, 如图 4 所示。

假设第 1 轮的查找表的输入为 $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4)$, $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \mathbf{t}_4)$, 经过合成变换后的输出为 $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4)$ 。由于 L 可以看做是 32×32 的矩阵, 所以可以将 L 分成 16 个 8×8 的子块:

$$L = \begin{pmatrix} L_{1,1} & L_{1,2} & L_{1,3} & L_{1,4} \\ L_{2,1} & L_{2,2} & L_{2,3} & L_{2,4} \\ L_{3,1} & L_{3,2} & L_{3,3} & L_{3,4} \\ L_{4,1} & L_{4,2} & L_{4,3} & L_{4,4} \end{pmatrix}$$

合成变换的输入 $\mathbf{y} = (\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4)$, $\mathbf{t} = (\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \mathbf{t}_4)$, 输出 $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4)$ 满足如下关系:

$$\mathbf{z}_r(\mathbf{y}, \mathbf{t}) = \mathbf{E}_{i+1,r} \mathbf{P}_r \left(\sum_{k=1}^4 L_{r,k} S(\mathbf{P}_k^{-1} \mathbf{E}_{i,k}^{-1}(\mathbf{y}_k)) \right) \oplus \mathbf{E}_{i+1,r} \mathbf{R}_{i-4,r}^{-1} \mathbf{t}_r$$

将输入 $\mathbf{y} = (\mathbf{y}_1, 0, 0, 0)$, $\mathbf{t} = (0, 0, 0, 0)$ 代入

上述关系:

$$\mathbf{z}_r(\mathbf{y}_1, 0, 0, 0, 0, 0, 0) = \mathbf{E}_{i+1,r} \mathbf{P}_r \mathbf{L}_{r,1} S(\mathbf{P}_1^{-1} \mathbf{E}_{i,1}^{-1}(\mathbf{y}_1)) \oplus \mathbf{c}_r$$

$$\mathbf{z}_j(\mathbf{y}_1, 0, 0, 0, 0, 0, 0) = \mathbf{E}_{i+1,j} \mathbf{P}_j \mathbf{L}_{j,1} S(\mathbf{P}_1^{-1} \mathbf{E}_{i,1}^{-1}(\mathbf{y}_1)) \oplus \mathbf{c}_r$$

式中:

$$\mathbf{c}_r = \mathbf{E}_{i+1,r} \mathbf{P}_r \left(\sum_{k=2}^4 L_{r,k} S(\mathbf{P}_k^{-1} \mathbf{E}_{i,k}^{-1}(0)) \right) \oplus \mathbf{E}_{i+1,r} \mathbf{R}_{i-4,r}^{-1}(0)$$

$$\mathbf{c}_j = \mathbf{E}_{i+1,j} \mathbf{P}_j \left(\sum_{k=2}^4 L_{j,k} S(\mathbf{P}_k^{-1} \mathbf{E}_{i,k}^{-1}(0)) \right) \oplus \mathbf{E}_{i+1,j} \mathbf{R}_{i-4,j}^{-1}(0)$$

则

$$\mathbf{z}_j(\mathbf{y}_1, 0, 0, 0, 0, 0, 0) = \mathbf{A}_{j,r} \mathbf{z}_r(\mathbf{y}_1, 0, 0, 0, 0, 0, 0) \oplus \mathbf{c}_{j,r}$$

$$\mathbf{A}_{j,r} = \mathbf{E}_{i+1,j} \mathbf{P}_j \mathbf{L}_{j,1} \mathbf{L}_{r,1}^{-1} \mathbf{P}_r^{-1} \mathbf{E}_{i+1,r}^{-1}$$

$$\mathbf{c}_{j,r} = \mathbf{A}_{j,r} \mathbf{c}_r \oplus \mathbf{c}_j$$

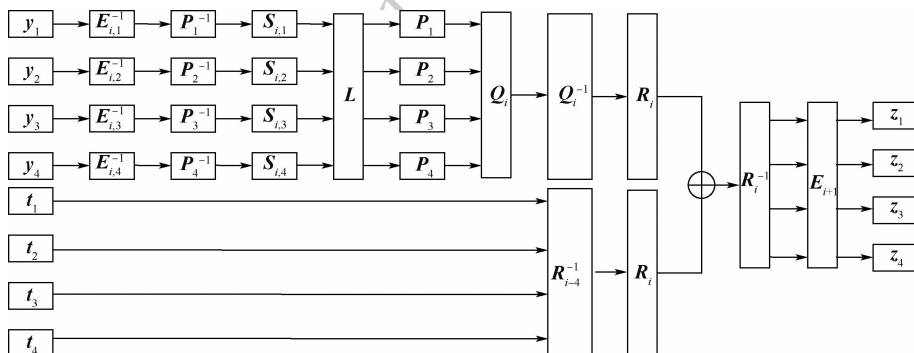


图 4 合成变换

Fig. 4 Synthetic transform

事实上, $A_{j,r}$ 可以看做是 8×8 的矩阵, 常数项 $c_{j,r}$ 有 8 个未知项, 共有 72 个未知项。对于一对 (z_j, z_r) , 将等式 $z_j(y_1, \mathbf{0}, \mathbf{0}, \mathbf{0}) = A_{j,r} z_r(y_1, \mathbf{0}, \mathbf{0}, \mathbf{0}) \oplus c_{j,r}$ 展开可获得 8 个线性方程, 则任取 9 个不同的 y_1 值, 可以组成 8 个含有 9 个未知数的方程组 (每个方程组有 9 个方程)。因此, 能以低于 8×9^3 的时间复杂度求出 $A_{j,r}$ 和 $c_{j,r}$ 。

同理可得, $z_r(\mathbf{0}, y_2, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}) = A'_{r,j} z_j(\mathbf{0}, y_2, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}) \oplus c'_{r,j}$, 其中, $A'_{r,j} = E_{i+1,r} P_r L_{r,2} L_{j,2}^{-1} \cdot P_j^{-1} E_{i+1,j}^{-1}$, $c'_{r,j} = A'_{r,j} c'_j \oplus c'_r$ 。那么, $A_{j,r} \cdot A'_{r,j} = (P_j^{-1} E_{i+1,j}^{-1})^{-1} L_{j,1} L_{r,1} L_{r,2} L_{j,2}^{-1} (P_j^{-1} E_{i+1,j}^{-1})$ 。

因为 $A_{j,r}$ 和 $A'_{r,j}$ 可以求出, 而各分块矩阵 $L_{i,j}$ 均已知, 因此可记为

$$XA = BX$$

式中: $A = A_{j,r} \cdot A'_{r,j}$, $B = L_{j,1} L_{r,1}^{-1} L_{r,2} L_{j,2}^{-1}$, 均为 8×8 的已知矩阵; $X = P_j^{-1} E_{i+1,j}^{-1}$ 为 8×8 的未知矩阵。因此, 方程可获得关于 X 的 64 个未知数的 64 个方程, X 能以低于 $4 \times 8^{2.4} + 64^3$ 的时间复杂度求出, 其中 $4 \times 8^{2.4}$ 为 4 个 8×8 矩阵相乘的复杂度, 64^3 为求解 64 阶方程组的复杂度。

经过 4 次相同的计算, 可以得到 $P_1^{-1} E_{i+1,1}^{-1}$, $P_2^{-1} E_{i+1,2}^{-1}$, $P_3^{-1} E_{i+1,3}^{-1}$, $P_4^{-1} E_{i+1,4}^{-1}$ 。

求出 $A_{j,r}$ 和 $A'_{r,j}$ 的时间复杂度低于 $2 \times 8 \times 9^3$, 因此能以低于 $4 \times (2 \times 8 \times 9^3 + 4 \times 8^{2.4} + 64^3) < 2^{20}$ 的复杂度计算出 $P^{-1} E_{i+1}^{-1} = \text{diag}(P_1^{-1} E_{i+1,1}^{-1}, P_2^{-1} E_{i+1,2}^{-1}, P_3^{-1} E_{i+1,3}^{-1}, P_4^{-1} E_{i+1,4}^{-1})$ 。

同理, 可计算出 $P^{-1} E_{i+2}^{-1}$ 。又因为线性变换 $E_{i+2} R_{i+1}^{-1}$ 已知, 故 $P^{-1} R_{i+1}^{-1}$ 可以求出。若能推算出 $R_{i+1} Q_{i+1}^{-1}$, 则 $P^{-1} Q_{i+1}^{-1}$ 是可计算的。若 S 盒的输出混淆 $P^{-1} E_{i+2}^{-1}$ 和输出混淆 $Q_{i+1} P$ 已知, 则密钥可推导出来。

然而, $R_{i+1} Q_{i+1}^{-1}$ 隐藏在第 2 张表中, 是未知的, 故 SM4 白盒加解密算法可以抵抗 BGE 攻击。

2 视频数据共享系统软件设计及实现

视频数据共享系统由基于 B/S 架构的服务端系统及白盒视频播放器组成, 系统提供共享视频文件审批、白盒加密共享数据输出、加密视频解密播放功能。共享视频文件审批完成对视频文件共享业务的审批, 并管理审批日志; 白盒加密共享数据输出服务将共享的视频文件转换为经 SM4 白盒算法加密保护的, 并提供下载功能, 实现视频数据的安全输出。共享视频文件审批服务可支持多层组织机构进行协同审批; 白盒加密共享数据输出系统的数据加密性能可达到 1 Gbit/s。

通过对原有视频文件播放器进行改造, 集成 SM4 白盒算法解密模块形成白盒视频文件播放器。

2.1 模块设计

服务端系统由用户身份管理模块、关键参数管理模块、视频安全共享处理模块、输出视频管理模块、商用基本密码处理模块、SM4 白盒密码处理模块组成。模块之间的关系如图 5 所示。

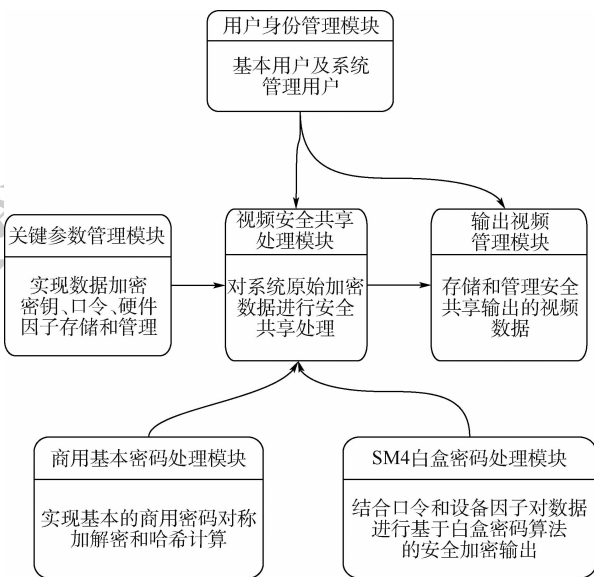


图 5 软件模块关系

Fig. 5 Software module relationship

1) 用户身份管理模块: 实现系统用户信息的管理, 包括用户注册、用户登录、用户权限管理及用户信息修改。

2) 关键参数管理模块: 对用户提交的视频数据的数据加密密钥、视频文件共享口令、硬件因子进行存储及管理。

3) 视频安全共享处理模块: 对待共享的视频文件进行审批处理, 并调用商用基本密码处理模块及 SM4 白盒密码处理模块提供的接口, 实现对共享视频文件的白盒加密处理。

4) 商用基本密码处理模块: 使用商用密码算法对待共享的加密敏感视频文件进行解密, 得到明文视频文件, 如待共享的视频文件已经是明文, 则直接返回; 同时提供 SM3 算法接口, 可实现对口令及硬件设备因子的哈希计算。

5) SM4 白盒密码处理模块: 结合共享口令和设备硬件因子对明文视频数据使用 SM4 白盒密码算法加密, 输出白盒密文视频文件。

6) 输出视频管理模块: 对白盒密文视频文件进行存储和管理, 并提供文件下载功能。

2.2 软件执行流程设计

2.2.1 视频共享

将待共享的视频文件名、共享者下载需要的

口令密码、分享对象的设备硬件因子信息、分享原因等信息上传至服务器,保存至数据库。

2.2.2 视频审核及白盒加密处理

分享者提出视频分享请求后,后台审核数据库会新增一条审核信息,审核者根据分享者提供的信息决定是否通过分享请求,分享请求审核通过后,系统向视频存储服务器请求相应视频文件,如该文件已被加密,则同时向该加密共享文件所属的密钥管理服务器请求解密所需要的密钥数据,如果视频库中没有相应视频文件信息,共享终止。

视频安全共享输出模块获得共享视频文件及密钥数据后,先使用原密钥解密视频文件,再使用SM4白盒密码算法对明文视频数据进行加密处理,生成白盒加密视频文件,以及对应的白盒加密信息表,最后将其打包存放于服务器并生成下载链接。具体流程如图6所示。

2.2.3 视频下载

用户提交正确的口令,从服务器下载已打包的白盒加密共享视频文件。如果用户提供的口令不正确,则不会下载文件。

2.2.4 白盒视频解密播放

用户将下载成功的白盒加密视频文件通过调用白盒解密库(需要输入口令)提供的接口输入,白盒解密库先计算播放端设备硬件因子及口令的哈希值,将其与白盒加密视频文件中保存的值对比,如正确则进行视频解密操作,解密成功后进行播放显示,否则返回口令或硬件因子不匹配的错误信息。

3 实验与性能评估

实验系统由一台视频共享输出服务器,两台访问客户端PC组成。服务器硬件配置CPU:Xeon(R) CPU E5-2630 V4@ 2.20 GHz;硬盘:2 TB;内存:32 GB。PC机硬件配置:CPU:Intel(R) Core(TM) i7-4710MQ CPU @ 2.50 GHz,内存:8 GB,硬盘:1 TB,硬件设备通过局域网连接。软件环境为:Windows Server 2008 R2 Standard 操作系统,Windows 7 SP1 操作系统,GoogleChrome 浏览器。实验环境连接示意图如图7所示。

实验内容为:视频数据文件共享审核有效性;视频数据共享加密口令及硬件设备因子绑定效果;使用SM4白盒算法对共享视频数据文件加密;被共享者下载经白盒算法加密的共享视频文件;白盒算法加密的共享视频文件安全播放;视频白盒软件数据加密输出速率性能评估及与标准SM4算法加密输出速率的比较。

系统软件安装部署后,默认存在一个系统管理员用户,由管理员在系统内分别创建1个视频共享操作者,一个视频分享对象。视频共享操作者选择需要进行共享的加密视频文件,并填写共享对象的名字、共享文件播放设备的标识信息、共享密码等信息,信息填写完毕后,进行提交。提交成功后,由系统管理员用户进行审核,审核通过后,后台自动执行共享文件的解密及使用SM4白盒算法进行加密,白盒加密成功后的共享视频文件自动保存在视频共享服务器中。此时,视频分享对象可登录视频共享服务器,查看到共享白盒视频文件列表,输入正确的共享密码后,分享对象下载共享文件,并在指定的共享文件播放设备中使用白盒视频播放器正常进行解密播放。

实验操作流程如图8所示。

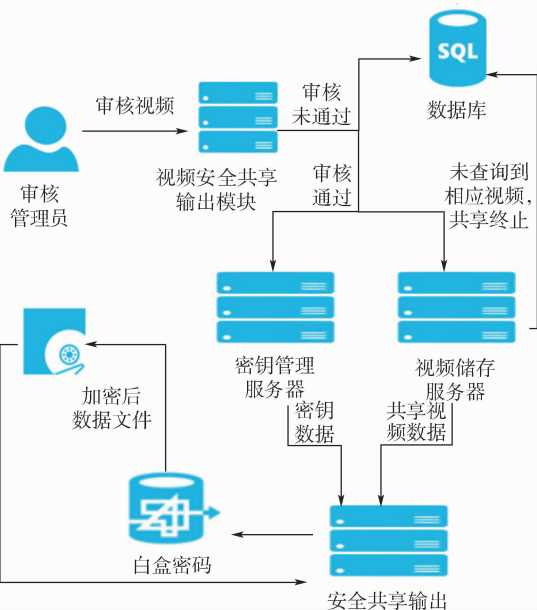


图6 视频审核流程
Fig.6 Video audit flowchart

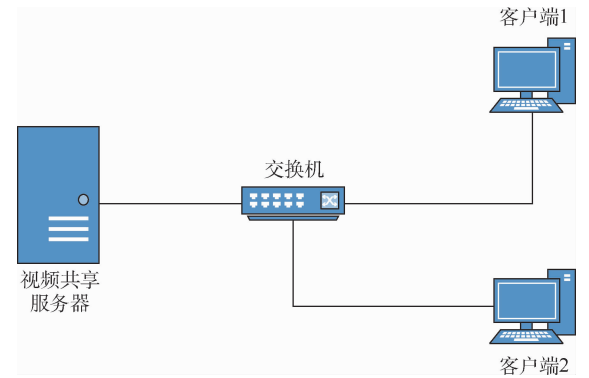


图7 实验设备部署图
Fig.7 Experimental equipment deployment diagram

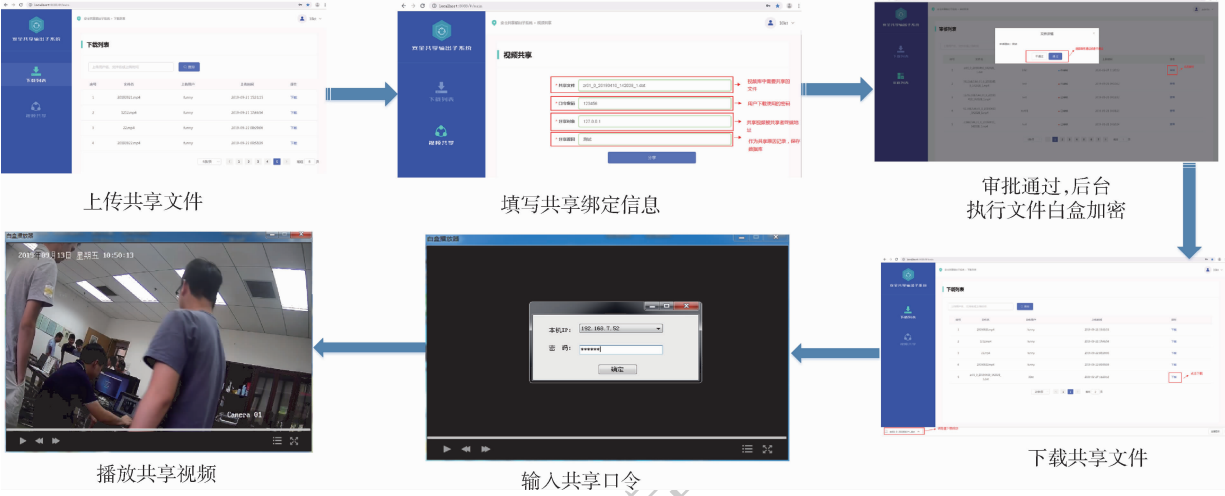


图 8 视频数据共享系统功能实验操作流程

Fig. 8 Experimental operation process of video data sharing system function

上述操作流程表明,视频数据共享系统能够使用 SM4 白盒算法实现加密视频数据文件的共享,经白盒算法处理后的视频共享文件只能在指定硬件设备播放,同时需要输入正确的共享口令,功能符合系统设计。

SM4 白盒算法的性能测试:编写脚本,在视频共享服务器端执行 64 个线程进行数据加密,每个线程加密数据量为 16 Mbit,即每次传输数据量为 $64 \times 16 \text{ Mbit} = 1 \text{ Gbit}$,执行 3 次。线程执行并输出加密所需的时间,如图 9 所示。

根据测试结果,白盒加密算法速率计算如表 1 所示。

SM4 算法的性能测试:编写程序,在视频共享服务器端执行 64 个线程对数据加密,每个线程加密数据量为 16 Mbit,即每次传输数据量为 $64 \times$

16 Mbit = 1 Gbit,执行 3 次,记录程序执行输出加密时间及加密速率,得到如表 2 所示 SM4 加密算法速率实验值。

表 2 SM4 加密算法速率实验值

Table 2 Experimental value of SM4 encryption algorithm rate

编号	加密数据量/Gbit	加密时间/s	加密速率/(Gbit · s ⁻¹)
1	1	0.840	1.190
2	1	0.825	1.212
3	1	0.836	1.196

- 根据表 1 及表 2 的性能实验结果可知:
- 1) 与标准 SM4 算法相比,SM4 白盒加密算法的速率有 10% 左右的降低,主要是生成白盒密钥表增加了时间消耗。
 - 2) 本文设计的 SM4 白盒加密算法速率平均在 1 Gbit/s 以上,性能满足设计要求。

4 结 论

- 1) 本文提出了一种基于 SM4 的白盒密码算法,实现了对称加密算法执行过程中的密钥隐藏。解决了非信任硬件环境下,软件加解密算法的密钥安全问题,保障了数据的安全性。
- 2) 对本文算法的安全性进行了理论分析,分析结果表明,该算法能够抵御 BGE 攻击。
- 3) 在算法基础上,设计并实现了基于白盒密码算法的视频数据共享系统,该系统能够实现共享视频文件的上传、审核、白盒加密处理、下载、安全解密播放。
- 4) 通过实验验证了系统的功能及性能。

参考文献 (References)

[1] 中华人民共和国公安部. 公共安全视频监控联网信息安全



图 9 白盒加密性能测试结果输出

Fig. 9 White-box encryption performance test result output

表 1 白盒加密算法速率实验值

Table 1 Experimental value of white-box encryption algorithm rate

编号	加密数据量/Gbit	加密时间/s	加密速率/(Gbit · s ⁻¹)
1	1	0.923	1.08
2	1	0.943	1.06
3	1	0.901	1.10

技术要求:GB 35114—2017[S]. 北京:中国标准出版社, 2017:4-5.

Ministry of Public Security of the PRC. Technical requirements for information security of video surveillance network system for public security:GB 35114—2017[S]. Beijing:China Standard Press,2017:4-5(in Chinese).

[2] COLLBERG C,NAGRA J. 软件加密与解密[M]. 崔孝晨, 译. 北京:人民邮电出版社,2012:5-6.

COLLBERG C,NAGRA J. Surreptitious software:Obfuscation, watermarking, and tamperproofing for software protection[M]. CUI X C, translated. Beijing:Posts & Telecom Press,2012:5-6 (in Chinese).

[3] ANDERSON R,KUHN M. Low cost attacks on tamper resistant devices[C]// Security Protocols 1997. Berlin:Springer,1997: 125-136.

[4] BIHAM E,SHAMIR A. Differential fault analysis of secret key cryptosystems[C]//17th Annual International Cryptology Conference Advances in Cryptology. Berlin:Springer,1997:513-525.

[5] BIHAM E,SHAMIR A. Power analysis of the key scheduling of the AES candidates[C]// 2nd AES Candidate Conference, 1999:22-23.

[6] BONEH D,DEMILLO R A,LIPTON R J. On the importance of eliminating errors in cryptographic computations[J]. Journal of Cryptology,2001,14(2):101-119.

[7] CHOW S,EISEN P,JOHNSON H, et al. White-box cryptography and an AES implementation[C]// Cryptography—SAC 2002. Berlin:Springer,2002:250-270.

[8] 周洁,白健. 基于余数系统的 SM2 签名算法的白盒实现[C]// 中国密码学会 2017 年会,2017:71-73.

ZHOU J,BAI J. White box implementation of the SM2 signature algorithm based on the remainder system[C]// Chinese Association for Cryptologic Research 2017 Annual Meeting,2017:71-73(in Chinese).

[9] 潘文伦,秦体红,贾音,等. 对两个 SM4 白盒方案的分析 [J]. 密码学报,2018,5(6):651-670.

PAN W L,QIN T H,JIA Y, et al. Analysis of two SM4 white box scenarios[J]. Journal of Cryptologic Research,2018,5 (6):651-670(in Chinese).

[10] CHOW S,EISEN P,JOHNSON H, et al. A white-box DES implementation for DRM applications[C]// Digital Rights Management—DRM 2002. Berlin:Springer,2003:1-15.

[11] XIAO Y Y,LAI X J. White-box cryptography and implementation of SMS4[C]// Proceedings of the 2009 CACR Annual Meeting. Beijing:Science Press,2009:24-34.

[12] BILLET O,GILBERT H,ECH-CHATBI C. Cryptanalysis of a white box AES implementation[C]// Selected Areas in Cryptography—SAC 2004. Berlin:Springer,2005:227-240.

[13] 国家密码管理局. SM4 分组密码算法:GM/T 0002—2012 [S]. 北京:中国标准出版社,2012.

State Cryptography Administration Office of Security Commercial Code Administration. SM4 block cipher algorithm:GM/T 0002—2012[S]. Beijing:China Standard Press,2012(in Chinese).

[14] 国家密码管理局. 信息安全技术 SM4 分组密码算法:GB/T 32907—2016[S]. 北京:中国标准出版社,2016.

State Cryptography Administration Office of Security Commercial Code Administration. Information security technology – SM4 block cipher algorithm:GB/T 32907—2016[S]. Beijing:China Standard Press,2016(in Chinese).

[15] 吕述望,苏波展,王鹏,等. SM4 分组密码算法综述[J]. 信息安全研究,2016(11):995-1007.

LU S W,SU B Z,WANG P, et al. Overview on SM4 algorithm [J]. Journal of Information Security Research,2016(11):995-1007(in Chinese).

[16] 国家密码管理局. SM3 密码杂凑算法:GM/T 0004—2012 [S]. 北京:中国标准出版社,2012.

State Cryptography Administration Office of Security Commercial Code Administration. SM3 cryptographic hash algorithm:GM/T 0004—2012[S]. Beijing:China Standard Press,2012 (in Chinese).

[17] MICHIELS W,GORISSEN P,HOLLMANN H D L. Cryptanalysis of a generic class of white-box implementations[C]// The Selected Areas in Cryptography. Berlin:Springer,2009:414-428.

作者简介:

吴震 女,硕士,高级工程师。主要研究方向:多媒体通信、网络信息安全技术。

白健 男,硕士,高级工程师。主要研究方向:密码算法、区块链技术。

李大双 男,博士,研究员。主要研究方向:密码算法、网络通信技术。

李斌 男,硕士,高级工程师。主要研究方向:网络信息安全技术、人工智能技术。

曾兵 男,硕士,高级工程师。主要研究方向:密码技术、信息安全技术。

张正强 男,工程师。主要研究方向:网络信息安全。

White-box cryptographic video data sharing system based on SM4 algorithm

WU Zhen^{1,2,*}, BAI Jian¹, LI Dashuang³, LI Bin^{1,2}, ZENG Bing^{1,2}, ZHANG Zhengqiang^{1,2}

(1. China Electronics Technology Cyber Security Co., Ltd., Chengdu 610041, China;

2. Chengdu Sanling Kaitian Communication Industry Co., Ltd., Chengdu 610041, China;

3. The 30th Research Institute of CETC, Chengdu 610041, China)

Abstract: White-box cryptographic video data sharing system based on SM4 algorithm is a system that guarantees the security of cross-level and cross-domain sharing of surveillance video data. In this paper, we propose a white-box cipher implementation method based on SM4 cryptographic algorithm, and analyze the security of the algorithm, solves the problem how the SM4 algorithm can compute safely in untrusted hardware environments. In addition, we developed a video data security sharing software system based on background permission control mechanism, including shared video data upload/download, sharing audit, data white-box encryption, access control functions, and shared video decryption player based on white-box cryptographic algorithm, realizing the security management and control of the entire process of video data sharing. Then we set up the experimental environment, and the system's functional performance experiment was performed. The experimental results show that the system's functional performance meets the design requirements.

Keywords: white-box cryptography; SM4 algorithm; encryption; video data; secure sharing

Received: 2020-03-03; **Accepted:** 2020-03-27; **Published online:** 2020-05-07 16:24

URL: kns.cnki.net/kcms/detail/11.2625.V.20200507.1149.003.html

Foundation item: National Key R & D Program of China (2016YFC0801004)

* **Corresponding author.** E-mail: wuzhen@30kt.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0076

基于 PF_RING 的多核视频流高性能传输模型



李欣^{1,2}, 樊志杰^{2,3,4,*}, 曹志威^{3,4}, 胡正梁^{3,4}, 陈国梁^{3,4}

(1. 中国人民公安大学 信息技术与网络安全学院, 北京 100038;
2. 安全防范技术与风险评估公安部重点实验室, 北京 100038;
3. 公安部第三研究所 信息安全技术部, 上海 200031; 4. 上海辰锐信息科技有限公司, 上海 200031)

摘 要: 在视频流高性能传输任务中, 视频流量的传输问题已成为一个研究的热点。为此, 提出了一种基于 PF_RING 技术的模型。通过使用 PF_RING + TNAPI 技术, 并结合内存路由表、多核、多队列多线程等相关技术确保视频流的高性能传输。同时为了保证不同网域间共享视频数据的安全传输, 提出了一种面向视频控制信令双向物理通道、视频流单向物理通道的视频传输模型。实验结果表明: 所提方法在有效吞吐量、CPU 使用率和平均误码率方面均取得了 10% 以上的提升。因此, 所提方法通过采用 PF_RING + TNAPI 技术, 同时结合视频传输模型, 确保了共享视频数据传输的安全性和高效性。

关 键 词: PF_RING + TNAPI; 内存路由表; 视频控制信令; 视频传输模型; 多核多通道
中图分类号: TP399
文献标志码: A **文章编号:** 1001-5965(2020)09-1670-07

随着国家大力推动“平安城市”“雪亮工程”的建设, 全国各地建设了大量的高清视频监控设备。同时, 伴随着视频传输网上视频流量的不断增长, 对视频网络流量的传输能力要求也越来越高, 并且视频流的传输是进行视频数据共享和视频数据分析的关键步骤。

现有的视频传输网流量传输模型主要有 2 种^[1-2]: ①基于硬件的方式, 包括 NetFlow 和 Net-Stream 等, 但该方式需要对网络设备进行专业的定制, 成本高, 且需要采用采样的方式抽取视频流量, 这样会造成分析结果不准确。②基于网络抓包嗅探的方式, 包括 Libpcap 和 Netmap 等, 但该方法往往只能够承载百兆级别的网络流量, 对于大规模的视频监控网络流量无法进行有效的传输、采集和分析。现阶段, PF_RING 技术在数据包抓取方面有其独特的优势, 并且已存在较多的

研究。例如, 刘峰^[3]提出 CPU 零参与方式, 提高了千兆网卡数据包的抓取效率。赵志强等^[4]通过零拷贝技术, 将 PF_RING 抓取数据包的方式集成到 Snort 中, 使 IDS 具备了 IPv6 检测的功能。同时, 考虑到视频传输网中的视频流数据具有实时性高、数据吞吐量大、数据误码率低等特性, 这就要求视频传输系统具有高性能视频流传输与分析的能力。

鉴于此, 本文采用 PF_RING 多核视频流高性能传输模型进行视频监控系统前端视频流数据包的处理, 以便减少视频流数据包的拷贝次数, 进而提高对视频传输网中视频流量传输、转发的速度和效率。

1 相关技术

PF_RING 的多核视频流高性能传输技术可以

收稿日期: 2020-03-03; 录用日期: 2020-04-24; 网络出版时间: 2020-05-11 10:23
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200509.1603.001.html
基金项目: 国家重点研发计划 (2018YFC0807105)
* 通信作者: E-mail: aaronzfan@126.com

引用格式: 李欣, 樊志杰, 曹志威, 等. 基于 PF_RING 的多核视频流高性能传输模型[J]. 北京航空航天大学学报, 2020, 46(9): 1670-1676. LI X, FAN Z J, CAO Z W, et al. High-performance multi-core video stream transmission model based on PF_RING [J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1670-1676 (in Chinese).

提升视频传输系统的传输性能^[5-9]。以下将对本文涉及的 PF_RING 机理和视频传输系统进行概述。

1.1 PF_RING 机理

PF_RING 是 Deri^[10] 研究出来的运行在 Linux 内核层的高效数据报文传输模型,其有助于对网络数据报文和网络流量进行深层次的分析研究。同时,用户级别的应用程序可以通过其提供的 API 来进行更为高效的网络数据传输。

在 Linux 传统的网络协议栈数据报文处理过程中,CPU 大量的处理能力使用在把网卡接收到的数据报文拷贝到 kernel 中的数据结构队列中,再从 kernel 的数据结构队列发送到用户空间的应用程序数据接收区中,导致了 CPU 处理能力的严重耗损。网络传输过程中,在网络设备上需要对数据报文 sk_buff 结构进行多次拷贝,并且涉及内核态和用户态的频繁系统调用,限制了传输报文的效率,尤其对小报文的传输性能影响更大。本

文引入的 PF_RING 技术,一方面减少了报文在网络设备上的拷贝次数,极大地释放出 CPU 的处理能力;另一方面与 TNAPI 技术的融合将网卡接收的数据报文分发到各个 CPU 核的 RX 队列中以充分利用多核的特性。通过这 2 种方式显著提高了网络设备的数据报文处理能力。

1.2 视频传输系统

视频传输系统是一种应用于不同视频传输网之间视频资源共享的通道^[11-12],其采用全新硬件设备和全新软件架构,由视频设备认证服务器、视频用户认证服务器和网络安全隔离设备 3 部分组成,是运行在各视频传输网之间的核心视频数据传输部件。本文提出的视频传输系统采用软硬一体化设计,主要用于解决用户在 2 个视频传输网域之间相互访问视频资源时所面临的视频资源可信性、用户合法性、数据安全性、访问可控性等问题。视频传输系统部署架构如图 1 所示。

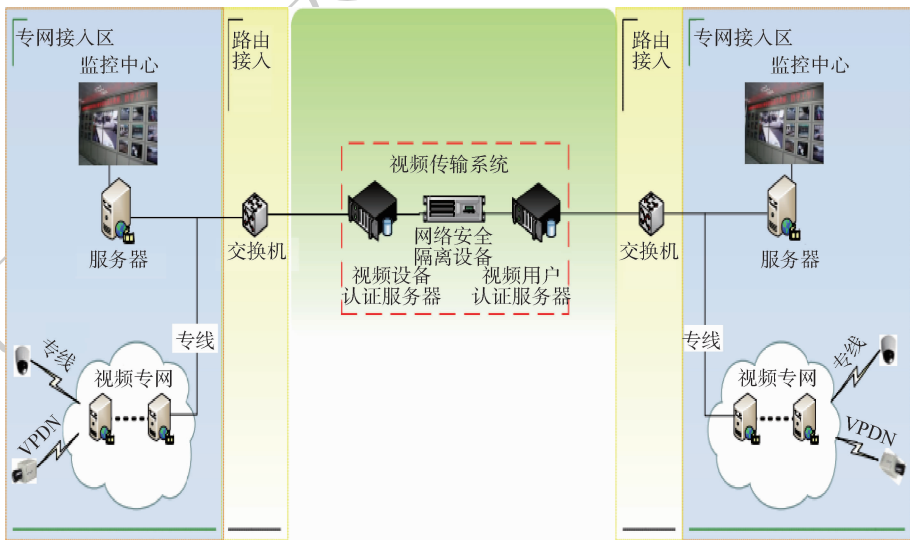


图 1 视频传输系统部署架构

Fig.1 Deployment framework of video transmission system

由于视频传输网的使用用户多,视频数据量大,在不同的视频传输网之间进行视频数据共享,就需要视频传输系统具有高性能的视频流传输机制。因此,视频传输系统要满足误码率低、延时小、吞吐量高的特性。本文研究使用基于 PF_RING 多核模型的高效数据包传输技术,进而提升视频传输系统接收视频流的能力,降低视频流的误码率,减少视频流在视频传输系统内部的延时,提高系统的网络数据吞吐量。

2 基于 PF_RING 多核模型的视频传输系统

本节详细分析了基于 PF_RING 多核模型的

共享视频流传输过程。

2.1 PF_RING 模型

PF_RING 的工作机制比 Linux 传统的网络协议栈处理网络数据报文更高效^[13-15],图 2 为两者工作机制的比较。

Linux 系统传统的处理网络数据报文过程为:网卡驱动将数据放入网卡的缓冲区中;驱动程序发送中断请求给 CPU,将网卡缓冲区中数据拷贝到 Linux 系统的内核网络协议栈缓冲区中;内核网络协议栈的 Ethernet 层、IP 层、TCP/UDP 层、Socket 层依次对数据报文进行处理;通过 Linux 的系统调用将数据从内核拷贝到用户空间。

PF_RING 是一种与 Linux 系统 Socket 不同的

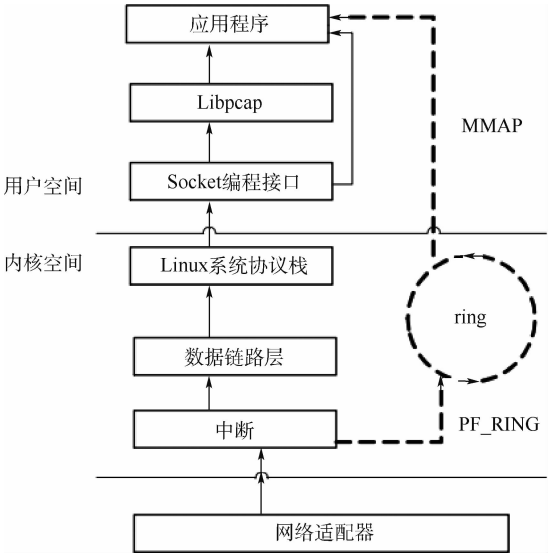


图2 传统协议栈与 PF_RING 工作机制的比较
Fig.2 Comparison of working mechanism between traditional protocol stack and PF_RING

新的网络 Socket,其可以极大地改进数据报文传输的速度,减少报文传输过程中在 Linux 系统内核态和用户态之间的拷贝次数。PF_RING 对数据传输方式优化的不同又分为 PF_RING + TNAPI 和 PF_RING 零拷贝 2 种实现机制。本文分别采用上述 2 种模式来提升视频传输系统的传输效率,并进行了对比分析。

随着多核 CPU 在网络设备上的广泛使用,Deri^[10] 在 NAPI (New Application Program Interface)基础上提出 TNAPI (Threaded New Application Program Interface),将网卡接收的数据报文分发到各个 CPU 核的 RX 队列中以充分利用多核的特性。PF_RING + TNAPI 的实现机制如图 3 所示,具体步骤如下:

步骤 1 创建对应多核的环形缓冲区 ring 的新型套接字 Virtual PF_RING。

步骤 2 TNAPI 将网卡上收到的数据分发到各核对应的 RX 队列中。

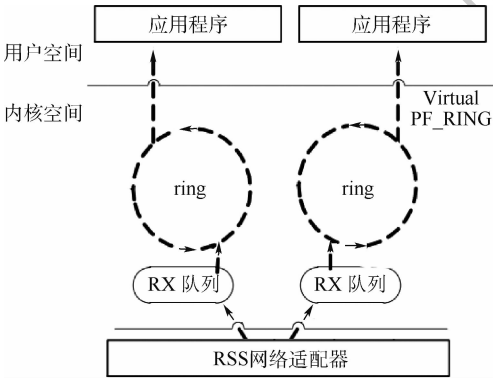


图3 PF_RING + TNAPI 的实现机制

Fig.3 Implementation mechanism of PF_RING + TNAPI

步骤 3 将 RX 队列中的数据报文拷贝到环形缓冲区 ring,若缓冲区 ring 已满,则将该报文丢弃。

步骤 4 在各核上运行用户空间应用程序,通过内存映射文件 MMAP 或 PF_RING 轮询直接访问各个环形缓冲区 ring 中的数据。

PF_RING 直接映射网卡内核和寄存器到用户空间的机制即 PF_RING 零拷贝模式,其实现机制如图 4 所示,具体步骤如下:

步骤 1 PF_RING 设置映射网卡内存 ring 到用户态的零拷贝方式。

步骤 2 仅有一次数据拷贝操作由网卡的网络处理单元完成。

步骤 3 应用程序使用 PF_RING 接口以 DMA 或应用轮询方式获取网卡环形缓冲区 ring 中的数据。

步骤 4 CPU 的处理能力能更多地分配到数据报文处理中,处理网络数据报文更高效。

PF_RING 零拷贝模式的缺点在于一次只允许一个应用程序访问 DMA 缓冲,不利于利用 CPU 多核的特性。

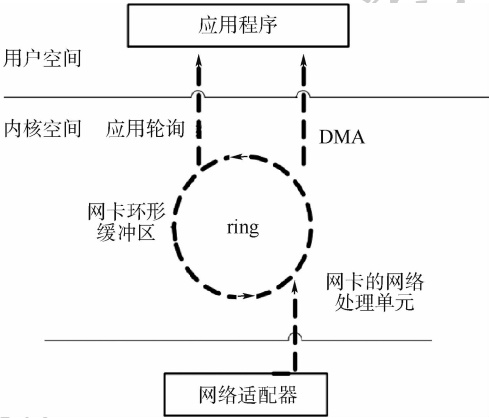


图4 PF_RING 零拷贝的实现机制

Fig.4 Implementation mechanism of PF_RING zero-copy

2.2 视频传输模型

视频传输系统的安全隔离设备是由 1 个视频控制信令双向物理通道、2 个正反向的视频流单向物理通道组成的专用设备^[16-18]。视频设备认证服务器主要用于对相互进行数据共享的视频传输网视频服务器等设备的身份认证。视频用户认证服务器主要用于对相互进行数据共享的视频传输网中的使用用户进行身份认证。同时,视频设备认证服务器、视频用户认证服务器需要对视频传输网间的视频流进行高效的传输。

视频传输系统为了实现视频传输网间的视频流高效的传输,使用了 PF_RING 技术结合内存路

由表、多通道多线程相关技术。

视频传输系统的内存路由表包含视频控制信令的内存路由表、由视频控制信令动态分配的视
频流内存路由表。其中,视频传输系统的控制信
令的内存路由表各字段如下:

- 1) 源地址、端口。指被调取视频源所在视频
传输网服务器的 IP 和端口。
- 2) 目标地址、端口。指视频接收方所在视频
传输网服务器的 IP 和端口。
- 3) 协议类型。UDP 或 TCP 协议。
- 4) 信令类型。指视频信令所采用的协议,如
SIP、RTSP 等协议。

视频流内存路由表是由上文中的视频控制信
令动态分配得到的内存路由表,且其是以视频传
输系统在设备上开辟的端口为 Key,以 DataNet-
PortVsas 结构体为 Value 的 map 内存表。视频传
输系统每个通道对应一个视频流内存路由表,系
统通道数不超过系统硬件 CPU 的总核数。Net-
PortVsas 为视频设备认证服务器、视频用户认证
服务器的端口。DataNetPortVsas 的数据结构如
表 1 所示。

2.3 基于 PF_RING 多核技术的视频流传输

视频传输系统的高效视频流传输过程如图 5
所示。

视频传输系统需要对 2 个专网间的视频资源
进行数据共享。视频用户认证服务器和视频设备

表 1 DataNetPortVsas 的数据结构

Table 1 Data structure of DataNetPortVsas		
属性名	数据类型	属性描述
timelast	time_t	数据报文时间戳
proxyType	IPProxyType	应用层协议类型
netSrcIPPort	NetIPPort	码流的视频源 IP 地址和端口
setNetDstIPPort	std::set<NetIPPort>	码流的视频目标 IP 地址和端口集合

认证服务器都具备视频码流的接收和转发能力。
如图 5 所示,SnifferThread 为视频传输系统的视
频流接收服务。通过反复实验发现,在现有的 8 核
CPU 硬件设备下,使用多通道能满足视频传输系
统高效视频流传输的要求。其中,瓶颈在视频传
输系统的 SnifferThread 接收视频码流能力。另
外,SnifferThread 线程接收视频码流的过程如图 6
所示。

鉴于此,本文分别使用 PF_RING 零拷贝机制
和 PF_RING + TNAPI 机制对 SnifferThread 服务进
行优化。PF_RING 零拷贝机制基本没有任何额
外的数据包复制工作,进而将 CPU 的处理能力更
多地运用在视频流数据报文的处理和传输功能
上。PF_RING + TNAPI 机制,视频传输系统 Sniff-
erThread 服务使用多线程多队列方式从网卡获取
数据报文,效率较高,可以更好地发挥硬件多核的
特性,显著降低了 CPU 的使用率。

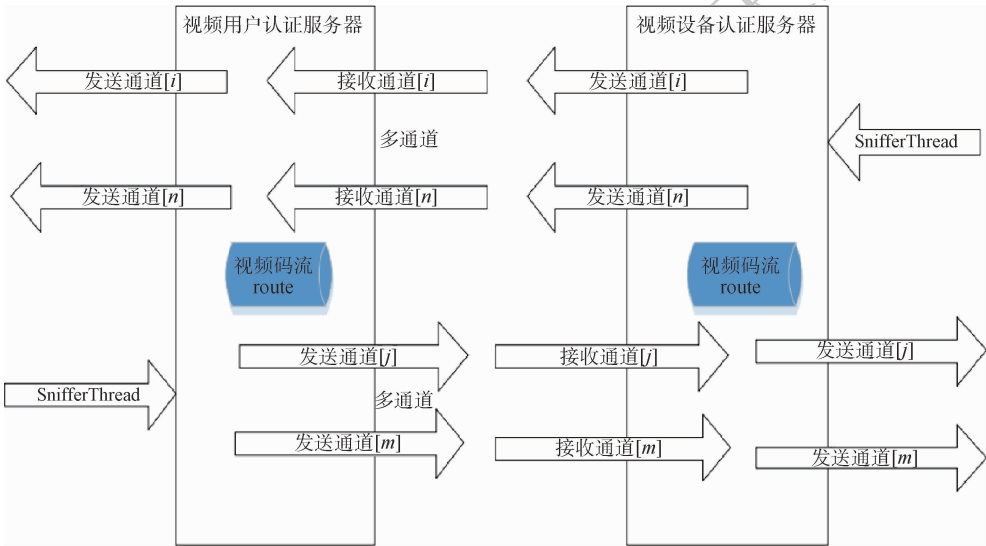


图 5 视频传输系统的视频流传输

Fig. 5 Video stream transmission of video transmission system

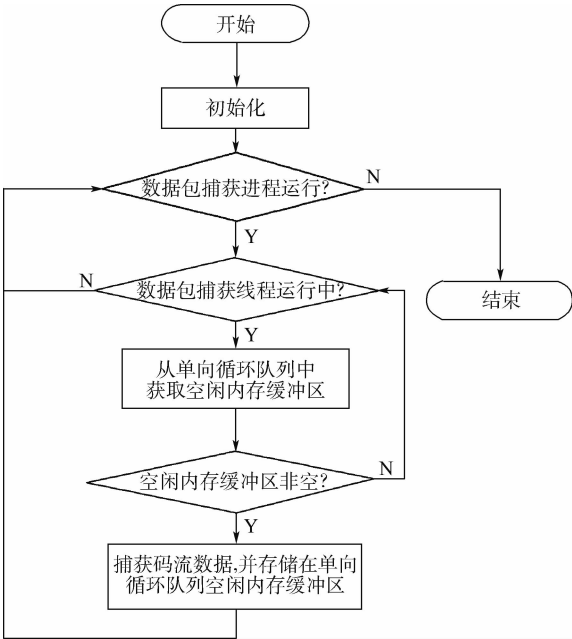


图6 SnifferThread 服务框架

Fig.6 Service framework of SnifferThread

3 实验

3.1 实验环境

为了将传统协议栈、PF_RING + TNAPI、PF_RING 零拷贝技术对视频传输系统的传输性能进行对比分析,搭建了如图 7 所示的测试环境。

测试环境中,硬件设备采用 Intel 的 E5-2623 v3@ 3.00 GHz 型号 CPU,32 GB 内存,英特尔® 82599 双口光口 10 GB 网卡;操作系统采用 CentOS 6.5。

本节实验的主要目的是验证 PF_RING 技术对视频传输性能的提高程度。鉴于此,实验中的每条测试用例将在 Linux 网络协议栈、PF_RING + TNAPI 和 PF_RING 零拷贝 3 种模式下进行对比分析,着重比较在相同软硬件环境下系统的有效吞吐量、CPU 使用率和平均误码率。

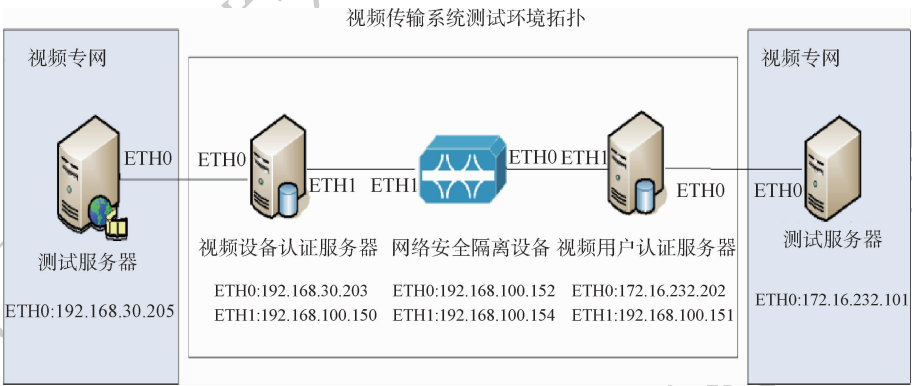


图7 测试环境拓扑

Fig.7 Topology of test environment

使用 iperf 工具发送 10 Mbit/s 的模拟视频流数据,测试用例如下:①第 1 条。400 路,每路 10 Mbit/s 的视频流。②第 2 条。500 路,每路 10 Mbit/s 的视频流。③第 3 条。600 路,每路 10 Mbit/s 的视频流。④第 4 条。700 路,每路 10 Mbit/s 的视频流。⑤第 5 条。800 路,每路 10 Mbit/s 的视频流。

3.2 实验结果与分析

实验从视频传输系统的有效吞吐量、CPU 使用率、平均误码率 3 个方面观察基于 Linux 网络协议栈、PF_RING + TNAPI 和 PF_RING 零拷贝 3 种模式下的视频传输能力,对比结果如图 8 ~ 图 10 所示,其中每项结果均为 5 次测试实验的平均值。

在上述 3 种模式下,当模拟发送 400 × 10 Mbit/s、500 × 10 Mbit/s 数据时,系统的有效吞吐量几乎相同;当模拟发送 700 × 10 Mbit/s、800 ×

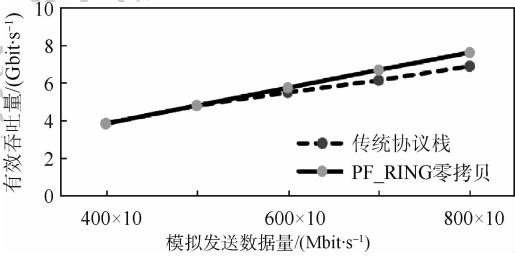


图8 有效吞吐量对比

Fig.8 Comparision of effective throughput

10 Mbit/s 时,PF_RING + TNAPI、PF_RING 零拷贝模式具有很大的优势。其中,PF_RING + TNAPI、PF_RING 零拷贝模式下有效吞吐量数据相近。

在接收视频流设备的 CPU 使用率对比中(见图 9),不管模拟发送的数据包有多大,PF_RING + TNAPI、PF_RING 零拷贝模式通常具有更低的 CPU 使用率。其中,PF_RING + TNAPI 能更好地发挥硬件多核的特性,在 CPU 使用率方面具备优势。

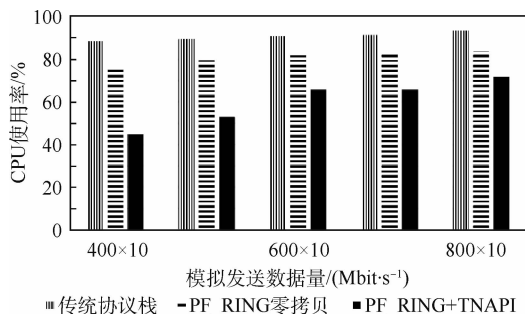


图9 CPU使用率对比

Fig.9 Comparision of CPU utilization

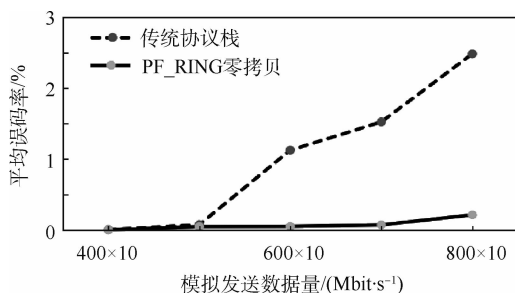


图10 平均误码率对比

Fig.10 Comparision of average bit error rate

图10显示了平均误码率对比分析,在基于Linux网络协议栈模式下,当模拟数据达到 600×10 Mbit/s时,平均误码率超过1%,将导致视频流失真较大。然而,PF_RING+TNAPI、PF_RING零拷贝模式通常具有更低的平均误码率,优势更大。其中,PF_RING+TNAPI、PF_RING零拷贝模式下平均误码率数据相近。

4 结 论

考虑到视频流量传输的特点,本文提出一种基于PF_RING的视频流高性能传输模型,其通过使用PF_RING+TNAPI机制结合视频传输系统多通道技术,进而减少视频流数据包在网络设备上的拷贝次数,更好发挥硬件多核的特性,极大地释放出CPU的处理能力。本文得出如下结论:

1) 通过合理地调整CPU的调度能力,本文模型有效解决了由于视频流波动导致的不同进程间相互影响造成的丢包率升高的问题。

2) 为了保证不同网域间视频流数据的安全传输,本文提出了视频控制信令双向传输、视频流单向传输的设计模式。

3) 通过大量的对比实验,本文验证了基于PF_RING+TNAPI的模型相比于传统的基于协议栈的模型在视频流传输方面的网络有效吞吐量、CPU使用率、平均误码率指标上均有很强的优势,在CPU使用率上也优于PF_RING零拷贝模式。

参考文献 (References)

- [1] 令瑞林,李峻峰,李丹,等.基于多核平台的高速网络流量实时捕获方法[J].计算机研究与发展,2017,54(6):1300-1313.
LING R L, LI J F, LI D, et al. Realtime capture of high-speed traffic on multi-core platform[J]. Journal of Computer Research and Development, 2017, 54(6): 1300-1313 (in Chinese).
- [2] 刘宝辰.高性能数据包捕获系统的研究与实现[D].上海:上海交通大学,2013:5-37.
LIU B C. Research and implementation of a high-performance packet capture system[D]. Shanghai: Shanghai Jiao Tong University, 2013: 5-37 (in Chinese).
- [3] 刘峰.利用PF_Ring零拷贝优化Spark Streaming网络性能[D].南宁:广西大学,2017:10-30.
LIU F. The optimization of network performance of Spark Streaming by using the PF_Ring zero_copy technology[D]. Nanning: Guangxi University, 2017: 10-30 (in Chinese).
- [4] 赵志强,易秀双,刘浩,等.基于IPv6的PF_RING ZC数据包捕获技术在IDS中的应用[J].通信学报,2018,39(1):1-6.
ZHAO Z Q, YI X S, LIU H, et al. Application of PF_RING ZC packet capture technology based on IPv6 in IDS[J]. Journal on Communications, 2018, 39(1): 1-6 (in Chinese).
- [5] 吴克河,王冬冬.基于PF_RING的高速网络数据捕获方法[J].计算机与数字工程,2019,47(3):598-604.
WU K H, WANG D D. Data capture method of high speed network based on PF_RING[J]. Computer and Digital Engineering, 2019, 47(3): 598-604 (in Chinese).
- [6] 刘小威,陈蜀宇,卢尧,等.零拷贝技术在网络分析工具中的应用[J].计算机系统应用,2012,21(4):169-173.
LIU X W, CHEN S Y, LU Y, et al. Application of zero-copy technology to network analysis system[J]. Computer Systems and Applications, 2012, 21(4): 169-173 (in Chinese).
- [7] WANG H P, HE D Z, WANG H, et al. Comparison of high-performance packet processing frameworks on NUMA[C]//2016 7th IEEE International Conference on Software Engineering and Service Science(ICSESS). Piscataway: IEEE Press, 2016: 54-58.
- [8] USUI T, KITATSUJI Y, YOKOTA H. A study on traffic management cooperating with IMS in MPLS networks[J]. Telecommunication Systems, 2013, 52(2): 671-680.
- [9] SAMOSIR J, INDRAWAN S M, HAGHIGHI P D. An evaluation of data stream processing systems for data driven applications[J]. Procedia Computer Science, 2016, 80(1): 439-449.
- [10] DERI L. High-speed dynamic packet filtering[J]. Journal of Network and Systems Management, 2007, 15(3): 401-415.
- [11] DENG R, LIU G Z. QoE driven cross-layer scheme for DASH-based scalable video transmission over LTE[J]. Multimedia Tools and Applications, 2018, 77(6): 6445-6469.
- [12] 刘文敏,张冬梅,李舒辰,等.基于Netmap的网络监听技术研究[J].信息安全,2013(11):45-48.
LIU W M, ZHANG D M, LI S C, et al. The network monitoring technology based on Netmap[J]. Information Network Security, 2013(11): 45-48 (in Chinese).

[13] 尚珍. 基于 NIDS 系统下 Snort 源码分析与数据结构研究 [J]. 网络安全技术与应用, 2017(8):64-65.
SHANG Z. Research on Snort source code analysis and data structure based on NIDS[J]. Network Security Technology and Application, 2017(8):64-65 (in Chinese).

[14] TYLMAN W. Native support for Modbus RTU protocol in Snort intrusion detection system[J]. Advances in Intelligent Systems and Computing, 2013, 224:479-487.

[15] 符程成, 赵荣彩, 邵铭, 等. 高负载网络下线速包捕获接口的设计与实现 [J]. 计算机工程与设计, 2010, 31(10):2203-2205.
GOU C C, ZHAO R C, TAI M, et al. Wire-speed packet capture interface design and implementation in high-load network[J]. Computer Engineering and Design, 2010, 31(10):2203-2205 (in Chinese).

[16] CHEN S, YANG J, RAN Y, et al. Adaptive layer switching algorithm based on buffer under-flow probability for scalable video streaming over wireless networks[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(6):1146-1160.

[17] 谢文韬. 基于无锁结构的大容量数据高性能检索系统研究 [D]. 南京: 东南大学, 2017:5-25.
XIE W T. Large capacity data high-performance retrieval system based on lock free structure[D]. Nanjing: Southeast University, 2017:5-25 (in Chinese).

[18] 尚彪. 网络数据包捕获技术的比较研究与实现 [D]. 武汉: 华中科技大学, 2014:1-26.
SHANG B. Comparative study and implementation of network packet capture technology[D]. Wuhan: Huazhong University of Science and Technology, 2014:1-26 (in Chinese).

作者简介:
李欣 男, 博士, 副教授。主要研究方向: 网络安全、智能视频处理等。
樊志杰 男, 博士, 助理研究员。主要研究方向: 网络安全、大数据安全、物联网安全、视频安全等。

High-performance multi-core video stream transmission model based on PF_RING

LI Xin^{1,2}, FAN Zhijie^{2,3,4,*}, CAO Zhiwei^{3,4}, HU Zhengliang^{3,4}, CHEN Guoliang^{3,4}

- (1. College of Information Technology and Network Security, People's Public Security University of China, Beijing 100038, China;
2. Key Laboratory of Security Prevention Technology and Risk Assessment of the Ministry of Public Security, Beijing 100038, China;
3. Department of Information Security Technology, The Third Research Institute of Ministry of Public Security, Shanghai 200031, China;
4. Shanghai Chenrui Information Technology Company, Shanghai 200031, China)

Abstract: The problem of video traffic transmission has become a research focus in the task of high-performance video stream transmission. In this paper, a model based on PF_RING is proposed to solve this problem. In this model, PF_RING + TNAPI technology, memory routing table, multi-core, multi-queue multi-threading and other related technologies are used to ensure the high-performance transmission of video stream. At the same time, in order to ensure the safe transmission of shared video data among different network domains, a video transmission model for video control signaling dual channel and video data single channel is proposed. The experimental results show that the performance of the proposed method has been increased by at least 10% in effective throughput, CPU utilization, and average bit error rate. Therefore, by fusing PF_RING + TNAPI technology and video transmission system model, the proposed method ensures the security and efficiency of shared video data transmission.

Keywords: PF_RING + TNAPI; memory routing table; video control signaling; video transmission model; multi-core and multi-channel

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0046

基于迁移学习的暴恐图像自动识别

陈猛夫*

(北京航空航天大学 公共管理学院, 北京 100083)



摘 要: 利用人工智能和深度学习技术自动化地分析互联网海量图片,快速、准确地识别有害的暴恐图像并及时处置是反恐工作的重要手段之一。研究了利用深度学习和迁移学习技术对暴恐图像进行分类识别。首先,定义了暴恐图像的主要概念特征,并针对性地构建数据集;其次,针对暴恐图像正样本较少的问题,设计深度神经网络模型和迁移学习方式;最后,基于构建的训练数据集进行模型训练和测试。结果显示:所提方法可以快速、准确地对互联网图片进行分类识别,平均分类准确率达到96.7%,从而有效降低人工检测的劳动强度,为反恐预警工作提供决策支持。

关 键 词: 暴恐图像;深度学习;图像识别;卷积神经网络;迁移学习

中图分类号: TP37; TP391.4

文献标志码: A

文章编号: 1001-5965(2020)09-1677-05

暴恐势力和暴恐事件是影响社会稳定与安全的最大毒瘤。随着 Web3.0、社交媒体、5G 网络的快速发展,网络上涉恐涉暴的图像传播也随之增多。由于图像与其他传统媒体相比更具鼓动性和传播力,因此这类图像传播造成的危害更大。近年来,各类恐怖组织在互联网上发布的暴力恐怖主义图片或视频日见增多,其试图利用此类图像煽动更多人参与恐怖活动,给社会稳定造成极大影响。因此,加强涉暴恐图像内容的网络监管正成为越来越紧迫的需求。

近年来,人工智能、深度学习技术快速发展,基于深度学习技术的图像分类、图像检测技术逐渐成熟,包括人脸识别、物体识别、黄色图片识别等,但是针对暴恐图像识别的研究成果还比较稀少。

本文基于当前主流的深度学习和迁移学习技术实现对互联网图片暴恐元素的自动化识别,从而提升“算法反恐”^[1]水平,加强互联网安全监管。

1 相关研究

当前,对互联网暴恐元素的判别主要采用文本识别、视频分析、音频分析等方法。文献[2]对大数据背景下暴恐信息挖掘方法进行了综述,包括文本分析、情感分析、音视频分析等。文献[3]通过对暴恐敏感标识的检测来识别暴恐图像,如蒙面人的检测,其利用蒙面人头像图片特征呈多峰分布的特点,采用高斯混合模型(GMM)结合方向梯度直方图(HOG)特征对蒙面人头部轮廓进行检测,使用基于像素点统计的局部像素差分法(Local Pixel Difference)完成蒙面人的检测,效率较低,准确率也不高。文献[4]提出一种基于投影特征算法的机场区域中人群暴恐动作智能识别方法。文献[5]对涉暴恐音视频犯罪进行了实证研究。文献[6-7]采用循环神经网络(Recurrent Neural Network, RNN)技术对暴恐信息实体和知识图谱进行研究。总体来看,对暴恐图像分析识别的研究较少。

收稿日期: 2020-02-22; 录用日期: 2020-02-28; 网络出版时间: 2020-03-20 16:59
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200320.1433.002.html
* 通信作者。E-mail: dhzhou2084@163.com

引用格式: 陈猛夫. 基于迁移学习的暴恐图像自动识别[J]. 北京航空航天大学学报, 2020, 46(9): 1677-1681.

CHEN M F. Automatic recognition for terrorism related image based on transfer learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1677-1681 (in Chinese).

广义的暴恐组织或暴恐势力包括多个方向,其图像特征种类繁多,差异很大。在图像表现上,主要体现为:①特定标志。②特定人群。持枪支者、持刀者、蓄胡须者等。③特定穿着。黑袍、黑布蒙面等。本文研究的图像包括图片和视频,对视频的检测一般有2种方式,分别是静态方式和动态方式,前者主要通过对视频中关键帧的检测实现视频检测。本文中,对视频的检测主要采用这种方式。因此,综合来讲,暴恐图像的识别检测可以归结为对图片的识别检测。

近年来,随着深度神经网络(Deep Neural Network,DNN)的发展,卷积神经网络(Convolution Neural Network,CNN)^[8-10]成为目前深度学习技术领域中非常具有代表性的神经网络之一,在人脸识别、医疗分类、场景分类、图像分析领域取得了众多突破性的进展,在学术界常用的标准图像标注集ImageNet^[11]上,基于CNN取得了很多成就,包括图像特征提取分类、场景识别等。CNN相较于传统的图像处理算法的优点之一在于:避免了对图像复杂的前期预处理过程,尤其是人工参与图像预处理过程,CNN可以采用端到端的方式,直接输入原始图像进行一系列工作,至今已经广泛应用于各类图像相关的应用中。

2 算法设计

2.1 构建数据集

根据问题定义,把暴恐图像识别作为一个分类问题看待,形成6个子类,分别为:①特定标志类;②持枪支者、持刀者;③蓄胡须者;④身着黑袍等特定穿着者;⑤黑布蒙面者;⑥正常图片。①~⑤类均为包含暴恐元素的类别,称为正样本,第⑥类属于负样本,所有图片均为JPEG或PNG格式。正样本均来自互联网监管部门,数量较少,为了增加样本量,笔者从部分暴恐视频中抽取关键帧,保存为JPEG图片,通过这种方法有效增加了正样本数量。然后将所有数据进行人工标注,每张图片归为以上6类中的1类或者多类。负样本从ImageNet的1000个分类中选取,抽样获得5000张,最后形成训练数据集,按照8:2的比例,划分为训练数据集和验证数据集,其数量统计如表1所示。

2.2 迁移学习

深层的CNN包含大量的网络参数,训练这些参数需要大量的训练数据,由于本文组建的数据库数据量小,在训练网络时网络参数容易过拟合。为解决此问题,采用迁移学习方法。

表 1 数据集数量统计			
Table 1 Dataset quantity statistics			
类别	样本总数量/张	训练数据集/张	验证数据集/张
特定标志类	960	768	192
持枪支或刀者	860	688	172
蓄胡须者	1 200	960	240
特定穿着者	640	512	128
黑布蒙面者	830	664	166
正常图片	5 000	4 000	1 000

迁移学习的核心思想是:将在A任务训练得到的模型移植到B任务上来。在大数据时代,要想训练深度学习模型,使模型收敛并得到理想的结果,须要耗费巨大的资源,包括大量的数据资源和计算机资源。但实际情况是,很多时候用于训练模型的数据集较小,仅仅基于有限的数据集进行训练极易造成模型不收敛或者过拟合(Over-fitting)。迁移学习的关键是源任务领域和目标任务领域要有相似性,如图1所示。假如目标任务是对猫科动物进行自动分类,那么可以选择脊椎动物作为原始训练资源,对模型进行预训练(Pre-training),之后再在该模型基础上,用有限的猫科动物数据集对模型进行微调(Fine-training)。本文中,由于用于训练的暴恐图像数据集有限,因此采用迁移学习的方式,先将模型在ImageNet数据集上进行预训练得到预训练模型,再用暴恐图像数据集对预训练后的模型进行微调。

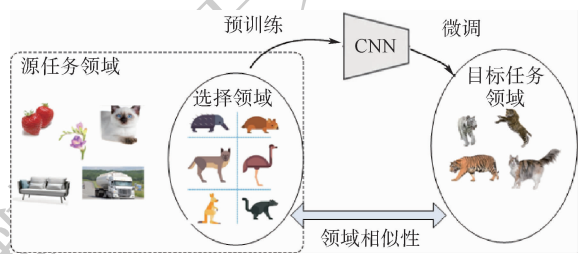


图 1 迁移学习
Fig. 1 Transfer learning

2.3 CNN 模型设计与训练

对于深度神经网络,在数据和训练算法一致的情况下,采用更先进的网络基础模型可以有效提升分类准确率,但是更先进的基础模型大部分情况下规模更大,计算量也更大,在生产系统中进行部署时,运算速度会更慢。综合考虑准确率和运行速度,选择当前ImageNet数据集上性能最好的模型,包括densenet121^[12]、resnext101^[13]、efficientnet-b3^[14]等。

模型训练和迁移学习过程如图2所示。CNN一般包括特征提取层和分类器。特征提取部分包

含多个级联的卷积层、池化层、归一层等(下文以卷积层为代表),分类器一般包括 1~3 个全连接层,如图 2 上半部分所示。本文实验中所采用的 3 个模型均用于特征提取,在图 2 中,特征提取层部分基于 3 个模型。以 densenet121 为例,先基于

ImageNet 数据集进行预训练,得到网络权重参数,再保留预训练模型网络结构与特征提取层的网络权重参数,将模型输出从 1 000 分类调整为 6 类,在自建的暴恐图像数据集上继续训练,得到微调后的网络模型,如图 2 下半部分所示。

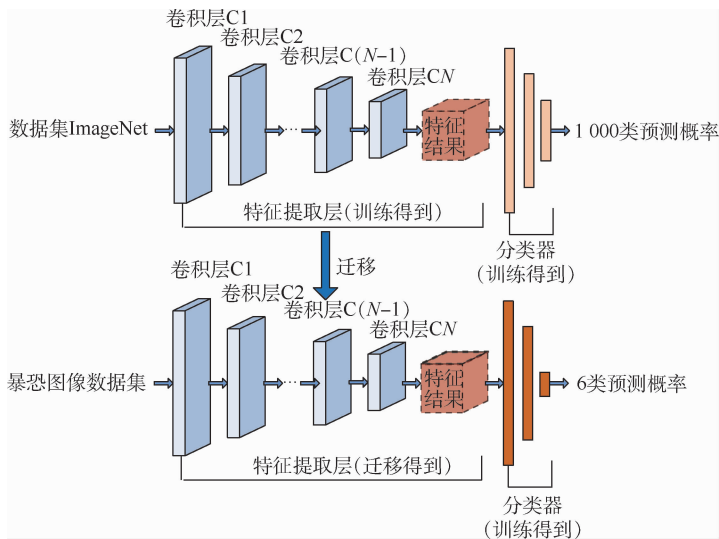


图 2 模型训练过程
Fig. 2 Model training process

3 实验分析

本文实验基于 ubuntu 16. 04 系统, 配备 NVIDIA GTXTitanX 显卡, 算法基于 Pytorch 框架实现。先使用 ImageNet 的预训练模型初始化网络权值, 再使用 Adam 算法^[15], 基于自有数据集进行网络参数学习。

自建数据集训练时, 批量大小设置为 32, 学习率初始值为 1×10^{-5} , 并且每隔 1 000 次迭代后更新学习率。学习速率 (Learning Rate) 每 20 轮 (epoch) 进行线性递减, 降低为前值的 1/10。当训练迭代 120 轮后, 结束训练。所有数据按照 8: 2 划分为训练数据集和验证数据集, 并进行 5 折交叉验证。

真实世界的图片数据变化多样, 包括亮度、饱和度、角度、位置等。在训练过程中, 考虑到原始训练数据集较小, 为避免过拟合, 同时应对真实环境中的图像多种变化, 对原始训练数据集进行多种复合数据扩增 (Data Augmentation), 图像变换方式包括随机亮度变化 (− 30% ~ 30%)、随机平移 (− 20% ~ 20%)、随机旋转 (90°、180°、270°)、随机高斯噪声、随机翻转 (水平或垂直翻转)、压缩处理 (压缩率不超过 50%)、尺寸变化 (− 60% ~ 100%)、添加水印 (加入图片水印)、裁剪 (边缘裁剪不超过 20%)

等。训练过程中, 每张图片在进入前向计算之前, 都随机地进行以上变换。实验发现, 本文方法可以有效提高模型的准确率, 同时避免过拟合, 尤其适合训练样本较小的场景。

图 3 为基于 densenet121、resnext101、efficientnet-b3 的模型训练结果。为了显示迁移学习的效果, 加入了每个模型在未做迁移学习时的平均分类准确率作为对比。

为验证本文方法的可行性, 同时对传统的支持向量机 (SVM) 算法与 CNN 进行比较。对于 SVM 模型, 通过提取全局特征和局部特征, 对暴恐图像进行识别。各算法的最优结果如表 2 所示。很明显, 3 个 CNN 模型的准确性均优于 SVM 算法, 其中 efficientnet-b3 的平均分类准确率最高, 达到了 94. 2%。

在训练过程中, 为了避免过拟合, 采用了 early stopping 策略, 也即一旦验证准确率进入稳定期, 错误率不再下降, 即停止训练。训练过程中, 每 5 轮进行一次验证, 输出一次验证结果。从训练时间来看, 到模型有效收敛, densenet121、resnext101、efficientnet-b3 分别用了 85、120、60 轮, 可以明显看出, efficientnet-b3 的训练效率是最高的。

对 3 个模型进行集成学习 (Ensemble Learning)。采取最简单直接的堆叠集成, 也即对参与集成的各个模型的输出结果进行加和后取平均

值。实验结果如表 3 所示。明显的,包含 3 个模型的集成方式平均分类准确率最高,达到了 96.7%,与单个模型最优性能相比,提升了 2.5%。

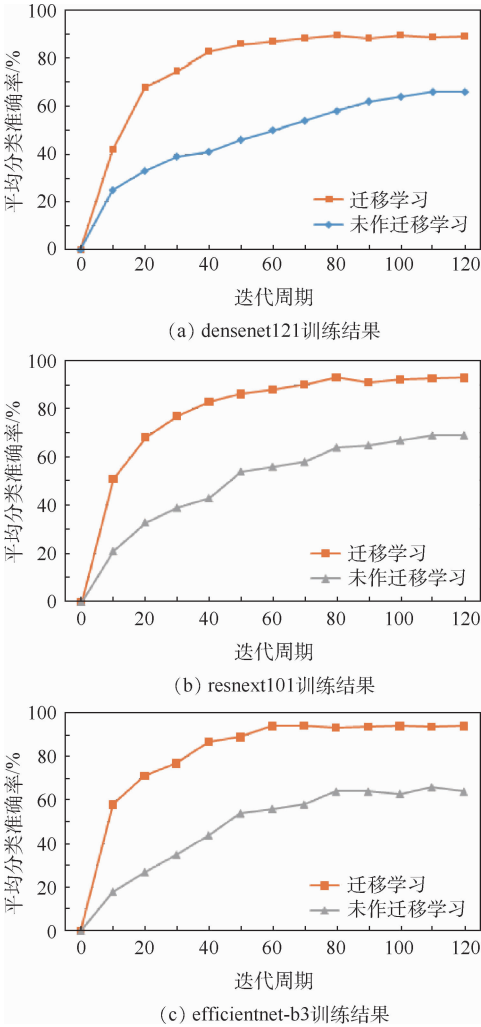


图 3 模型训练结果

Fig. 3 Model training results

表 2 不同模型平均分类准确率及训练收敛时间

Table 2 Average classification accuracy rate and training time of different models

模型	输入图像 尺寸/像素	平均分类 准确率/%	训练收敛 时间/轮
SVM		52.3	
densenet121	224	89.6	85
resnext101	224	92.9	120
efficientnet-b3	300	94.2	60

表 3 模型组合的平均分类准确率

Table 3 Average classification accuracy rate of ensemble model

模型组合	平均分类准确率/%
efficientnet-b3 + densenet121	94.1
efficientnet-b3 + resnext101	96.5
resnext101 + densenet121	93.2
efficientnet-b3 + resnext101 + densenet121	96.7

进一步分析模型分类结果(以 efficientnet-b3 模型预测结果为例),重点分析被错误预测为暴恐的图像(False Positive,假阳)和被错误预测为正常的图像(False Negative,假阴),得到:①旗帜和徽章是识别暴恐图片最有效的特征,这个类别样本的预测结果中,假阴率和假阳率分别为0.7%和1.6%,为所有类别中最低的;②存在正常的持枪和蓄胡须人物被错误分类为暴恐图像,因此该特征需要与其他特征结合在一块,用于暴恐图像的识别;③假阴样本中,主要是图片包含的暴恐元素比较单一,如仅仅是穿着上黑布蒙面,但并没有持枪或者蓄胡须等,经常被分类为正常类别。因此,在以后的工作中,应进一步考虑多个特征融合分类,以进一步降低错误率。

4 结束语

本文针对互联网上暴恐图像的检测识别,将问题收敛为图像是否包含几个显著特征,并以此为基础构建训练数据集选择合适的 CNN 基础模型。针对训练数据集不足的问题,提出使用迁移学习的方法,先在 ImageNet 数据集进行模型预训练,在此基础上,使用有限的暴恐图像数据集进行模型微调,配合大量的数据扩增技术,进行模型训练。结果显示,在训练数据较少的情况下,采用本文方法可以有效提高分类准确率。同时,也要看到暴恐图像种类繁多,形式变化,在检测结果中依然存在不少假阳和假阴样本,要进一步提高检测准确率,需要在训练数据数量和多样性方面进一步丰富,同时优化模型网络模型和训练方法。这是今后将要着力研究的方向。

参考文献 (References)

[1] 李龙,支庭荣.“算法反恐”:恐怖主义媒介化与人工智能应对[J]. 现代传播(中国传媒大学学报),2018(9):13-18.
LI L,ZHI T R. “Algorithmic anti-terrorism”: Terrorism media and the response based on artificial intelligence [J]. Modern Communication (Journal of Communication University of China),2018(9):13-18(in Chinese).

[2] 倪叶舟,张鹏,扈翔,等.大数据背景下暴恐信息挖掘方法综述[J]. 中国公共安全(学术版),2018(4):91-95.
NI Y Z,ZHANG P,HU X, et al. Summarization of the methods of information mining in the background of big data [J]. China Public Security(Academy Edition),2018(4):91-95(in Chinese).

[3] 符亚彬. 基于 Logo 标志检测的暴恐视频识别系统的设计与实现[D]. 北京:北京交通大学,2016:15-30.
FU Y B. Design and implementation of the violent-terrorist video recognition system based on Logo markers detection [D]. Bei-

- jing: Beijing Jiaotong University, 2016:15-30 (in Chinese).
- [4] 张宁, 朱金福. 机场区域中人群涉暴恐动作智能识别方法仿真[J]. 计算机仿真, 2015, 32(6): 67-70.
- ZHANG N, ZHU J F. Intelligent recognition method simulation of critical action of people involved in airport areas[J]. Computer Simulation, 2015, 32(6): 67-70 (in Chinese).
- [5] 王胜华. 涉暴恐音视频犯罪实证研究——以中国裁判文书网公开的48个判例为分析样本[J]. 江西警察学院学报, 2019(6): 89-96.
- WANG S H. An empirical study on audio and video of violent terrorist crimes—Take 48 cases published by China judicial document network as the analysis sample[J]. Journal of Jiangxi Police College, 2019(6): 89-96 (in Chinese).
- [6] 黄炜, 黄建桥, 李岳峰. 基于BiLSTM-CRF的暴恐信息实体识别模型研究[J]. 情报杂志, 2019, 38(12): 149-156.
- HUANG W, HUANG J Q, LI Y F. Research on entity identification model of terrorism-related information based on BiLSTM-CRF[J]. Journal of Intelligence, 2019, 38(12): 149-156 (in Chinese).
- [7] 廖浚斌, 周欣, 何小海, 等. 面向暴恐领域的知识图谱构建方法[J]. 信息技术与网络安全, 2019, 38(9): 34-38.
- LIAO J B, ZHOU X, HE X H, et al. Construction method of knowledge graph for terrorism domain[J]. Information Technology and Network Security, 2019, 38(9): 34-38 (in Chinese).
- [8] YAN L C, BERNHARD E B, JOHN S D, et al. Backpropagation applied to handwritten zip code recognition[J]. Neural Computation, 1989, 1(4): 541-551.
- [9] BENGIO Y, COURVILLE A, VINCENT P. Representation learning: A review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828.
- [10] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504-507.
- [11] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2012: 1097-1105.
- [12] HUANG G, LIU Z, WEINBERGER K Q, et al. Densely connected convolutional networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2261-2269.
- [13] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7132-7141.
- [14] TAN M, LE Q. EfficientNet: Rethinking model scaling for convolutional neural networks[C]// Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 2019: 6105-6114.
- [15] KINGMA D, BA J. Adam: A method for stochastic optimization[C]// International Conference on Learning Representations, 2014.

作者简介:

陈猛夫 男, 硕士, 高级工程师。主要研究方向: 数据挖掘、信息融合。

Automatic recognition for terrorism related image based on transfer learning

CHEN Mengfu*

(School of Public Administration, Beihang University, Beijing 100083, China)

Abstract: Using AI and deep learning technology to automatically analyze massive Internet pictures, quickly and accurately identifying harmful images related to terrorism and dealing with them in time is one of the important means for anti-terrorism work. This paper studies how to use deep learning and transfer learning technology to classify and recognize the images related to terrorism. First, we define the main concept features of the image related to terrorism and collect the relevant positive samples to construct dataset. Second, we design suitable deep neural network model and transfer learning method for the problem of less positive samples of the image related to terrorism. Finally, using the constructed training dataset to fine-tune the final model. The results show that, based on the proposed method in this paper, we can classify and recognize the Internet pictures which have terrorism content quickly and accurately with average classification accuracy rate of 96.7%, and thus the labor intensity of manual monitoring will reduce effectively, which can provide support for decision-making in the work of anti-terrorism early warning.

Keywords: terrorism related image; deep learning; image recognition; convolutional neural network; transfer learning

Received: 2020-02-22; Accepted: 2020-02-28; Published online: 2020-03-20 16:59

URL: kns.cnki.net/kcms/detail/11.2625.V.20200320.1433.002.html

* Corresponding author. E-mail: dhzhou2084@163.com

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0072

基于局部语义特征不变性的跨域行人重识别



张晓伟*, 吕明强, 李慧

(青岛大学 计算机科学技术学院, 青岛 266071)

摘 要: 行人重识别是刑侦案件中重要的侦查手段,而跨域是行人重识别的主要挑战之一,也是制约其实际应用的瓶颈问题。在带标签的源域和无标签的目标域学习跨域行人局部语义不变性特征模型。首先,在源域上通过只含有行人标识无部件标签的监督学习方式学习行人的各部件特征,并在源域和目标域上采用无监督学习方式对齐行人部件。然后,基于对齐后的行人全局与局部特征,引入特征模板池存储对齐后的目标域全局和局部特征,并设计了跨域不变性损失函数进行特征不变性约束,提高行人重识别的跨域适应能力。最后,在 Market-1501、DukeMTMC-reID 和 MSMT17 数据集之间开展了跨域行人重识别验证实验,实验结果表明,所提方法在跨域行人重识别上取得了显著的性能提升。

关 键 词: 行人重识别; 全局特征; 局部特征; 语义对齐; 特征模板池; 跨域不变性

中图分类号: TP37; TP277

文献标志码: A

文章编号: 1001-5965(2020)09-1682-09

行人重识别(Person Re-identification)也称行人再识别,是利用计算机视觉技术判断图像或者视频序列中是否存在某一特定行人的图像识别技术,其在智能化的公共安全领域扮演着重要角色。跨域行人重识别也被认为是图像检索的一个子问题,旨在弥补单一、固定摄像头的视觉局限,可与行人检测、行人跟踪技术相结合,检索跨摄像机设备下的同一标识的行人图像,被广泛应用于智能视频监控、智能安保等领域,具有重要的研究意义和应用价值。

近年来,伴随深度卷积神经网络的发展,行人重识别在同一场景(单域)数据集下取得了巨大的进展^[1-2]。行人重识别技术从基于深度卷积神经网络的全局特征开始向更细粒度的局部特征方向发展。基于全局特征的行人重识别主要基于深度卷积神经网络结构输出的整体行人区域作为行人图像的重识别特征,研究如何基于全局特征设计更为合理的网络损失函数。其中,为解决行人

检测对行人重识别的影响,Han等^[3]开发了一个可区分的ROI变换层,将行人检测器产生的边界框从原始图像中转换出来生成更可靠的边界框。在损失函数的设计方面,主要有对比损失(Contrastive Loss)^[4]、三元组损失(Triplet Loss)^[5]、四元组损失(Quadruplet Loss)^[6]和中心聚类损失(Center Loss)^[7]等。

由于全局特征不具有细粒度的区分能力,在光照条件、背景、人体姿势和成像视角的变化下,同一ID标识的行人在不同视图之间的特征辨识度明显降低。因此,基于局部特征的行人重识别受到研究者的日益重视。基于行人局部区域划分方式的不同,行人语义部件模型可以分为人体语义分块和物理切割分块。近年来,人体姿势估计技术取得了较大的进展^[8-11],基于人体语义分块方法^[12-14]通过借鉴人体姿态估计的结果用于行人区域划分来进行行人重识别。但姿势估计和行人重识别之间的数据集标注偏差仍然是阻碍行人

收稿日期: 2020-03-02; 录用日期: 2020-03-28; 网络出版时间: 2020-04-15 10:39

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200415.0953.001.html

基金项目: 国家自然科学基金(61902204); 山东省自然科学基金(ZR2019BF028)

* 通信作者。E-mail: by1306114@buaa.edu.cn

引用格式: 张晓伟, 吕明强, 李慧. 基于局部语义特征不变性的跨域行人重识别[J]. 北京航空航天大学学报, 2020, 46(9): 1682-1690. ZHANG X W, LYU M Q, LI H. Cross-domain person re-identification based on partial semantic feature invariance[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1682-1690 (in Chinese).

重识别图像语义划分的障碍。不同于人体语义分块方法需要对行人的强语义标注,物理切割分块不需要行人每个部件区域的训练标签,受到了研究者的重视^[15]。其中,可见部件感知模型(VPM)^[16]聚焦在2幅行人图像共同可见的部件区域上学习细粒度的区域特性,以解决行人部件与行人整体图像间的空间位置错位问题,显著地改善了特征表征的学习能力。Sun等^[17]提出了基于行人局部特征的部件卷积池化方法,以增强行人重识别中行人部件区域的语义一致性。在行人重识别模型中引入基于深度学习的注意力机制^[18-19],自动地给每一幅行人图像赋予一个权重,也进一步推动了行人重识别的性能。Chen等^[20]通过卷积神经网络特征通道聚合和空间位置感知提出了一种基于双重注意力机制的行人重识别模型,其集成了注意力模块和多样性正则化,以学习具有代表性、健壮性和更具区分性的特征。

上述基于监督学习的行人重识别方法^[4,21-23]依靠丰富的标记数据和深度网络^[24-26]在单域数据集上取得了巨大的成功。在闭合的单域行人重识别上,其源域数据集和目标域数据集具有完全相同的标识类别,但在跨域的无标签数据集上,由于跨域的数据集偏移问题导致行人重识别性能出现显著下降。为了解决这个问题,研究者们提出使用带有标记的源域数据来提取行人特征学习初始阶段的行人重识别模型。在此基础上,通过在目标域上进行无监督聚类^[27]或学习度量^[28]来改进Re-ID模型,然而这种方法在源域到目标域的特征迁移过程中并没有有效利用带有标记的源域数据作为有效监督信息。为了克服这一问题,目前跨域行人重识别方法主要是基于图像级行人图像整体特征风格^[29-31]和行人局部属性语义特征^[32-33]进行数据集之间的跨域迁移。

在图像级特征风格迁移方面,Zhong等^[34]提出了相机不变性以缓解跨域行人重识别的跨域偏移问题。Yu等^[35]提出一种基于无监督多标签学习的MAR模型,开发未标记目标对的多标签视觉特征的相似一致性,基于强化学习增强不同相机视图之间的跨视角一致性。Chen等^[36]提出了一个对偶条件生成对抗网络,利用大量未标记的目标域行人实例作为生成图像的上下文指导,将源域目标转移到目标域上下文中,通过迁移丰富的上下文信息来增强行人再识别的辨别力。在行人局部属性语义特征方面,EANet模型^[37]利用行人的关键点将行人进行分块和对准,在此基础上进行跨域迁移学习。但是,这些方法却忽略了在

目标域内同一标识行人存在的视觉差异,这种同一标识行人的类内差异将对无监督的域自适应问题产生重要的影响。对此,本文联合相机不变性和行人图像局部语义特征不变性将目标域中每个ID标识局部语义对齐的行人特征存储于特征模板池中,对目标域行人进行不变性约束,更好地发挥跨域特征不变性的作用。

综上所述,本文提出一种基于局部语义特征不变性的跨域行人重识别方法,其主要贡献如下:

1) 根据行人图像在源域与目标域上语义特征的一致性,基于弱标注的源域行人图像提出一种行人局部特征语义对齐方法,以迁移源域行人局部语义特征到无标注的目标域,增强跨域行人重识别的特征表达能力。

2) 针对同一ID标识行人图像存在的表现特征变化,通过估计目标域图像之间的相似性,采用特征模板池存储并更新目标域中行人图像的全局与局部特征,将跨域不变性加入到深度网络中提高对跨域特征不变性的学习。

3) 鉴于行人重识别在源域与目标域上的特征差异,基于行人局部特征语义对齐和行人特征模板池提出了一种联合实例不变性、相机不变性和类别(标识)不变性的损失函数,无监督地学习跨域不变性特征,提高行人重识别的跨域适应能力。

1 局部语义特征不变性网络

1.1 网络框架

本文提出的基于局部语义特征不变性的网络结构框架如图1所示。基础骨干网络采用ImageNet^[38]上预训练的ResNet-50^[25]网络,其将带标签的源域数据和无标签的目标域数据输入到网络中获取行人实例的特征表示。本文保留ResNet-50的Pooling-5层及之前的网络结构,深度特征的提取分为全局特征模块和局部特征模块。在全局特征模块中添加全局平均池化(Global Average Pooling, GAP)层以获得全局特征,使用交叉熵损失函数与三元组损失函数来优化模型。在局部特征模块添加自适应平均池化层,将行人图像提取到的三维特征张量 T 分成 P 个部分,并降维到256维,分别为每个部分添加分类器来进行训练,使用交叉熵损失函数来优化模型。在目标域中,引入特征模板池保存每个行人样本的全局与局部特征,通过估计小批量的目标样本与特征模板池中的特征的相似性,计算无标签目标域的不变性损失。

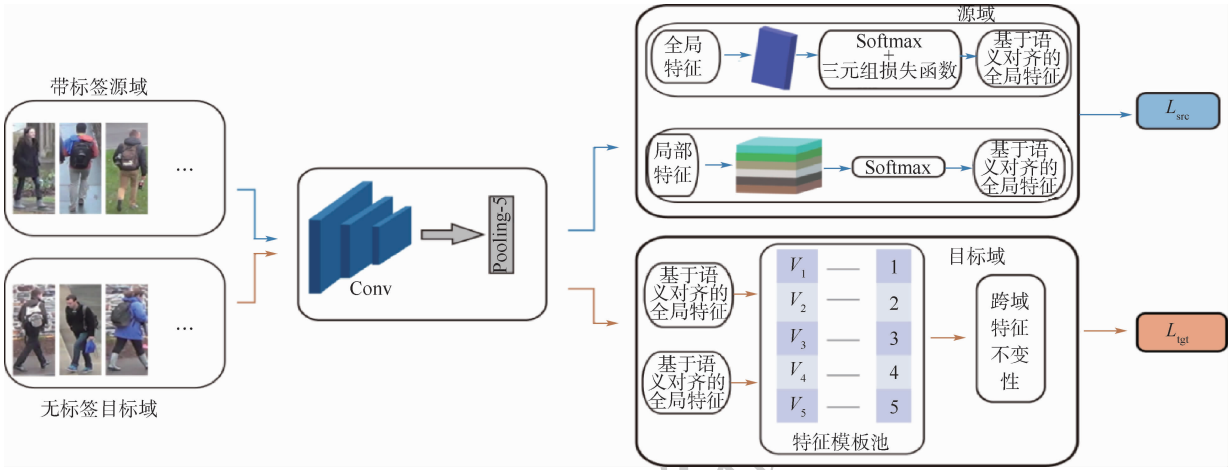


图1 基于局部语义特征不变性的跨域行人重识别框架

Fig. 1 Framework of cross-domain person re-identification based on partial semantic feature invariance

1.2 基于语义对齐的全局与局部特征

1.2.1 全局特征

本文在 ResNet-50 的 Pooling-5 层之后,采用 GLAD 模型^[14]的全局平均池化来提取全局特征,之后再经过 Softmax 分类器预测行人样本的标识,使用交叉熵损失函数 L_s 和三元组损失函数 L_t 来共同计算源域全局损失 L_{src} ,其计算公式为

$$L_{src} = \eta_1 L_s + \eta_2 L_t \tag{1}$$

式中: η_1 和 η_2 为损失函数平衡因子,本文取值为 $\eta = [0.96, 1]$ 。

1.2.2 基于语义对齐的局部特征

本文行人局部特征的提取以 PCB 模型^[3]为基础,将行人图像从上到下均匀划分为 6 个部件区域,如图 2(a)所示。为了实现行人部件的语义对齐,本节基于行人局部部件的内部特征,把每个局部部件分成 4×8 的特征块 f ,每副行人图像被分为 6 个语义部件,共包含 24×8 个特征块 f ,如图 2(b)所示。首先,在源域数据集上,将基于行人图像划分的 6 个部件区域进行训练学习,形成行人部件分类器 W_i 。然后,对每个特征块 f 使用线性分类 Softmax 函数进行动态分类:

$$p(P_i | f) = \text{softmax}(W_i^T f) = \exp(W_i^T f) / \sum_{j=1}^P \exp(W_j^T f) \tag{2}$$

式中: $p(P_i | f)$ 为属于 P_i 部件的预测概率; P 为预定义行人部件的数量(即 $P = 6$); W 为部件分类器的可训练权重矩阵。

为保证部件分类的准确性,为每个部件分类器的预测结果设立一个阈值。经过此分类,每个 f 得到 P 个预测概率 $p(P_i | f)$,取其中的最大值 $p(P_i | f)$,并且 $p(P_i | f)$ 要大于阈值 T_i ,满足该条件就把 f 分为 P_i 部件中,经过这步处理可以筛选出

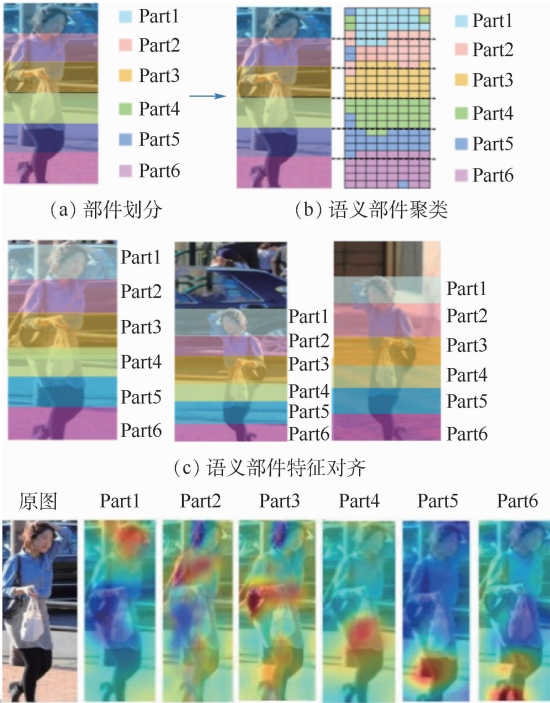


图2 基于语义特征的行人部件示意图
Fig. 2 Schematic diagram of person parts based on semantic features

属于各个部件的 f ;之后,对每幅行人图像中分类预测为同一部件类别的 f 做局部平均池化,获取行人各个部件的特征表示。

为对齐行人部件区域,采用成对样本对计算最小总距离^[39],形成 2 幅图像的最短路径。假定待查询行人图像和搜索库中的某一图像经式(2)提取的局部特征分别为 $F = \{h_1, h_2, \dots, h_p\}$ 和 $G = \{g_1, g_2, \dots, g_p\}$,则 2 幅行人图像成对部件元素之间的距离规范化为

$$d_{i,j} = \exp(\|h_i - g_j\|_2) - 1 / \exp(\|h_i - g_i\|_2) + 1 \tag{3}$$

式中: $d_{i,j}$ 为待查询行人图像的第 i 个部件和搜索库中某一行人图像的第 j 部件之间的距离,其取值范围为 $[0,1]$ 。

本文基于 $P \times P$ 个距离 $d_{i,j} (i, j \in (1, 2, \dots, P))$ 形成的距离矩阵 $\mathbf{D}$ 中的最短路径来对齐待查询行人图像与搜索库中行人图像的语义局部区域。其中,2 张行人图像相似性距离 $S_{i,j}$ 是通过动态规划计算距离矩阵 $\mathbf{D}$ 中每行或每列对应最短路径的总距离,即为

$$S_{i,j} = \begin{cases} d_{i,j} & i = 1, j = 1 \\ S_{i-1,j} + d_{i,j} & i \neq 1, j = 1 \\ S_{i,j-1} + d_{i,j} & i = 1, j \neq 1 \\ \min(S_{i-1,j}, S_{i,j-1}) + d_{i,j} & i \neq 1, j \neq 1 \end{cases} \quad (4)$$

经过行人部件划分和语义特征对齐后的部件区域如图 2(c) 所示,每个部件区域的局部特征注意力位于行人的不同部位,如图 2(d) 所示。基于语义对齐后的行人部件局部特征,其行人重识别源域局部损失 L_{src2} 为

$$L_{src2} = \gamma_1 S_1 + \gamma_2 S_2 + \gamma_3 S_3 + \gamma_4 S_4 + \gamma_5 S_5 + \gamma_6 S_6 \quad (5)$$

式中: S_i 为部件 P_i 的交叉熵损失;权重向量 γ 的取值,通过实验分析分别取 $[0.18, 0.15, 0.15, 0.17, 0.16, 0.19]$ 重识别性能最佳。

1.3 特征模板池

为了提高网络在目标测试集上的泛化能力,提出通过估计目标图像之间的相似性来增强网络的特征不变性。为此,构造了一个具有键值结构用于存储所有目标域图像的全局与局部特征的特征模板池,其具有键池(K)和值池(V)。其中,键池存放在行人图像经 ResNet-50 全连接后 L_2 规范化的特征,包括语义对齐后全局特征和局部特征的串联聚合特征,分别为 1×4096 维和 6×256 维;值池存储行人图像的标签。

假定目标域中包含 N_t 副无标签的行人图像,特征模板池就有 N_t 个键池,每个键池存放 5 632 维的聚合特征,其值池存放行人实例的类别标注。在初始化过程中,将键池中所有功能的值初始化为零,将相应的索引指定为目标样本的标签,并存储在值池中。例如,值池中的第 i 个目标图像的类被分配给 $V[i] = i$ 。在整个训练过程中,值池中的标签是固定的,在每次训练迭代过程中将目标域行人图像训练样本 $x_{t,i}$ 经 ResNet-50 得到的特征 $h(x_{t,i})$ 进行 L_2 归一化。在反向传播过程中,更新训练样本 $x_{t,i}$ 的键池中的特征权重矩阵:

$$\mathbf{R}[i] = a\mathbf{R}[i] + (1 - a)\mathbf{R}[i] \quad (6)$$

式中: $\mathbf{R}[i]$ 为图像 $x_{t,i}$ 的特征模板池中键池存放的特征, i 为标签;超参数 $a \in [0,1]$ 控制更新速率。

1.4 特征不变性学习

鉴于仅使用源域的监督学习得到的深度 Re-ID 模型对目标域内的变化较为敏感,考虑目标域内的差异变化,研究了 3 个跨域特征不变性,即实例不变性、相机不变性、类别不变性。

1) 实例不变性。在行人重识别中,即使是同一 ID 标识的图像间也会有差距,为了学习每一张图像都有别于其他图像,本文在 Re-ID 模型中增加实例不变性。将 n_t 个目标域图像看做 n_t 个独立不同的类别,计算某一给定目标图像特征 $h(x_{t,i})$ 与已保存在特征模板池中的特征之间的余弦相似度,并使用 Softmax 函数预测其归属类别:

$$p(i | x_{t,i}) = e^{(\mathbf{R}[i]^T h(x_{t,i}) / \beta)} / \sum_{j=1}^{n_t} e^{(\mathbf{R}[j]^T h(x_{t,i}) / \beta)} \quad (7)$$

式中: $\beta \in (0,1]$ 为平衡因子。

实例不变性的目标是使目标域上训练图像的负对数似然最小化:

$$L_{ei} = -\ln p(i | x_{t,i}) \quad (8)$$

式中: L_{ei} 为实例不变性损失。

2) 相机不变性。相机风格的变化是影响行人图像识别的一个重要因素,尤其是在源域与目标域之间往往存在迥异的相机风格。因此,本文基于相机不变性拉近同一 ID 标识的行人图像与其经过相机风格变换后的距离。假设每个图像的摄像机 ID 是已知的,将每台摄像机视为一个相机域,利用 starGAN^[40] 来训练一个相机风格转换模型^[34],利用学习到的相机风格转换 Re-ID 模型。假设 C 为目标域中相机风格的数量,一张行人图像则可以得到 $C - 1$ 张其他相机风格的图像,且这 C 张图片属于同一行人 ID,相机不变性的损失函数为

$$L_{ci} = -\ln p(i | \bar{x}_{t,i}) \quad (9)$$

式中: L_{ci} 为相机不变性的损失; $\bar{x}_{t,i}$ 为 $x_{t,i}$ 及其 $C - 1$ 张生成图像的集合中随机选取的一张图像,拉近目标域行人图像集中具有同一标识、不同相机风格的行人图像。

3) 类别不变性。基于目标域中属于同一 ID 标识的行人图像样本集拉近彼此间的距离克服目标域中的变化,以提高 Re-ID 模型的稳定性。先计算图像特征 $h(x_{t,i})$ 与其特征模板池中键池的余弦相似性。在特征模板池中找到与行人图像 $x_{t,i}$ k 邻近的行人图像索引,定义为 $M(x_{t,i}, k)$ 。本文提出的类别不变性是假定目标域行人图像 $x_{t,i}$

的 ID 标识应该隶属于 k 邻近图像索引 $M(x_{t,i}, k)$ 的候选类。因此,将目标域行人图像 $x_{t,i}$ 属于某一 ID 标识 j 的概率权重赋值为

$$\omega_{i,j} = \begin{cases} \frac{1}{k} & i \neq j \\ 1 & i = j \end{cases} \quad \forall j \in M(x_{t,i}, k) \quad (10)$$

类别不变性的损失函数为

$$L_{ni} = - \sum_{j \neq i} \omega_{i,j} \ln p(j|x_{t,i}) \quad \forall j \in M(x_{t,i}, k) \quad (11)$$

式中: L_{ni} 为类别不变性损失。

基于上述实例不变性、相机不变性、类别不变性,跨域特征不变性学习的总体损失可以写为

$$L_{lgt} = - \frac{1}{n_t} \sum_{i=1}^{n_t} \sum_j \omega_{i,j} \ln p(j|\hat{x}_{t,i}) \quad (12)$$

式中: $j \in M(\hat{x}_{t,i}, k)$; $\hat{x}_{t,i}$ 为从 $x_{t,i}$ 经相机风格迁移模型生成后的随机采样图像。当 $i=j$ 时,使用实例不变性和相机不变性来优化网络;当 $i \neq j$ 时,使用相机不变性和类别不变性来优化网络。

合并源域和目标域的特征不变性网络的损失函数为

$$L = \lambda (L_{src1} + L_{src2}) + (1 - \lambda) L_{lgt} \quad (13)$$

式中: λ 为平衡源域与目标域损失的权重系数,本文取值为 0.3。

2 实验结果分析

2.1 数据集与实验设置

本文在国际公开的行人重识别 (Re-ID) 基准数据集 Market-1501^[41]、DukeMTMC-reID^[6,42] 和 MSMT17^[30] 上评估了本文方法。本文采用 ResNet-50^[25] 作为模型的骨干网络,并使用在 ImageNet^[43] 上预先训练的参数初始化模型,输入图像的大小采样到 384×128 。在前 40 次 epoch 迭代训练阶段中,使用 0.01 的学习率对 ResNet-50

的基础网络层进行训练,在接下来的 20 次 epoch 迭代中,学习率每个阶段都以除以 10 的比例缩放。源域行人图像和目标域行人图像的批量大小都设置为 128,网络模型学习更新率 α 初始化为 0.01,并随着训练次数的增加而线性增大, $\alpha = 0.01 \times \text{epoch}$ 。本文设置行人部件 $P=6$,每个部件的权重向量 γ 依次设为 $[0.18, 0.15, 0.15, 0.17, 0.16, 0.19]$,参数 $\beta=0.05$,特征模板池候选正性样本数 $k=6$,损失权重 $\lambda=0.3$ 。在测试中,提取 Pooling-5 层的 L_2 规范化输出作为图像特征,并采用欧几里德距离度量查询图像与图库图像的相似度。

2.2 消融实验

2.2.1 基于语义对齐的特征不变性分析

针对目标域提出的语义部件特征对齐与跨域特征不变性学习的有效性,本文在 Market-1501 和 DukeMTMC-reID 数据集上开展了验证实验。R- k 为返回置信度最高的 k 张行人图像具有正确结果的概率, mAP 为平均准确率。将全局特征作为实验基线,记为 GAP,基于语义部件对齐的局部特征记为 PAF (Part Aligned Feature),跨域特征不变性记为 Inv (Invariance)。在表 1 由 DukeMTMC-reID $\rightarrow$ Market-1501 数据集的跨域行人重识别实验中,联合语义部件对齐后的局部特征和全局特征 (GAP + PAF) 较全局特征基线 (GAP) mAP 提高 1.6%, R-1 提高 1.5%。而且无论对全局特征还是局部特征,加入跨域特征不变性 (Inv) 均提高了行人重识别的性能,其平均准确率 mAP 分别提高 13.1% 和 23.8%, R-1 的性能分别提高了 20.6% 和 29.8%。实验发现,联合全局与局部特征和特征不变性取得了最优的行人重识别性能,其 R-1 和 mAP 分别达到了 77.6% 和 45.0%。同样的实验结论在 Market-1501 $\rightarrow$ DukeMTMC-reID 数据集的跨域行人重识别中得到了验证。

表 1 在 Market-1501 和 DukeMTMC-reID 数据集上的跨域行人重识别消融实验
Table 1 Ablation experiment of cross-domain person re-identification on Market-1501 and DukeMTMC-reID datasets

行人重识别 跨域方法	DukeMTMC-reID $\rightarrow$ Market-1501					Market-1501 $\rightarrow$ DukeMTMC-reID				
	R-1/%	R-5/%	R-10/%	R-20/%	mAP/%	R-1/%	R-5/%	R-10/%	R-20/%	mAP/%
GAP	45.1	62.4	68.9	75.2	18.8	29.6	43.9	50.2	57.0	15.1
PAF	45.9	63.5	69.9	75.7	19.7	30.0	45.4	51.8	58.1	15.5
GAP + PAF	46.6	63.9	70.6	76.3	20.4	30.3	46.7	52.4	58.8	16.0
Inv + GAP	65.7	81.5	86.0	89.9	31.9	52.7	64.7	69.7	73.4	27.0
Inv + PAF	75.7	89.4	92.0	94.3	43.5	64.1	76.5	80.3	83.8	41.0
Inv + GAP + PAF	77.6	88.7	92.0	95.2	45.0	65.5	77.6	81.1	84.5	42.8

行人重识别 单域方法	Market-1501 $\rightarrow$ Market-1501					DukeMTMC-reID $\rightarrow$ DukeMTMC-reID				
	R-1/%	R-5/%	R-10/%	R-20/%	mAP/%	R-1/%	R-5/%	R-10/%	R-20/%	mAP/%
Inv + GAP + PAF	88.2	95.6	97.4	98.3	67.0	74.7	85.5	89.1	91.8	54.1

值得注意的是,相对于单域行人重识别,跨域行人重识别的性能均出现了显著下降。从表 1 中可以观察到,在 Market-1501 单域上本文行人重识别方法的平均准确率 mAP 为 67.0%,R-1 为 88.2%;但是在跨域 DukeMTMC-reID 作为源域,Market-1501 作为目标域,行人重识别的平均准确率 mAP 下降到 45.0%,R-1 下降到 77.6%。同理,Market-1501 作为源域,DukeMTMC-reID 作为目标域,行人重识别的平均准确率 mAP 也显现出大幅性能下降。由此可以看出,跨域行人重识别仍然是一个重大挑战。

2.2.2 特征模板池的重要性分析

为验证本文提出的特征模板池对跨域行人重识别的有效性,在全局与局部特征不变性的基础上检验特征模板池的作用。从表 2 中可以看出,添加特征模板池的方法优于基于小批量的跨域行人重识别方法,其在 R-1 和 mAP 分别提高近 11% 和 8.8%。但是,使用特征模板池的缺点是会增加有限的额外训练时间和 GPU 内存的开销。

表 2 特征模板池与 Mini-batch 的比较
Table 2 Comparison of feature template pooling memory and Mini-batch

方法	DukeMTMC-reID→Market-1501				
	R-1/%	R-5/%	R-10/%	R-20/%	mAP/%
Mini-batch	66.4	83.5	88.3	90.6	36.2
Mini-batch + Memory	77.6	88.7	92.0	95.2	45.0

2.3 对比实验

为横向比较本文提出的基于局部语义特征不变性的跨域行人重识别有效性,在 Market-1501、DukeMTMC-reID 和 MSMT17 数据集上开展跨域行人重识别性能的对比验证。在 Market-1501→MSMT17 跨数据集的行人重识别实验中,表 3 显

示了本文方法相较于当前一流的行人重识别方法 ECN^[44],R-1 提高 0.3%,R-5 提高 0.8%,R-10 提高 1.2%,mAP 提高 0.4%。在 DukeMTMC-reID→MSMT17 跨数据集的实验中,本文方法相较于行人重识别方法 ECN^[44],R-1 提高 0.4%,R-5 提高 0.3%,R-10 提高 0.8%,mAP 提高 0.5%。从表 4 中可以看到,在 DukeMTMC-reID→Market-1501 的跨数据集行人重识别实验中,行人重识别的性能 R-1 为 77.6%,R-5 为 88.7%,R-10 为 92.0%,mAP 为 45.0%,相较于当前一流的跨域行人重识别方法 ECN^[44]均有较为明显的性能提升,分别提高 2.5%、1.1%、0.4% 和 2.0%。在跨数据集 Market-1501→DukeMTMC-reID 的实验中,本文方法相较于行人重识别方法 ECN^[44],R-1 提高 2.2%,R-5 提高 1.8%,R-10 提高 0.7%,mAP 提高 2.4%。由此可以看出,本文基于局部语义特征不变性的跨域行人重识别方法有效地提高了跨域行人重识别的性能。

表 3 在 Market-1501 和 DukeMTMC-reID 到 MSMT17 跨域数据集上与当前先进方法的实验比较

Table 3 Experimental comparison with other advanced methods from Market-1501 and DukeMTMC-reID to MSMT17 cross-domain datasets

方法	Market-1501→MSMT17			
	R-1/%	R-5/%	R-10/%	mAP/%
PTGAN ^[30]	10.2		24.4	2.9
ECN ^[44]	25.3	36.3	42.1	8.5
本文	25.6	37.1	43.3	8.9

方法	DukeMTMC-reID→MSMT17			
	R-1/%	R-5/%	R-10/%	mAP/%
PTGAN ^[30]	11.8		27.4	3.3
ECN ^[44]	30.2	41.5	46.8	10.2
本文	30.6	41.8	47.6	10.7

表 4 在 Market-1501 和 DukeMTMC-reID 跨数据集上与当前先进方法的实验比较

Table 4 Experimental comparison with other advanced methods on Market-1501 and DukeMTMC-reID cross-domain datasets

行人重识别跨域方法	DukeMTMC-reID→Market-1501				Market-1501→DukeMTMC-reID			
	R-1/%	R-5/%	R-10/%	mAP/%	R-1/%	R-5/%	R-10/%	mAP/%
MMFA ^[33]	56.7	75.0	81.8	27.4	45.3	59.8	66.3	24.7
SPGAN + LM ^[29]	57.7	75.8	82.4	26.7	46.4	62.3	68.0	26.2
TJ-AIDL ^[32]	58.2	74.8	81.1	26.5	44.3	59.6	65.0	23.0
CamStyle ^[45]	58.8	78.2	84.3	27.4	48.4	62.5	68.9	25.1
HHL ^[46]	62.2	78.8	84.0	31.4	46.9	61.0	66.7	27.2
ECN ^[44]	75.1	87.6	91.6	43.0	63.3	75.8	80.4	40.4
本文	77.6	88.7	92.0	45.0	65.5	77.6	81.1	42.8

3 结束语

本文通过弱监督学习方式对齐行人语义特征,并结合局部特征与全局特征提高了行人的表征能力;联合跨域不变性损失函数约束行人表观特征的不变性提高了行人重识别的跨域适应能力。但是,相较于单域行人重识别,由于源域与目标域行人图像存在的特征分布差异导致跨域行人重识别仍然具有较大的挑战性。

参考文献 (References)

- [1] ZHENG L, YANG Y, HAUPTMANN A G. Person re-identification: Past, present and future [EB/OL]. (2016-10-10) [2020-03-01]. <https://arxiv.org/abs/1610.02984>.
- [2] 田萱, 王亮, 丁琪. 基于深度学习的图像语义分割方法综述 [J]. 软件学报, 2019, 30(2): 440-468.
TIAN X, WANG L, DING Q. Review of image semantic segmentation based on deep learning [J]. Journal of Software, 2019, 30(2): 440-468 (in Chinese).
- [3] HAN C C, YE J C, ZHONG Y S, et al. Re-ID driven localization refinement for person search [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 9814-9823.
- [4] VARIOR R R, HALOI M, WANG G. Gated Siamese convolutional neural network architecture for human re-identification [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 791-808.
- [5] CHENG D, GONG Y H, ZHOU S P, et al. Person re-identification by multi-channel parts-based CNN with improved triplet loss function [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1335-1344.
- [6] ZHENG Z D, ZHENG L, YANG Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 3754-3762.
- [7] WEN Y, ZHANG K, LI Z F, et al. A discriminative feature learning approach for deep face recognition [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 499-515.
- [8] 蔡铁珩, 王雪艳, 胡绍斌, 等. 基于多源图像弱监督学习的3D人体姿态估计 [J]. 北京航空航天大学学报, 2019, 45(12): 2375-2384.
CAI Y H, WANG X Y, HU S B, et al. Three-dimensional human pose estimation based on multi-source image weakly-supervised learning [J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2375-2384 (in Chinese).
- [9] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 7291-7299.
- [10] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 4724-4732.
- [11] 郭琳, 秦世引. 遥感图像飞机目标高效搜检深度学习优化算法 [J]. 北京航空航天大学学报, 2019, 45(1): 159-173.
GUO L, QIN S Y. Deep learning and optimization algorithm for high efficient searching and detection of aircraft targets in remote sensing images [J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(1): 159-173 (in Chinese).
- [12] ZHENG L, HUANG Y, LU H, et al. Pose invariant embedding for deep person re-identification [J]. IEEE Transactions on Image Processing, 2019, 28(9): 4500-4509.
- [13] SU C, LI J, ZHANG S, et al. Posedrive deep convolutional model for person re-identification [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 3960-3969.
- [14] WEI L, ZHANG S, YAO H, et al. GLAD: Global-local-alignment descriptor for pedestrian retrieval [C] // ACM International Conference on Multimedia. New York: ACM, 2017: 420-428.
- [15] YAO H, ZHANG S, HONG R, et al. Deep representation learning with part loss for person re-identification [J]. IEEE Transactions on Image Processing, 2019, 28(6): 2860-2871.
- [16] 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] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 393-402.
- [17] SUN Y F, ZHENG L, YANG Y, et al. Beyond part models: Person retrieval with refined part pooling [C] // European Conference on Computer Vision. Berlin: Springer, 2019: 480-496.
- [18] LIU X, ZHAO H, TIAN M, et al. HydraPlus-Net: Attentive deep features for pedestrian analysis [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 350-359.
- [19] ZHAO L, LI X, WANG J, et al. Deeply-learned part-aligned representations for person re-identification [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 3219-3228.
- [20] CHEN T L, DING S J, XIE J Y, et al. ABD-Net: Attentive but diverse person re-identification [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 8351-8361.
- [21] LI W, ZHU X T, GONG S G, et al. Harmonious attention network for person re-identification [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2285-2294.
- [22] SUN X X, ZHENG L. Dissecting person re-identification from the viewpoint of viewpoint [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 608-617.
- [23] ZHENG Z D, YANG X D, YU Z D, et al. Joint discriminative and generative learning for person re-identification [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2138-2147.
- [24] LUO Y W, ZHENG L, GUAN T, et al. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2138-2147.

- and Pattern Recognition. Piscataway: IEEE Press, 2019: 2507-2516.
- [25] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
- [26] DONG X Y, YANG Y. Searching for a robust neural architecture in four GPU hours [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 1761-1770.
- [27] FAN H H, ZHENG L, YAN C G, et al. Unsupervised person re-identification: Clustering and finetuning [J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2018, 14(4): 1-18.
- [28] YU H X, WU A C, ZHENG W S. Crossview asymmetric metric learning for unsupervised person re-identification [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 994-1002.
- [29] DENG W J, ZHENG L, YE Q X, et al. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 994-1003.
- [30] WEI L H, ZHANG S J, GAO W, et al. Person transfer GAN to bridge domain gap for person re-identification [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 79-88.
- [31] BAK S, CARR P, LALONDE J F. Domain adaptation through synthesis for unsupervised person re-identification [C] // European Conference on Computer Vision. Berlin: Springer, 2018: 189-205.
- [32] WANG J Y, ZHU X T, GONG S G, et al. Transferable joint attribute-identity deep learning for unsupervised person re-identification [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2275-2284.
- [33] LIN S, LI H L, LI C T, et al. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification [C] // British Machine Vision Conference. Berlin: Springer, 2018: 1-13.
- [34] ZHONG Z, ZHENG L, ZHENG Z D, et al. Camera style adaptation for person re-identification [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 5157-5166.
- [35] YU H X, ZHENG W S, WU A C, et al. Unsupervised person re-identification by soft multilabel learning [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 2148-2157.
- [36] CHEN Y B, ZHU X T, GONG X T, et al. Instance-guided context rendering for cross-domain person re-identification [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 232-242.
- [37] HUANG H J, YANG W J, CHEN X T, et al. EANet: Enhancing alignment for cross domain person re-identification [EB/OL]. (2018-12-29) [2020-03-01]. <https://arxiv.org/abs/1812.11369>.
- [38] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.
- [39] ZHANG X, LUO H, FAN X, et al. AlignedReID: Surpassing human-level performance in person re-identification [EB/OL]. (2017-11-22) [2020-03-01]. <https://arxiv.org/abs/1711.08184>.
- [40] CHOI Y, CHOI M, KIM M, et al. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8789-8797.
- [41] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: A benchmark [C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1116-1124.
- [42] RISTANI E, SOLERA F, ZOU R, et al. Performance measures and a data set for multi-target, multi-camera tracking [C] // European Conference on Computer Vision. Berlin: Springer, 2016: 17-35.
- [43] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge [J]. International Journal of Computer Vision, 2015, 115: 211-252.
- [44] ZHONG Z, ZHENG L, LUO Z M, et al. Invariance matters: Exemplar memory for domain adaptive person re-identification [C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 598-607.
- [45] ZHONG Z, ZHENG L, ZHENG Z D, et al. CamStyle: A novel data augmentation method for person re-identification [J]. IEEE Transactions on Image Processing, 2019, 28(3): 1176-1190.
- [46] ZHONG Z, ZHENG L, LI S Z, et al. Generalizing a person retrieval model hetero- and homogeneously [C] // European Conference on Computer Vision. Berlin: Springer, 2018: 172-188.

作者简介:

张晓伟 男,博士,讲师。主要研究方向:图像/视频分析和理解、计算机视觉和机器学习。

吕明强 男,本科生。主要研究方向:行人重识别。

李慧 女,硕士研究生。主要研究方向:行人重识别和计算机视觉。

Cross-domain person re-identification based on partial semantic feature invariance

ZHANG Xiaowei^{*}, LYU Mingqiang, LI hui

(College of Computer Science and Technology, Qingdao University, Qingdao 266071, China)

Abstract: Cross-domain is one of the main challenges of person re-identification which is an important investigative method in criminal investigation cases, and that restricts the practical application of re-identification. In this paper, cross-domain invariance feature model based on pedestrian partial semantics is learned from the labeled source domain and the unlabeled target domain. First, features of person parts are learned by supervised learning with only pedestrian signs without labels of parts, and pedestrian parts are aligned by unsupervised learning in the source and target domains. Second, based on the aligned global and local features, feature template pooling is introduced to store the aligned global and local partial features of the target domain, and cross-domain invariance loss function is designed for feature invariance constraints to improve the cross-domain adaptability of person re-identification. Finally, verification experiments of cross-domain person re-identification are conducted on the Market-1501, DukeMTMC-reID and MSMT17 datasets. The experiment results show that the proposed method achieves significant performance improvements in cross-domain person re-identification.

Keywords: person re-identification; global feature; local feature; semantic alignment; feature template pooling; cross-domain invariance

Received: 2020-03-02; **Accepted:** 2020-03-28; **Published online:** 2020-04-15 10:39
URL: kns.cnki.net/kcms/detail/11.2625.V.20200415.0953.001.html
Foundation items: National Natural Science Foundation of China (61902204); Natural Science Foundation of Shandong Province of China (ZR2019BF028)
*** Corresponding author.** E-mail: by1306114@buaa.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0057



基于三元组网络的单图三维模型检索

杜雨佳^{1,2,3}, 李海生^{1,2,3,*}, 姚春莲^{1,2,3}, 蔡强^{1,2,3}

(1. 北京工商大学 计算机与信息工程学院, 北京 100048; 2. 农产品质量安全追溯技术及应用国家工程实验室, 北京 100048;
3. 食品安全大数据技术北京市重点实验室, 北京 100048)

摘 要: 随着媒体数据的多样化发展, 联合图像与三维模型的跨域检索成为三维模型检索问题的一个新挑战。针对图像与三维模型差异大、难匹配问题, 提出了一种基于三元组网络的跨域数据检索方法。以端到端的方式构建真实图像与三维模型的特征联合嵌入空间, 通过特征间的距离度量不同模态数据之间的相似性, 实现从单张图像检索相似的三维模型。为了提高跨域检索准确度, 将三维模型用一组顺序视图表示, 结合门控循环单元 (GRU) 聚合视图级特征, 同时引入注意力机制提取图像特征, 缩小真实图像与投影视图间的语义差异。实验结果表明: 相比于同类方法, 所提方法在两个跨域数据集上的检索平均准确率至少提升 2.98% ~ 3.05%。

关键词: 三维模型检索; 深度学习; 跨域检索; 三元组网络; 门控循环单元 (GRU); 注意力机制

中图分类号: TP183; TP391

文献标志码: A

文章编号: 1001-5965(2020)09-1691-10

信息时代的到来为计算机硬件的发展提供了强大的助力, 各种媒体数据如音频、视频、图像、三维数据等呈井喷式增长。如何对已获得的海量媒体数据进行智能分析, 并设计快速有效的检索方法成为目前的热点问题。如今, 三维模型被广泛应用于计算机辅助设计、影视动画、医疗诊断、自动驾驶^[1-2]等众多领域, 由于三维视角更贴近于人类日常生活, 因此开发便捷的 3D 应用程序有助于提高工作效率, 促进生产发展。公共安全监控视频是维护治安稳定的重要信息资源, 利用视频中的图像数据检索三维模型在增强现实、3D 打印、3D 场景重建等研究中有重要的应用^[3-4], 有助于提高对公共安全监控视频数据内容分析效率和质量, 实现警务实战应用系统向高效化、三维化、

智能化发展。

近年来, 很多学者在三维模型检索方面做了卓有成效的工作, 取得了一定的成果^[5-7]。但是, 已有工作往往属于基于实例的三维模型检索, 根据输入三维模型检索同类相似的三维模型, 属于同域检索。在实际应用中, 用于查询的三维模型往往并不容易获取, 相比之下, 图像数据的获取则方便得多。因此, 基于图像的三维模型跨域检索具有重要的研究意义和应用前景。

基于图像检索三维模型属于跨域检索问题, 其输入可以是 RGB 图像、手绘草图或 RGB-D 图像, 输出为与图像对应的三维模型, 本文主要关注基于单幅 RGB 图像的三维模型检索。近年来, 深度学习在二维图像领域取得了显著的成绩^[8-9],

收稿日期: 2020-02-28; 录用日期: 2020-03-28; 网络出版时间: 2020-04-17 12:56

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200417.1206.003.html

基金项目: 国家自然科学基金 (61877002); 北京市自然科学基金-丰台轨道交通前沿研究联合基金 (L191009); 北京市教委科研团队建设项目 (PXM2019_014213_000007)

* 通信作者: E-mail: lihsh@btau.edu.cn

引用格式: 杜雨佳, 李海生, 姚春莲, 等. 基于三元组网络的单图三维模型检索[J]. 北京航空航天大学学报, 2020, 46(9): 1691-1700. DU Y J, LI H S, YAO C L, et al. Monocular image based 3D model retrieval using triplet network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1691-1700 (in Chinese).

随着大规模三维数据集的建立,深度神经网络在三维模型分析上展现出其强大的特征表示和学习能力。相比于基于手工特征的传统分析方法,利用深度学习解决基于图像的三维模型检索问题能在较短时间内获取更准确的检索结果。

基于图像的三维模型检索方法为模型检索工作带来便利的同时也具有很多挑战:①如何对二维图像和三维模型提取完整有效的特征描述符,为后续的特征相似度度量工作打下良好的基础;②二维图像和三维模型是2种不同模态的数据,如何融合多模态数据,减小其间的语义差异,从而提高模型检索准确度。近期的研究工作通常利用深度学习将图像和三维模型特征映射到同一高维空间来完成跨域检索问题,并取得了不错的成果^[10]。但是这些工作大多是针对手绘草图作为输入来检索三维模型,而在实际生活中,人们更多接触的是带有背景、光照等噪声信息的真实图像,这些噪声信息同时也给检索准确度带来了新的挑战。相较于二维图像而言,三维模型数据包含更加丰富的信息,其特征提取工作也具有更大的难度。目前,用于跨域检索的主流特征提取方法是用一组多角度投影视图表示三维模型^[11],通过组合视图特征完成对三维模型的特征表示,但是如何减小视图表示三维模型过程中产生的特征损失依然是值得关注的问题。

针对基于图像检索三维模型问题的难点和已有工作的不足,本文设计异构的三元组网络提取图像和三维模型的特征表示向量。其中,三维模型采用一组顺序视图表示,通过距离度量构建特征联合嵌入空间,减小不同模态数据特征的分布差异,实现基于单幅RGB图像的三维模型检索。具体来说,跨域检索三元组网络包含2种结构的分支网络,分别对应于2种模态的输入数据。对于图像输入,设计基于注意力机制^[12]的图像精确特征提取分支网络,实现图像自适应特征细化学习,去除图像中的背景噪声,缩小真实图像与三维模型视图间的语义差异。对于三维模型,将其投影为一组顺序视图,并结合门控循环单元(Gate Recurrent Unit, GRU)^[13]设计三维模型顺序视图特征提取分支网络,在获取每个视图内容信息的同时得到视图间的空间信息,增强三维模型特征表示能力,提高跨域检索准确度。在获取到输入数据的特征表示后,即可通过三元组损失函数将不同模态的数据映射到特征联合嵌入空间,在此空间中,直接计算特征间距离即可度量不同模态数据的相似性。

分别在IM2MN^[14]和MI3DOR^[15]2个跨域检索数据集上进行实验,证明了本文方法的有效性,并提供了定量和定性的评估结果。总体来说,本文的主要贡献如下:

1) 提出了一种跨域检索三元组网络,用于图像与三维模型跨域检索。针对不同模态数据设计适应的特征提取分支网络,实现单幅真实图像到三维模型的端到端检索。

2) 将三维模型投影为一组顺序视图进行表示,设计三维模型顺序视图特征提取分支网络,在图像深度神经网络的基础上加入GRU结构,通过在获取每个视图信息的同时关注视图之间的联系,得到更加完整的三维模型全局特征描述符,从而提高跨域检索准确度。

3) 针对具有复杂背景的真实图像与投影视图之间的语义差异,将图像卷积神经网络(Convolutional Neural Networks, CNN)的注意力模块用于图像到三维模型的跨域检索任务,针对真实图像数据设计基于注意力机制的图像精确特征提取分支网络,使得网络在学习图像特征时更关注于对象信息,而不是背景信息,提高跨域检索准确度。

1 相关工作

1.1 基于实例的三维模型检索

目前的三维模型检索工作大多是基于实例的三维模型检索,这种方法需要提供一个待查询的三维模型,通过设计三维模型特征描述符提取方法,比较查询模型与模型库中三维模型的相似度完成检索过程,属于同域检索问题^[16-17]。在深度学习发展之前,三维模型的特征表示主要采用几何距离,如欧氏距离^[18]、测地距离^[19]等,还有基于谱形状分析的特征提取方法,如热核特征^[20]和波核特征^[21]等传统方法。随着深度学习的迅速发展,主流的三维模型检索研究利用CNN、自动编码器等深度学习网络模型提取三维形状特征,其可根据三维模型表达方式分为基于体素^[22]、网格^[23]、点云^[24]和视图^[25]的方法。体素表示方法较为简单,可以看作是二维像素到三维的扩展,缺点是需要消耗大量的计算和存储资源^[26]。网格由若干顶点和与其相连的边构成,是三维模型的重要原始表示形式,拥有比其他表达方式更为强大的描述能力,但是由于网格的复杂性和不规则性,目前的方法大多只对平滑的流形网格有效^[27]。点云数据是分布在三维空间中点的集合^[28],因获取方便而被广泛应用,但是点云数据的旋转性和不规则性给当前研究带来了较大的困

难。基于视图的方法用不同角度的投影视图表示三维模型,通过融合二维视图特征得到三维模型特征描述符用于检索任务。Su等^[29]提出了多视图卷积神经网络(Multi-View Convolutional Neural Networks, MVCNN)用于学习多视图表示的三维模型。在此基础上,Feng等^[30]提出了分组卷积神经网络(Group-View Convolutional Neural Networks, GVCNN),通过加入分组模块考虑了视图间的相似性和区分性,提取更完整的三维模型特征。考虑到同一平面视图组的局限性,Jiang等^[31]提出多环视卷积神经网络(Multi-Loop-View Convolutional Neural Networks, MLVCNN)模型,通过分层视图-循环-形状3个层次提取三维模型特征,完成模型检索任务。基于实例的三维模型检索研究中的特征提取方法也可以应用到跨域三维模型检索任务中来,鉴于投影视图与查询图像较为相近,采用基于视图的方法表示三维模型。

1.2 跨域三维模型检索

与基于实例的同域检索相比,跨域三维模型检索更适应于媒体数据多样化的发展趋势。Tasse和Dodgson^[32]构建了多模态三维模型检索系统,实现从文本、草图、深度图、真实图像和三维模型等多模态数据检索相似的三维模型。由于基于文本的输入通常过于笼统,无法精确表达目标三维形状,因此激发了对基于图像的三维模型检索研究。

一种方法是利用从图像中提取的手工特征或者学习特征训练分类器,该分类器为每个细粒度类提供三维模型^[33-34],但是这种方法的检索结果只选择训练过程中出现过的三维模型。为了突破此局限,近期研究的主流方法是将图像和三维模型映射到公共嵌入空间中,并通过距离度量来执行检索任务^[35]。Xiang等^[36]使用单个CNN将RGB图像和RGB渲染图像映射到嵌入空间,并采用基于欧氏距离的损失函数优化训练。Wang等^[37]将孪生网络用于基于草图的三维模型检索,其中一个分支作用于草图域,另一个分支作用于三维模型域,这种方法选择一个“最佳”投影视图表示三维模型,但实际上用单张视图表示三维模型会产生很大的损失,影响检索准确度。Li等^[38]提出了将图像和三维形状共同嵌入同一空间的框架,其中嵌入空间由通过AlexNet^[8]网络提取的图像特征与三维模型多视图的方向梯度直方图(Histogram of Oriented Gradient, HoG)特征之间的距离构建,但是由于需要计算所有训练三维模型对的距离,这种方法不可用于大规模三维形状数

据集。Dai等^[39]使用预训练的AlexNet^[8]对草图提取特征,并对三维模型提取具有局部约束线性编码(Locality-constrained Linear Coding, LLC)约束的SIFT特征,再对草图和三维模型得到的特征使用判别性损失函数和相关性损失函数进行度量学习。在后续工作中,Dai等^[40]在此基础上继续提出了DHML(Deep Correlated Holistic Metric Learning)方法,在隐藏层加入损失训练,以鼓励隐藏层中的功能也具有所需的特性,避免梯度消失,帮助网络稳定收敛,从而提高性能。鉴于三元组网络在图像分类检索领域取得了很好的成绩^[41],Lee等^[14]将其扩展到图像到三维模型的跨域检索领域中,使用2个CNN将RGB图像和投影视图组映射到嵌入空间,但是没有关注真实图像的有效特征提取,而且其提出的跨视图卷积方法会损失视图内容信息,影响检索准确度。

与以上方法不同,本文采用异构的三元组网络构建RGB图像与三维模型的特征联合嵌入空间,并分别设计基于注意力机制的图像精确特征提取算法和基于顺序视图的三维模型特征提取算法,提高不同模态数据特征描述符的辨别力,从而提高跨域检索准确度。

2 基于三元组网络的跨域数据检索方法

本文提出一种采用三元组深度神经网络匹配不同模态数据特征的方法,用于基于单幅RGB图像的三维模型检索,图1给出了本文方法的详细框架。首先,对图像数据和三维模型数据进行预处理,将三维模型投影为一组顺序视图。然后,将图像数据和视图组数据以三元组的形式输入到网络中进行训练,跨域检索三元组网络由2种网络结构组成:一种是用于处理二维图像数据的图像精确特征提取分支网络,另一种是用于视图组数据的三维模型顺序视图特征提取分支网络。最后,在获取到数据的特征向量后,利用三元组损失构建特征联合嵌入空间,在此空间上即可根据距离度量跨域数据相似性,完成检索任务。

2.1 数据预处理

图像与三维模型跨域检索数据集由RGB图像和三维模型组成。先将图像数据统一处理为 256×256 ,通过多尺度数据增强,随机裁剪成 227×227 的图片,防止模型过拟合。再将RGB图像转换为灰度图,去除图像颜色信息对检索的干扰。

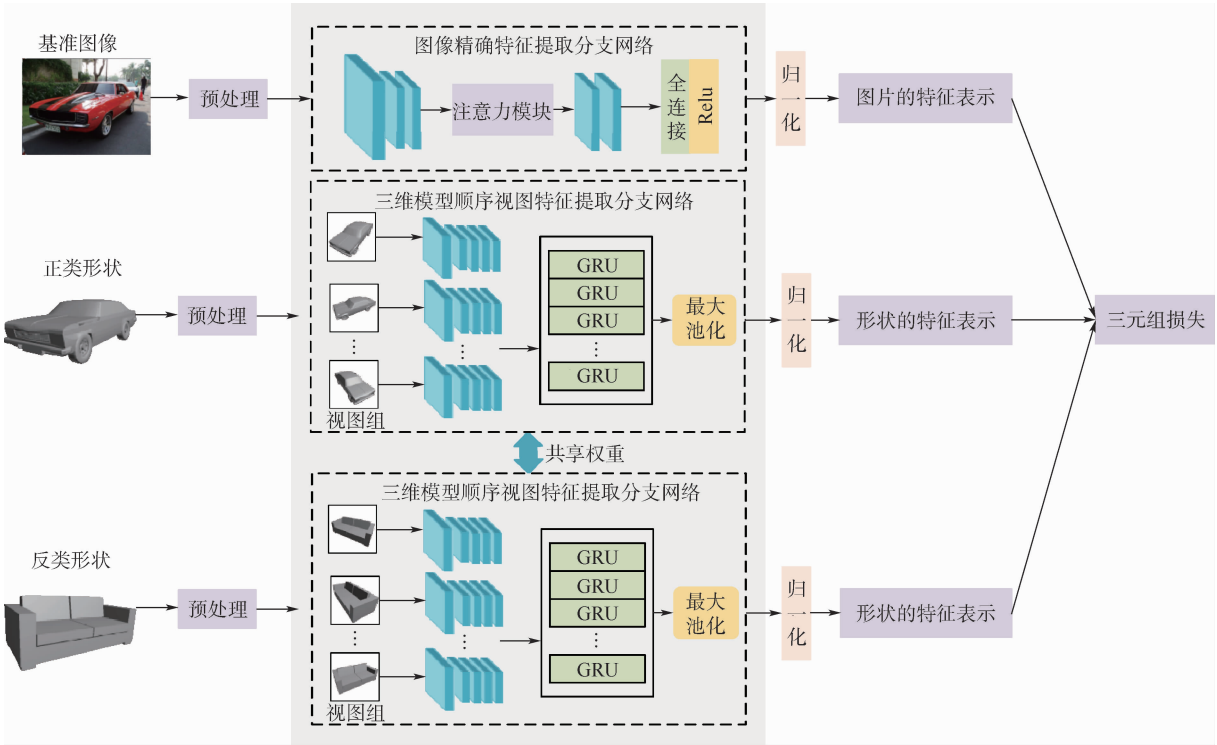


图 1 跨域检索三元组网络架构

Fig. 1 Architecture of cross-domain retrieval triplet network

对三维模型数据进行顺序投影处理,在以模型为中心的圆上均匀设置虚拟相机阵列,从而捕获三维模型的顺序视图组。这种视图获取方法简单有效,被广泛应用于三维模型预处理任务中^[27,31]。在投影视图数量的选取上,鉴于 12 个不同角度的投影视图被证明可以很好地表达三维模型,并且在三维模型的分类、检索等任务中取得了不错的效果^[29-30],设置相机阵列包含 12 个虚拟相机,它们放置在三维模型周围的圆平面上,指向三维模型质心,与水平面呈 30°夹角,相机间隔为 30°。第一个视图的视角可以选在圆上任意一个固定位置,再以顺时针方向获取后续的视图,由此组成三维模型的顺序视图组。

2.2 三元组网络结构

数据预处理完成后,将其中的图像数据和三维模型的视图组数据以三元组的形式输入到跨域检索三元组网络中,如图 1 所示。三元组数据格式为:(基准图像-正类视图组-反类视图组)。其中,基准图像来自图像数据集,正类视图组表示的三维模型与基准图像中的对象同类,反类视图组则与基准图像异类。对应于输入数据,跨域检索三元组网络有 3 个分支网络。分支 1 为图像精确特征提取分支网络,其输入为具有复杂背景的基准图像,经过 CNN 完成对图像重要特征自适应细化学习。分支 2 和分支

3 为三维模型顺序视图特征提取分支网络,输入分别为正类模型和反类模型的视图组,其网络结构相同,权值共享,在提取每个视图特征的同时关注不同视图之间的联系,获得三维模型的特征描述符。由于基准图像与视图均属于二维图像,因此 3 个分支网络均以在 ImageNet 数据集上预训练好的 VGG19^[42]网络作为基础网络提取基准图像特征和三维模型的视图级特征,如图 1 中的卷积块所示。

2.2.1 图像精确特征提取分支网络

真实图像包含与检索无关的复杂背景信息,这些噪声会对检索准确度产生负面影响。而且,在通过特征学习构建跨域特征联合嵌入空间时,真实图像与“干净”的投影图像之间的语义鸿沟也会导致检索准确度降低。文献[12]通过叠加全连接层和 Relu 层,对提取到的图像特征执行仿射变换来缩小真实图像与投影图像的域差异。在此基础上,笔者关注于在特征提取阶段实现图像特征自适应细化学习。

通过加入注意力机制实现对基准图像有效特征的精确提取。注意力机制的本质是定位特征中的感兴趣信息,并抑制其中的无用信息。本文主要关注于对基准图像的有效区域特征提取,因为对于图像与三维模型的跨域检索任务来说,并非图像中的所有区域对任务的贡献都是同等重要

的,只有与被检索对象相关的像素才是特征提取网络的重点学习区域。因此,选择空间注意力模型来降低图像背景信息对检索任务的影响,从而完成基准图像精确特征提取。

图像精确特征提取分支网络以 VGG19 网络为基础,VGG19 网络由 16 个卷积层和 3 个全连接层组成,这 16 个卷积层被分成 5 段,段间采用一个最大池化层来缩小图片尺寸。在每一段的每 2 个卷积层之间加入空间注意力模块^[12],实现对图像中重要对象的特征提取。由于 VGG19 网络结构较长,在图 1 中仅显示一个注意力模块作为示例。

空间注意力模块利用特征在空间维度的关系生成空间注意力权重图,用于在图像信息中找出需要被关注的区域。如图 2 所示,对于输入特征 $F_1 \in \mathbb{R}^{H \times W \times C}$ (H 为高度, W 为宽度, C 为通道数量),空间注意力模块在通道维度上对其执行最大池化 (Max-pooling) 和平均池化 (Average-pooling),得到 2 个 $H \times W \times 1$ 的空间描述符。采用 2 种池化方法的好处在于:最大池化可以更多地保留纹理信息,而平均池化对特征图上的每一个像素点都有反馈,可以在减少维度的同时完整传递信息。池化完成后,按照通道维度进行拼接得到通道数为 2 的特征描述符,应用一个 7×7 的卷积层和 Sigmoid 激活函数生成 $H \times W \times 1$ 的空间注意力权重图 A_s 。经过区域过滤得到的新特征 $F'_1 \in \mathbb{R}^{H \times W \times C}$ 即可参与接下来的特征提取过程,最终得到图像的精确特征。

$$F'_1 = A_s(F_1) \otimes F_1 \tag{1}$$

$$A_s(F_1) = \varphi(g^{7 \times 7}([\text{MaxPool}(F_1); \text{AvgPool}(F_1)])) \tag{2}$$

式中: $\otimes$ 表示逐个元素相乘; φ 表示 Sigmoid 函数; g 表示卷积计算。

通过加入空间注意力模块,图像精确特征提取分支网络在学习图像特征时更关注于对象信息,而非背景信息。采用文献[12]中由全连接层和 Relu 层叠加组成的特征自适应层,对获取到的图像特征进行仿射变换,进一步消除背景信息,从而得到图像的精确特征。

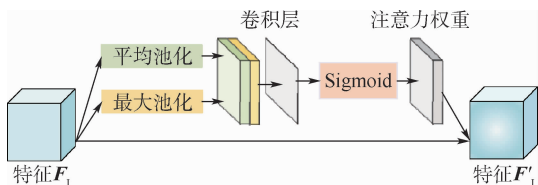


图 2 注意力模块详细结构

Fig. 2 Detailed structure of attention module

2.2.2 三维模型顺序视图特征提取分支网络

由于视图与图像相近,多视图方法可以在特征提取部分减小图像与三维模型的语义差异,因此在跨域检索任务中具有良好的效果。但是,用视图表示三维模型的过程会产生较多信息损失,因此,如何减少特征损失,提取三维模型更完整有效的特征仍然是一个困难的问题。相比于传统 CNN,循环神经网络 (Recurrent Neural Network, RNN) 具有强大的记忆功能,善于处理序列信息。GRU 网络是 RNN 最成功的变体,其通过加入门控单元解决了长序列训练过程中的梯度爆炸和梯度消失问题,具有更强大的记忆功能。因此,通过 GRU 结构聚合顺序视图,在聚合每个视图内容信息的同时保留视图序列的空间信息,获取更具有辨别力的三维模型特征描述符。

三维模型顺序视图特征提取分支网络由 VGG19 网络和 GRU 网络组成。对于输入的顺序视图组 V ,先将 VGG19 网络作用于其中的每个顺序视图 v_k ($k = 1, 2, \dots, 12$),获取它们的低级特征 f_k ($k = 1, 2, \dots, 12$),再采用 GRU 网络按照视图序列逐步聚合视图级特征。如图 3 所示,以第 j 时间步为例,视图 v_k 的低级特征向量 f_j 被输入到 GRU 中,其中隐状态 h_j 作为“记忆”储存学习过的信息。在重置门控 r_j 的控制下,GRU 利用 f_j 与前一个时间步的隐状态 h_{j-1} 生成重置数据 h'_j 。

$$r_j = \varphi(W_r \cdot [h_{j-1}, f_j]) \tag{3}$$

$$h'_j = \tanh(W \cdot [r_j * h_{j-1}, f_j]) \tag{4}$$

式中: W_r 和 W 为权重矩阵; $\tanh$ 为激活函数。

通过更新门控 z_j 有针对性地将 h'_j 中的信息添加到隐状态 h_{j-1} ,从而学习到了新的“记忆”,生成新的隐状态 h_j 。

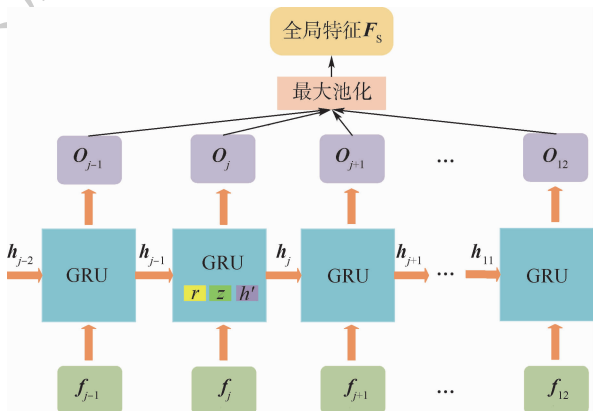


图 3 GRU 网络聚合视图级特征

Fig. 3 Aggregation of view-level features using GRU networks

$$\mathbf{z}_j = \varphi(\mathbf{W}_z \cdot [\mathbf{h}_{j-1}, \mathbf{f}_j]) \quad (5)$$

$$\mathbf{h}_j = \mathbf{z}_j \circ \mathbf{h}_{j-1} + (1 - \mathbf{z}_j) \circ \mathbf{h}'_j \quad (6)$$

式中: $\mathbf{W}_z$ 为更新门控对应的权重矩阵; “ $\circ$ ” 表示 Hadamard 乘积。

更新“记忆”后, GRU 即可将隐状态 $\mathbf{h}_j$ 向下一个步骤传递并经过维度变换输出为 $\mathbf{O}_j$ 。

$$\mathbf{O}_j = \mathbf{W}_o \mathbf{h}'_j + b \quad (7)$$

式中: $\mathbf{W}_o$ 和 b 分别为可学习的权重和偏置参数。

在 GRU 完成逐步学习后, 采用最大池化层聚合每个视图对应的 GRU 隐状态的输出 $\mathbf{O}_k (k=1, 2, \dots, 12)$, 得到三维模型的全局特征 $\mathbf{F}_s$ 。

2.3 构建特征联合嵌入空间

通过图像精确特征提取分支网络和三维模型顺序视图特征提取分支网络, 三元组网络的 3 个分支分别获得基准图像、正类三维模型和反类三维模型的特征表示向量, 采用 L_2 正则化函数分别对其进行标准化, 从而控制模型复杂度, 防止过拟合。

$$\mathbf{x} = \frac{\mathbf{x}}{\max(\|\mathbf{x}\|_2, \varepsilon)} \quad (8)$$

式中: $\mathbf{x}$ 为特征向量中的元素; ε 取值为 1×10^{-12} 。

利用三元组损失函数对图像特征和三维模型特征构建联合嵌入空间, 通过控制特征距离将不同模态的数据特征映射到同一高维空间中。具体来说, 网络训练的目的是使基准图像特征 $\mathbf{f}(X_i^a)$ 与反类三维模型特征 $\mathbf{f}(X_i^n)$ 的距离大于它和正类三维模型特征 $\mathbf{f}(X_i^p)$ 的距离, 使得域间数据特征差异减小的同时, 域内同类数据相近, 异类数据疏远。损失函数定义如下:

$$d_p = \|\mathbf{f}(X_i^a) - \mathbf{f}(X_i^p)\|_2^2 \quad (9)$$

$$d_n = \|\mathbf{f}(X_i^a) - \mathbf{f}(X_i^n)\|_2^2 \quad (10)$$

$$\text{Loss} = \max(\|d_p - d_n + \text{margin}\|, 0) \quad (11)$$

式中: d_p 为基准图像与正类三维模型的特征距离; d_n 为基准图像与反类三维模型的特征距离; margin 为避免模型在训练过程中走捷径导致错误结果而设定的相对距离。

在特征联合嵌入空间中, 图像数据与三维模型数据特征按照类别呈簇状分布, 通过欧氏距离等方法直接计算特征间距离即可度量不同模态数据的相似性。

3 实验

检索过程使用跨域检索三元组深度网络, 对查询图像和模型库中的三维模型同时提取特征并构建特征联合嵌入空间, 特征在此空间中的距离代表数据的相似度, 利用欧氏距离计算查询图像

与三维模型之间的特征距离并排序, 选择前 5 个三维模型作为检索结果。本文以 CentOS7.3 为工作平台, 在一个 Intel Xeon E5-2609 v4 CPU 和两个 NVIDIA GTX1080 GPU 上采用 TensorFlow 深度学习框架实现模型。采用 Adam 算法优化网络, 设置三元组损失函数中 margin 为 0.2, 网络学习率为 1×10^{-6} , 衰减因子设为 0.1。

3.1 数据集

分别在 SHREC 2019 提供的 MI3DOR 数据集和 IM2MN 数据集上进行实验来验证本文方法。其中, MI3DOR 数据集有 21 个类, 包含 7 690 个三维模型和 21 000 个 RGB 图像, 其中, 训练集包括 3 842 个三维模型和 10 500 个图像, 测试集包括 3 848 个三维模型和 10 500 个图像。IM2MN 数据集包含 12 311 个三维模型和 10 000 个 RGB 图像, 共分为 40 个类, 其中, 训练集包括 9 843 个三维模型和 8 000 个图像, 测试集包括 2 468 个三维模型和 2 000 个图像。由于本文的三维模型预处理方法与 IM2MN 数据集中的视图子数据集生成方法一致, 所以直接将其用于视图特征提取网络。相比于 MI3DOR 数据集, IM2MN 数据集涉及的类别范围更广, 而且包含较多容易混淆的数据, 因此更具有挑战性。

3.2 消融实验

为了验证提出的基于注意力机制的图像精确特征提取算法和基于 GRU 网络的三维模型顺序视图特征提取算法对跨域检索的有效性, 分别在 IM2MN 数据集和 MI3DOR 数据集上做消融实验, 其中检索性能通过平均准确率 (mean Average Precision, mAP) 进行评估, 如表 1 和表 2 所示。针对网络是否包含注意力模块和 GRU 网络模块, 分别设计不同的网络结构进行测试。其中, 图像精确特征提取分支网络在采用文献 [14] 所提出的特征自适应层的基础上, 加入空间注意力模块, 进一步减少图像中的背景噪声对检索任务的影响, 缩小跨域数据特征间的语义差异。为区分注意力模块与特征自适应层对提升检索结果的贡献程度, 通过控制是否包含特征自适应层模块变量, 进一步体现各模块对跨域检索任务的影响。从表 1、表 2 可以看出, 注意力模块和 GRU 网络模块均对检索准确度产生了积极的影响, 仅加入注意力模块使得检索平均准确率提高 3.19% ~ 4.08%, 仅加入 GRU 网络模块使得检索平均准确率提高 5.57% ~ 5.74%, 而同时包含 2 种结构的网络模型的检索平均准确率最高, 说明了本文方法的有效性。

表 1 IM2MN 数据集消融实验测试结果

Table 1 Test results of ablation experiment on IM2MN dataset

自适应层	注意力模块	GRU	mAP/%
无	无	无	42.16
有	无	无	48.74
有	有	无	51.93
有	无	有	54.48
有	有	有	55.65

表 2 MI3DOR 数据集消融实验测试结果

Table 2 Test results of ablation experiment on MI3DOR dataset

自适应层	注意力模块	GRU	mAP/%
无	无	无	42.78
有	无	无	49.67
有	有	无	53.75
有	无	有	55.24
有	有	有	56.53

3.3 基于图像的三维模型跨域检索

为验证本文方法的有效性,在 IM2MN 数据集上与 CDTNN^[14]算法、MVCNN^[29]算法和基于三元组网络的增强 MVCNN 算法的检索结果进行定量比较,检索性能通过平均准确率 mAP 进行评估,测试结果如表 3 所示。其中,基于三元组网络的增强 MVCNN 算法是在基础三元组网络结构的基础上,利用 MVCNN 算法提取三维模型特征并作用于网络的正类和负类分支上得到检索结果。

从表 3 中可以发现,CDTNN 算法在基线方法中的检索平均准确率最高,而本文方法的检索平均准确率较 CDTNN 算法提高 2.98%,达到 55.65%。为验证本文方法在其他公开数据集上的有效性,在 MI3DOR 数据集上与 CDTNN 基线方法进行检索平均准确率评估,相较于 MI3DOR 数据集,IM2MN 数据集中包含更多的类别和容易产生混淆的数据,因此其检索准确度略微低一些。如表 3 所示,本文方法在 MI3DOR 数据集上检索平均准确率较 CDTNN 算法提高 3.05%,达到 56.53%。

表 3 基于图像的三维模型检索性能

Table 3 Performance for image-based 3D model retrieval

数据集	方法	mAP/%
IM2MN	MVCNN ^[29]	7.92
	三元组 + MVCNN ^[14]	40.85
	CDTNN ^[14]	52.67
	本文	55.65
MI3DOR	CDTNN ^[14]	53.48
	本文	56.53

图 4 可视化地展示了一些检索结果示例。其中,图 4 的顶部展示了一些成功的检索结果,底部则展示了 2 种典型的检索失败情况,其中错误的检索结果用方框圈出。从图 4 中可以看出,本文方法在处理单幅图像到三维模型的跨域检索任务上有不错的效果,但同时也存在一些失败情况,如多目标场景(显示器与桌子共同出现在查询图像中),图像中的检索对象不完整及存在遮挡物体(查询图像中书架没有被完整显示,并且被大量物体遮盖,导致其与桌子类别产生混淆),均会对检索产生负面影响。

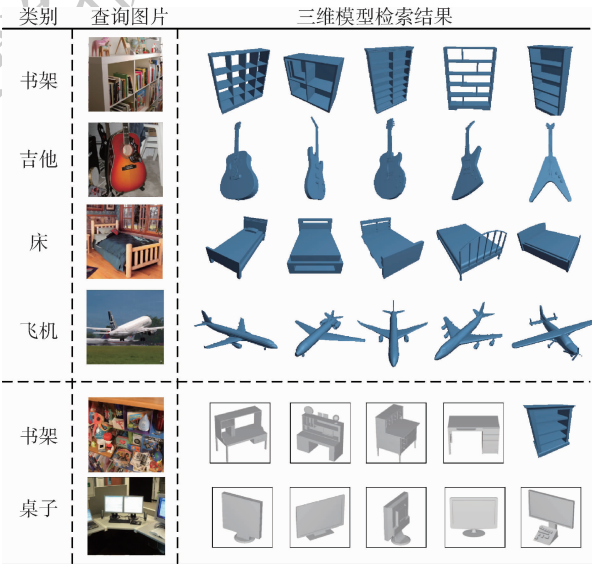


图 4 基于单张图像的三维模型检索结果示例

Fig. 4 Examples of monocular image based 3D model retrieval results

4 结 论

1) 本文提出了跨域检索三元组网络来解决基于单幅图像的三维模型检索任务,为提高检索准确度,针对网络的不同输入数据设计图像精确特征提取分支网络和三维模型顺序视图特征提取分支网络,提高从不同模态数据学习得到的特征描述符的适配能力。通过对比实验可以看出,相较于最优基线方法,本文方法在 2 个数据集上的检索平均准确率均提高 3% 左右,体现了本文方法的有效性。

2) 为了验证提出的图像精确特征提取算法和三维模型顺序视图特征提取算法对提高跨域检索准确度的有效性,设置了消融实验针对不同网络情况进行测试。从实验结果可以看出,2 种特征提取算法均对检索有积极的作用。

3) 本文方法对多目标场景和存在遮挡物体

的图像检索可能会出现失败的情况,而这类图像在实际生产应用中也较为常见。因此,在后续工作中,可以考虑结合图像识别等领域的知识对算法作进一步改进。

参考文献 (References)

- [1] BU S H, WANG L, HAN P C, et al. 3D shape recognition and retrieval based on multi-modality deep learning[J]. *Neurocomputing*, 2017, 259: 183-193.
- [2] 蔡铁珩, 王雪艳, 胡绍斌, 等. 基于多源图像弱监督学习的 3D 人体姿态估计[J]. *北京航空航天大学学报*, 2019, 45(12): 2375-2384.
CAI Y H, WANG X Y, HU S B, et al. Three-dimensional human pose estimation based on multi-source image weakly-supervised learning[J]. *Journal of Beijing University of Aeronautics and Astronautics*, 2019, 45(12): 2375-2384 (in Chinese).
- [3] GIRDHAR R, FOUHEY D F, RODRIGUEZ M, et al. Learning a predictable and generative vector representation for objects[C]// *European Conference on Computer Vision*. Berlin: Springer, 2016: 484-499.
- [4] TULSIANI S, GUPTA S, FOUHEY D F, et al. Factoring shape, pose, and layout from the 2d image of a 3d scene[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2018: 302-310.
- [5] IYER N, JAYANTI S, LOU K, et al. Three-dimensional shape searching: State-of-the-art review and future trends[J]. *Computer-Aided Design*, 2005, 37(5): 509-530.
- [6] XIE J, FANG Y, ZHU F, et al. Deepshape: Deep learned shape descriptor for 3d shape matching and retrieval[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2015: 1275-1283.
- [7] MIAN A S, BENNAMOUN M, OWENS R A. Matching tensors for pose invariant automatic 3D face recognition[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2005: 120.
- [8] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]// *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, 2012: 1097-1105.
- [9] 杨思晨, 王华锋, 王月海, 等. 深度学习机制与小波融合的超分辨率重建算法[J]. *北京航空航天大学学报*, 2020, 46(1): 189-197.
YANG S C, WANG H F, WANG Y H, et al. Super-resolution reconstructing algorithm based on deep learning mechanism and wavelet fusion[J]. *Journal of Beijing University of Aeronautics and Astronautics*, 2020, 46(1): 189-197 (in Chinese).
- [10] GRABNER A, ROTH P M, LEPETIT V. Location field descriptors: Single image 3D model retrieval in the wild[C]// *Proceedings of the 2019 International Conference on 3D Vision (3DV)*. Piscataway: IEEE Press, 2019: 583-593.
- [11] WU Z Z, ZHANG Y H, ZENG M, et al. Joint analysis of shapes and images via deep domain adaptation[J]. *Computers & Graphics*, 2018, 70: 140-147.
- [12] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]// *European Conference on Computer Vision*. Berlin: Springer, 2018: 3-19.
- [13] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[EB/OL]. (2014-06-03) [2020-02-25]. <https://arxiv.org/abs/1406.1078>.
- [14] LEE T, LIN Y L, CHIANG H Y, et al. Cross-domain image-based 3D shape retrieval by view sequence learning[C]// *Proceedings of the 2018 International Conference on 3D Vision (3DV)*. Piscataway: IEEE Press, 2018: 258-266.
- [15] LI W, LIU A, NIE W Z, et al. SHREC 2019-Monocular image based 3D model retrieval[EB/OL]. (2019-01-28) [2020-02-25]. <https://www.iti-tju.org/MI3DOR19/>.
- [16] FANG Y, XIE J, DAI G, et al. 3D deep shape descriptor[C]// *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE Press, 2015: 2319-2328.
- [17] 李海生, 武玉娟, 郑艳萍, 等. 基于深度学习的三维数据分析理解方法研究综述[J/OL]. *计算机学报*, 2019, 42: 1-25. (2019-07-09) [2020-02-21]. <http://kns.cnki.net/kcms/detail/11.1826.TP.20190709.1509.002.html>.
LI H S, WU Y J, ZHENG Y P, et al. A survey of 3D data analysis and understanding based on deep learning[J/OL]. *Chinese Journal of Computers*, 2019, 42: 1-25. (2019-07-09) [2020-02-21]. <http://kns.cnki.net/kcms/detail/11.1826.TP.20190709.1509.002.html> (in Chinese).
- [18] OSADA R, FUNKHOUSER T, CHAZELLE B, et al. Shape distributions[J]. *ACM Transactions on Graphics (TOG)*, 2002, 21(4): 807-832.
- [19] MAHMOUDI M, SAPIRO G. Three-dimensional point cloud recognition via distributions of geometric distances[J]. *Graphical Models*, 2009, 71(4): 22-31.
- [20] SUN J, OVSIJANIKOV M, GUIBAS L. A concise and provably informative multi-scale signature based on heat diffusion[J]. *Computer Graphics Forum*, 2009, 28(5): 1383-1392.
- [21] AUBRY M, SCHLICKWEI U, CREMERS D. The wave kernel signature: A quantum mechanical approach to shape analysis[C]// *Proceedings of 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. Piscataway: IEEE Press, 2011: 1626-1633.
- [22] WANG P S, SUN C Y, LIU Y, et al. Adaptive O-CNN: A patch-based deep representation of 3D shapes[J]. *ACM Transactions on Graphics (TOG)*, 2018, 37(6): 1-11.
- [23] FENG Y, FENG Y, YOU H, et al. MeshNet: Mesh neural network for 3D shape representation[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*. Palo Alto: AAAI Press, 2019: 8279-8286.
- [24] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[C]// *Advances in Neural Information Processing Systems*. Cambridge: MIT Press, 2017: 5099-5108.
- [25] HUANG H B, KALOGERAKIS E, CHAUDHURI S, et al. Learning local shape descriptors from part correspondences with multiview convolutional networks[J]. *ACM Transactions on Graphics (TOG)*, 2017, 37(1): 1-14.

- [26] WANG P S, LIU Y, GUO Y X, et al. O-CNN: Octree-based convolutional neural networks for 3d shape analysis [J]. ACM Transactions on Graphics (TOG), 2017, 36(4): 1-11.
- [27] HAN Z, SHANG M, LIU Z, et al. SeqViews2SeqLabels: Learning 3D global features via aggregating sequential views by RNN with attention [J]. IEEE Transactions on Image Processing, 2018, 28(2): 658-672.
- [28] LAN S Y, YU R C, YU G, et al. Modeling local geometric structure of 3D point clouds using Geo-CNN [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 998-1008.
- [29] SU H, MAJI S, KALOGERAKIS E, et al. Multi-view convolutional neural networks for 3d shape recognition [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 945-953.
- [30] FENG Y, ZHANG Z, ZHAO X, et al. GVCNN: Group-view convolutional neural networks for 3D shape recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 264-272.
- [31] JIANG J, BAO D, CHEN Z, et al. MLVCNN: Multi-loop-view convolutional neural network for 3D shape retrieval [C] // Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019, 33: 8513-8520.
- [32] TASSE F P, DODGSON N. Shape2Vec: Semantic-based descriptors for 3D shapes, sketches and images [J]. ACM Transactions on Graphics (TOG), 2016, 35(6): 1-12.
- [33] AUBRY M, MATURANA D, EFROS A A, et al. Seeing 3d chairs: Exemplar part-based 2d-3d alignment using a large dataset of cad models [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 3762-3769.
- [34] MOTTAGHI R, XIANG Y, SAVARESE S. A coarse-to-fine model for 3d pose estimation and sub-category recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 418-426.
- [35] KULIS B. Metric learning: A survey [J]. Foundations and Trends in Machine Learning, 2013, 5(4): 287-364.
- [36] XIANG Y, KIM W, CHEN W, et al. ObjectNet3D: A large scale database for 3d object recognition [C] // European Conference on Computer Vision. Berlin: Springer, 2016: 160-176.
- [37] WANG F, KANG L, LI Y. Sketch-based 3d shape retrieval using convolutional neural networks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1875-1883.
- [38] LI Y Y, SU H, QI C R, et al. Joint embeddings of shapes and images via CNN image purification [J]. ACM Transactions on Graphics (TOG), 2015, 34(6): 1-12.
- [39] DAI G, XIE J, ZHU F, et al. Deep correlated metric learning for sketch-based 3d shape retrieval [C] // Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2017: 4002-4008.
- [40] DAI G X, XIE J, FANG Y. Deep correlated holistic metric learning for sketch-based 3d shape retrieval [J]. IEEE Transactions on Image Processing, 2018, 27(7): 3374-3386.
- [41] SCHROFF F, KALENICHENKO D, PHILBIN J. FaceNet: A unified embedding for face recognition and clustering [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 815-823.
- [42] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [EB/OL]. (2014-09-04) [2020-02-25]. <https://arxiv.org/abs/1409.1556>.

作者简介:

杜雨佳 女, 硕士研究生。主要研究方向: 计算机图形学。

李海生 男, 博士, 教授, 博士生导师。主要研究方向: 计算机图形学。

姚春莲 女, 博士, 副教授。主要研究方向: 视频、图像处理、嵌入式系统设计。

蔡强 男, 博士, 教授, 博士生导师。主要研究方向: 计算机图形学。

Monocular image based 3D model retrieval using triplet network

DU Yujia^{1,2,3}, LI Haisheng^{1,2,3,*}, YAO Chunlian^{1,2,3}, CAI Qiang^{1,2,3}

- (1. School of Computer and Information Engineering, Beijing Technology and Business University, Beijing 100048, China;
- 2. National Engineering Laboratory For Agri-product Quality Traceability, Beijing 100048, China;
- 3. Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing 100048, China)

Abstract: With the diversified development of media data, the cross-domain retrieval between images and 3D models becomes a new challenge for 3D model retrieval. In view that images and 3D models are extremely different and hard to match, a cross-domain retrieval algorithm based on triple network is proposed to construct a joint embedding space for real images and 3D shapes in an end-to-end manner. Then the similarity between different modal data could be effectively computed by the distance in the space, leading to accurate retrieval of similar 3D models from single image. In order to improve the accuracy of cross-domain retrieval, the 3D model was represented by a set of sequential views, and the Gate Recurrent Unit (GRU) was utilized for view-level features to generate the global feature. In addition, an attention mechanism was introduced to extract image features and bridge the semantic gaps between the real image and the rendered 3D views. Experimental results show that the mean average precision can be improved by at least 2.98% – 3.05% on two cross-domain datasets compared with other similar algorithms.

Keywords: 3D model retrieval; deep learning; cross-domain retrieval; triplet network; Gate Recurrent Unit (GRU); attention mechanism

Received: 2020-02-28; **Accepted:** 2020-03-28; **Published online:** 2020-04-17 12:56
URL: kns.cnki.net/kcms/detail/11.2625.V.20200417.1206.003.html
Foundation items: National Natural Science Foundation of China (61877002); Beijing Natural Science Foundation and Fengtai Rail Transit Frontier Research Joint Fund (L191009); Beijing Municipal Education Commission Research Team Construction Project (PXM2019_014213_000007)
*** Corresponding author.** E-mail: lihsh@bttu.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0061

基于结构加权低秩近似的泊松图像去模糊

吴庆波^{1,2}, 任文琦^{1,2,*}

(1. 中国科学院信息工程研究所 信息安全国家重点实验室, 北京 100193;

2. 中国科学院大学 网络空间安全学院, 北京 100049)



摘 要: 针对由高斯模糊和泊松噪声引起的图像降质问题, 提出了一种基于结构加权低秩近似的图像去模糊方法。首先, 通过依次组合缩放、旋转、剪切和翻折等四种基本操作引入结构变换, 以增加搜索空间内候选图像块的相似性。然后, 构造新的目标函数, 利用相似图像块的低秩性, 在正则项中使用加权核范数(WNN)对结构变换后的图像块进行惩罚, 以在去模糊的同时抑制泊松噪声。最后, 基于半正定二次分裂(HQS)方法设计交替优化方案, 用于求解目标函数, 从泊松图像中去除模糊。实验结果表明: 在多种泊松噪声强度下, 所提方法取得的峰值信噪比(PSNR)和结构相似性(SSIM)都高于当前同类去模糊方法。

关 键 词: 图像去模糊; 泊松噪声; 结构变换; 加权低秩近似; 半正定二次分裂(HQS)

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2020)09-1701-10

当今, 视频监控技术在社会治安管理中起着非常重要的作用, 尤其是在夜间等人们不常活动的时间段。然而, 成像设备固有的点扩散效应往往导致视频帧(或图像)中存在不同程度的模糊。同时, 夜间有限的光照又不可避免地在图像中引起泊松噪声污染。这些因素通常会降低原始图像的质量, 造成图像结构信息受损, 给后续图像水印和图像取证等任务^[1-3]增加了困难。因此, 如何有效地从泊松图像中去除模糊和抑制噪声已经成为图像处理领域的重要研究课题。

为从泊松图像中去除模糊, 现有工作^[4-5]把降质图像的形成过程表示为数学模型: $y = \mathcal{P}(Hx)$, y , x , H 和 $\mathcal{P}$ 分别为观测图像、清晰图像、模糊核和泊松噪声添加算子。因此, 泊松图像去模糊问题归结为从观测图像 y 中恢复出清晰图像 x , 这是一个典型的不适定问题。为克服问题的不适定性, 大部分泊松图像去模糊方法^[6-10]建立

在变分模型的基础上, 先用正则化理论将不适定问题转化为适定问题, 再借助优化技术恢复清晰图像。相关工作中的模型可以概括为: $\min_x D(x, y) + R(x)$, 函数 $D(x, y)$ 和 $R(x)$ 分别为数据项(Data Term)和正则项(Regularization Term)。数据项 $D(x, y)$ 用于约束复原图像的内容与观测图像接近, 防止恢复结果失真, 主要来自泊松概率密度函数^[11]: $f(y|Hx) = \prod_n \frac{e^{-[Hx]_n} ([Hx]_n)^{y_n}}{y_n!}$, n 为模糊图像中像素的索引, $[Hx]_n$ 和 y_n 分别为向量 Hx 和 y 中的第 n 个元素。正则项 $R(x)$ 通常基于某种先验假设对清晰图像进行惩罚, 以克服逆问题的不适定性, 所以在很大程度上依赖于人们对清晰图像的先验知识(如平滑和低秩)。

现有泊松图像去模糊方法^[12-16]基于不同的先验知识引入了多种正则项。Harmany 等^[13]假设清晰图像是平滑的, 在正则项中使用全变分

收稿日期: 2020-03-02; 录用日期: 2020-03-20; 网络出版时间: 2020-04-01 09:38

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200331.1722.002.html

基金项目: 国家自然科学基金(U1605252, U1803264, 61802403); 国家重点研发计划(2019YFB1406500); 北京市自然科学基金(L182057, KZ201910005007, L182057)

* 通信作者: E-mail: renwenqi@iie.ac.cn

引用格式: 吴庆波, 任文琦. 基于结构加权低秩近似的泊松图像去模糊[J]. 北京航空航天大学学报, 2020, 46(9): 1701-1710.
WU Q B, REN W Q. Structural weighted low-rank approximation for Poisson image deblurring[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1701-1710 (in Chinese).

(Total Variation, TV)^[17]抑制泊松噪声。全变分方法虽然能够保持物体的主要边缘,但在非平滑区域表现不佳,往往导致恢复结果中纹理细节的过平滑。为避免全变分的缺陷,Chen^[15]在小波域使用 L_1 -范数约束清晰图像的小波系数(Wavelet Coefficients),虽然恢复结果中的纹理细节有所改善,但是该方法采用固定的字典,仍然会导致恢复结果中存在模糊细节。为进一步改进去模糊结果的质量,Ma等^[14]通过字典学习(Dictionary Learning)^[18]的方法从观测图像中获得与图像内容相适应的基元,并借助学习到的字典从泊松图像中去除模糊,但是该方法在字典学习阶段比较耗时。为克服该问题,Buades等^[19]认为清晰图像具有非局部自相似性,并基于此假设提出了非局部均值(Nonlocal Means, NLM)方法,从输入图像中去除高斯噪声。随后,研究人员^[20-21]基于相似图像块的低秩假设使用加权核范数最小化(WNNM)方法^[20]去除加性高斯噪声。但是,这些方法^[19-21]仅仅通过滑动(平移)窗口在正视角度下搜索相似图像块。然而,实际中获得的图像常常含有大量的透视场景^[22]。因此,对候选图像块进行合适的变换可以在一定程度上避免现有方法在正视角度下搜索相似图像块的局限。

为得到更高质量的复原图像,本文提出了基于结构加权低秩近似的泊松图像去模糊模型。在该模型中,首先,引入结构变换(Structural Transformation, ST),通过增加候选图像块的相似性在透视场景中更好地利用图像的非局部自相似性。然后,构造新的目标函数,基于相似图像块的低秩假设在正则项中使用加权核范数(WNN)^[20]对结构变换后的相似图像块进行惩罚,从而使模型在去除高斯模糊和泊松噪声的同时有利于保持物体的结构信息。由于本文模型是具有可分性的非光滑优化问题,提出使用半正定二次分裂(Half-Quadratic Splitting, HQS)方法有效地求解目标函数。相关工作^[22-24]在图像去噪和超分辨任务中也使用了某些结构变换,但是所采用的两步策略相对简单。相反,本文方法建立在统一的目标函数上,并借助优化技术进行泊松图像去模糊。超分辨方法^[25]在残差块(Residual Block)中嵌入空间特征变换(SFT),以避免SRMD方法^[26]在输入端串联模糊核和图像引起的与图像无关的干扰。SFT只含有缩放和平移2种操作,而本文提出的结构变换组合了4种基本操作,以提供更相似的图像块。

1 结构变换

除了传统的平移操作(滑动窗口)^[19-21],本文引入另外4种基本操作,即旋转、缩放、剪切和翻折。给定某正方形候选图像块 $P(u, v, r)$, (u, v) 为图像块中心点在整个图像上的坐标, r 为图像块的半径。

对于以 (u, v) 为坐标中心的像素点 $(i, j) \in P(u, v, r)$,翻折操作定义为 $F_p \cdot (i, j)^T, i, j \in [-r, r]$ 。 F_p 有以下5种情况:

$$\begin{cases} F_{i=0} = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \\ F_{i=0} = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} \\ F_{i=j} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \\ F_{i=-j} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \\ F_o = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \end{cases} \quad (1)$$

式中: F_o 表示无翻折操作。

剪切操作定义为

$$S(\Delta i, \Delta j) = \begin{bmatrix} 1 & \Delta i \\ \Delta j & 1 \end{bmatrix} \quad (2)$$

式中: Δi 和 Δj 分别为 i -和 j -轴上的剪切量。当 $\Delta i = \Delta j = 0$ 时, $S(\Delta i, \Delta j)$ 为无剪切操作。

旋转操作定义为

$$R(\theta) = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix} \quad (3)$$

式中: $0 \leq \theta < 2\pi$ 为旋转角度。特别地, $\theta = 0$ 意味着不进行旋转操作。

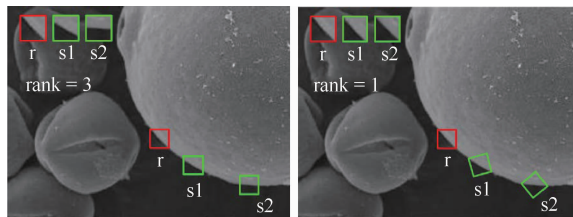
缩放操作定义为数量 $c > 0$ 。当 $c > 1$ 和 $c < 1$ 时,该操作分别放大和缩小图像块。当 $c = 1$ 时,该操作保持原始图像块的大小和形状。

一般地,翻折操作只适用于具有对称性的正方形图像块,因此可以最先施加在图像块 $P(u, v, r)$ 中的每个像素上。旋转操作能够作用于任何形状的图像块,故而可以最后执行。剪切操作适合放置在前两类操作的中间。鉴于此,本文通过有序地联合基本操作引入结构变换。

$$T_s(p, \Delta i, \Delta j, \theta, c) = c \cdot R(\theta) \cdot S(\Delta i, \Delta j) \cdot F_p \quad (4)$$

式中: $p \in \{j=0, i=0, i=j, i=-j, o\}$ 。对于参数 $p=o, \Delta i = \Delta j = \theta = 0$ 和 $c=1$,结构变换使原始图像块 $P(u, v, r)$ 保持不变。为直观地说明结构变

换的作用,对图1所示的参考图像块(r 方框),用不同的方法搜索与其不同的相似图像块($s1$ 、 $s2$ 方框)。从图1(a)、(b)中可以看到,传统的滑动窗口方法产生的相似图像块(含参考图像块自身)有较大的秩(rank),而本文结构变换的结果有较小的秩。



(a) 传统滑动窗口

(b) 本文结构变换

图1 不同方法搜索相似图像块结果

Fig. 1 Results of similar patches that are searched by different methods

2 泊松图像去模糊模型

为在透视场景中有效地去除模糊和抑制泊松噪声,考虑到相似图像块具有低秩性,在正则项 $R(\mathbf{x})$ 中使用 WNN^[20] 对结构变换后的相似图像块进行惩罚。基于此,构造新的目标函数如下:

$$\min_{\mathbf{x}} \langle \mathbf{H}\mathbf{x} - \mathbf{y} \ln(\mathbf{H}\mathbf{x}), \mathbf{1} \rangle + \lambda \sum_n \|\mathbf{X}_n^{\text{ST}}\|_{w,*} \quad (5)$$

式中:第一项和第二项分别为数据项和正则项; $\langle \cdot, \cdot \rangle$ 和 $\lambda > 0$ 分别为内积运算符和超参数; $\mathbf{1}$ 为元素都是1的列向量;矩阵 $\mathbf{X}_n^{\text{ST}}$ 含有 k 列,为图像块 $\mathbf{x}_n$ 的 k 个相似图像块经结构变换后的列向量,WNN 定义为

$$\|\mathbf{X}_n^{\text{ST}}\|_{w,*} = \sum_m w_m \sigma_m(\mathbf{X}_n^{\text{ST}}) \quad (6)$$

其中: σ_m 为矩阵 $\mathbf{X}_n^{\text{ST}}$ 的第 m 个奇异值; w_m 为权重向量 $\mathbf{w}$ 的第 m 个元素。

此外,式(5)与相关工作^[20]中的目标函数有以下2点区别:①文献[20]中的任务是去除加性高斯噪声,所以其目标函数可以建立在图像块的基础上;而本文的任务是去除高斯模糊,故而式(5)只能建立在整幅图像上。②文献[20]把相似图像块的搜索方式限制为平移操作,而式(5)因含有结构变换而更适用于恢复含透视场景的图像。

为了用清晰图像 $\mathbf{x}$ 替换中间矩阵 $\mathbf{X}_n^{\text{ST}}$,式(5)等价地转化为

$$\min_{\mathbf{x}} \langle \mathbf{H}\mathbf{x} - \mathbf{y} \ln(\mathbf{H}\mathbf{x}), \mathbf{1} \rangle + \lambda \sum_n \|\mathcal{T}_s \epsilon_n \mathbf{x}\|_{w,*} \quad (7)$$

式中:算子 ϵ_n 的作用是从清晰图像 $\mathbf{x}$ 中提取出与参考图像块 $\mathbf{x}_n$ 最相似的 k 个正方形图像块;算子 $\mathcal{T}_s$ 的作用与结构变换 $\mathcal{T}_s(p, \Delta i, \Delta j, \theta, c)$ 相同,可

以对提取出的 k 个正方形图像块进行变形。

3 优化求解方案

由于式(7)不可微但具有可分性,基于 HQS 方法提出交替方向最小化(Alternating Direction Minimization, ADM)方案,用于求解目标函数。

式(7)松弛化为

$$\min_{\mathbf{x}, \mathbf{h}, \mathbf{Z}_n} \langle \mathbf{h} - \mathbf{y} \ln(\mathbf{h}), \mathbf{1} \rangle + \lambda \sum_n \|\mathbf{Z}_n\|_{w,*} + \frac{\alpha}{2} \|\mathbf{h} - \mathbf{H}\mathbf{x}\|^2 + \frac{\beta}{2} \sum_n \|\mathbf{Z}_n - \mathcal{T}_s \epsilon_n \mathbf{x}\|_F^2 \quad (8)$$

式中: $\mathbf{h}$ 和 $\mathbf{Z}_n$ 分别为辅助向量和辅助矩阵; $\alpha, \beta > 0$ 为超参数。当 $\alpha, \beta \rightarrow \infty$ 时,式(8)收敛到式(7)。

更新辅助矩阵 $\mathbf{Z}_n$ 。固定式(7)中与 $\mathbf{Z}_n$ 相关的项,得到如下子问题:

$$\min_{\mathbf{Z}_n} \|\mathbf{Z}_n - \mathcal{T}_s \epsilon_n \mathbf{x}\|_F^2 + \frac{2\lambda}{\beta} \|\mathbf{Z}_n\|_{w,*} \quad (9)$$

该问题类似于 WNNM 方法^[20]中的目标函数,其解可以通过变量替换求得。于是,式(8)有如下闭合解:

$$\mathbf{Z}_n = \mathbf{U}\mathbf{S}_w(\mathbf{\Sigma})\mathbf{V}^T \quad (10)$$

式中: $(\mathbf{S}_w(\mathbf{\Sigma}))_{m,m} = \max(\sigma'_m - w_m, 0)$, σ'_m 为矩阵 $\mathbf{\Sigma}$ 主对角线上的第 m 个元素,权重向量 $\mathbf{w}$ 中的第 m 个元素定义为 $w_m = \sqrt{k}/(\sigma'_m + 10^{-16})$; $[\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}] = \text{SVD}(\mathcal{T}_s \epsilon_n \mathbf{x}) = \text{SVD}(\mathcal{T}_s \mathbf{X}_n^{\text{ST}})$, SVD 为奇异值分解(Singular Value Decomposition)。

式(10)为抑制泊松噪声的操作,即从较大(小)的奇异值 σ'_m 中减去较小(大)的权重 w_m 。需要指出的是,式(9)中第一项使用 F -范数是合理的,因为图像块中的泊松噪声在局部区域可近似为高斯分布^[27-28]。

更新辅助向量 $\mathbf{h}$ 。固定与 $\mathbf{h}$ 有关的项,式(8)可以简化为

$$\min_{\mathbf{h}} \langle \mathbf{h} - \mathbf{y} \ln(\mathbf{h}), \mathbf{1} \rangle + \frac{\alpha}{2} \|\mathbf{h} - \mathbf{H}\mathbf{x}\|^2 \quad (11)$$

通过求导可获得问题(11)的闭合解:

$$\mathbf{h} = \frac{1}{2} \left(\mathbf{H}\mathbf{x} - \frac{1}{\alpha} + \sqrt{\left(\mathbf{H}\mathbf{x} - \frac{1}{\alpha} \right)^2 + \frac{4\mathbf{y}}{\alpha}} \right) \quad (12)$$

更新清晰图像 $\mathbf{x}$ 。固定式(8)中与 $\mathbf{x}$ 有关的项,得到如下子问题:

$$\min_{\mathbf{x}} \frac{\alpha}{2} \|\mathbf{h} - \mathbf{H}\mathbf{x}\|^2 + \frac{\beta}{2} \sum_n \|\mathbf{Z}_n - \mathcal{T}_s \epsilon_n \mathbf{x}\|_F^2 \quad (13)$$

令 $\mathbf{Z}_n^q$ 为矩阵 $\mathbf{Z}_n$ 的第 q 列($1 \leq q \leq k$), N_n^q 为第 q 个相似图像块 $\mathbf{x}_n^q$ 在图像 $\mathbf{x}$ 中的索引, $\mathbf{E}_{N_n^q}$ 为从 $\mathbf{x}$ 中抽取 $\mathbf{x}_n^q$ 的矩阵, $\mathcal{T}_s^q$ 为对 $\mathbf{x}_n^q$ 进行结构变换的矩阵。由于 $\mathcal{T}_s$ 和 ϵ_n 是可逆的, $\mathbf{E}_{N_n^q}$ 和 $\mathcal{T}_s^q$ 也是可逆的。于是,式(13)的闭合解为

$$\mathbf{x} = \left(\alpha \mathbf{H}^T \mathbf{H} + \beta \sum_n \sum_{q=1}^k (\mathbf{T}_s^q \mathbf{E}_{N_n^q})^T \mathbf{T}_s^q \mathbf{E}_{N_n^q} \right)^{-1} \cdot \left(\alpha \mathbf{H}^T \mathbf{h} + \beta \sum_n \sum_{q=1}^k (\mathbf{T}_s^q \mathbf{E}_{N_n^q})^T \mathbf{T}_s^q \mathbf{E}_{N_n^q} \mathbf{Z}_n^q \right) \quad (14)$$

式(14)通过对去模糊图像 $\mathbf{H}^T \mathbf{h}$ 和无噪声图像块 $\mathbf{Z}_n^q$ 拼成的图像 $\sum_n \sum_{q=1}^k (\mathbf{T}_s^q \mathbf{E}_{N_n^q})^T \mathbf{T}_s^q \mathbf{E}_{N_n^q} \mathbf{Z}_n^q$ 进行加权平均而恢复出清晰图像 $\mathbf{x}$ 。

综上所述,本文提出的基于结构加权低秩近似的泊松图像去模糊方法步骤如图 2 所示。

步骤 1 输入含有泊松噪声的模糊图像 $\mathbf{y}$ 和模糊核 $\mathbf{H}$ 。

步骤 2 把清晰图像 $\mathbf{x}$ 初始化为输入图像 $\mathbf{y}$, 置 $\alpha = 1.5, \beta = 2$ 。

步骤 3 对 $\mathbf{x}$ 中每一个图像块,搜索并存储 k 个最相似的图像块。

步骤 4 通过计算式(10)更新辅助矩阵 $\mathbf{Z}_n$ 。

步骤 5 通过计算式(12)更新辅助向量 $\mathbf{h}$ 。

步骤 6 通过计算式(14)更新清晰图像 $\mathbf{x}$ 。

步骤 7 如果 $\mathbf{x}$ 的更新率 $\frac{\|\mathbf{x}^{t+1} - \mathbf{x}^t\|}{\|\mathbf{x}^t\|}$ (t 为迭代次数)大于阈值 10^{-3} ,转到步骤 8。否则,输出清晰图像 $\mathbf{x}$ 。

步骤 8 如果 $t \leq \maxIter$,把 α 和 β 乘以 2 后转到步骤 3。否则,输出清晰图像 $\mathbf{x}$ 。

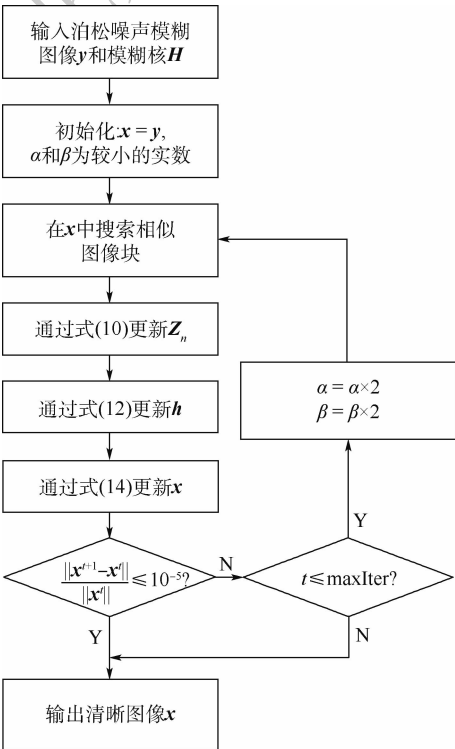


图 2 基于结构加权低秩近似泊松图像去模糊方法流程
Fig. 2 Flowchart of structural weighted low-rank approximation method for Poisson image deblurring

4 实验结果与分析

4.1 实验条件

由于本文主要考虑透视场景中的非局部自相似性,故在量化评估实验中选取如图 3 所示的 20 幅图像作为测试图像。其中,前两行的 10 幅自然图像^[28]含有三维真实场景的透视结构,后两行的 10 幅生物医学图像中不仅存在一定程度的透视结构,还含有细微的边缘信息和纹理细节(<http://cellimagelibrary.org/>)。

为合成含有泊松噪声的模糊图像,根据 $\mathbf{y} = \mathcal{P}(\mathbf{H}\mathbf{x})$ 执行 2 个步骤:①用大小 $s = 9$ 的高斯模糊核(均值和标准差 σ 分别设置为 0 和 3)对图 3 中的测试图像进行卷积,得到模糊图像。②根据

$$f(\mathbf{y} | \mathbf{H}\mathbf{x}) = \prod_n \frac{e^{-[\mathbf{H}\mathbf{x}]_n} ([\mathbf{H}\mathbf{x}]_n)^{y_n}}{y_n!}$$

在模糊图像中添加泊松噪声,从而获得最终合成的退化图像。为验证本文方法在不同泊松噪声强度上的去模糊性能,合成图像中泊松噪声的峰值 P , 分别设置为 63、127、255、511 和 1023。这里,较小的泊松噪声峰值表示较大的噪声强度。此外,为验证本文方法的泛化能力,也在真实图像上进行了去模糊实验。

为验证本文方法在泊松图像去模糊中的有效性,选用当前最新的方法^[6-9]作为比较对象,也 compared 本文方法不含结构变换时的情形。PURE-LET^[8]采用一种端到端的线性框架,其反卷积过程为基函数的线性组合,因此该方法相当于求解线性系统。文献[6-7,9]方法与本文方法类似,也基于迭代方案对泊松图像去模糊。特别地, Hess^[6]和 PDS^[7]采用了非局部自相似性先验,但是这些方法仅仅在滑动窗口中搜索与参考图像块最相似的 k 个图像块。本文方法不含结构变换时的情形可视为改造后用于泊松图像去模糊的 WNNM 方法。此外,由于基于深度网络的方法需要大量的训练数据,而这些数据通常较难获取^[29],



图 3 量化评估中的 20 张测试图像
Fig. 3 Twenty test images for quantitative evaluation

目前还没有发现基于深度学习的泊松图像去模糊方法。

本文方法中的关键参数设置如下:①结构变换 $T_s(p, \Delta i, \Delta j, \theta, c)$ 中的参数, $p \in \{j = 0, i = 0, o\}$, $\Delta i, \Delta j \in \{-0.2, 0, 0.2\}$, $\theta \in \{0, \pi/2, \pi, 3\pi/2\}$, $c \in \{0.8, 0.9, 1, 1.1, 1.25\}$;②对于泊松噪声峰值 63、127、255、511 和 1 023,迭代次数分别设置为 14、14、12、10 和 8,相应的超参数 λ 分别为 0.024、0.02、0.018、0.014 和 0.01;③图像块的大小和相似图像块的个数分别设置为 7×7 和 60。

为评估不同方法的性能,使用客观评价和主观评价 2 种方法比较去模糊结果的质量。在客观评价时,选用峰值信噪比 (PSNR) 和结构相似性 (SSIM) 作为度量指标。一般地,指标的值越大,复原图像的质量越高。所有实验都在配置为 2.4 GHz Core i7 CPU 和 32 GB RAM 的机器上完成。

4.2 模型参数讨论

本文方法主要包含以下参数:式 (4) 中的翻转操作类型 p 、剪切操作在 i -和 j -轴上的偏移量 Δi 和 Δj 、旋转角度 θ 、缩放量 c 及式 (5) 中的超参数 λ 。式 (4) 中参数的量化粗细度对本文方法抑制泊松噪声的性能有较大影响。如果式 (4) 中参数的量化越细,那么候选图像块的数量越多。相应地,前 k 个相似图像块的相似度越大。对较大的奇异值,权重 w_m 就越小,从而相应向量 (表示物体边缘等主要信息) 被越多地保留。对较小的奇异值,相应向量 (表示噪声等次要信息) 被越多地约减,所以方法的去噪性能越强。相反,如果式 (4) 中参数的量化越粗,方法的去噪性能越低。从图 4(a) 中的曲线可以看出,PSNR 随着旋转角度间隔 $\Delta\theta$ 的减小而增加。当 $\Delta\theta = \pi/2$ 时,PSNR 接近最大值,并且趋于平稳。据此,旋转角度的值设置为 $\theta \in \{0, \pi/2, \pi, 3\pi/2\}$ 。此外,用该方法确定式 (4) 中其他参数的量化值。

式 (5) 中超参数 λ 的作用是调节正则项的重要性,其值严重地影响去模糊结果的质量。如果 λ 的取值偏小,数据项起主导作用,那么恢复的清晰图像与输入图像 y 更接近,所以往往含有较多的模糊和噪声。相反,如果 λ 的取值偏大,正则项起主要作用,那么复原结果中的边缘信息通常会被过度平滑。如图 4(b) 所示,对不同强度的泊松噪声,较小和较大的 λ 值都导致了较低的 PSNR。但是平稳的曲线也表明,本文方法对参数 λ 的取值具有较强的鲁棒性。因此,对泊松噪声峰值 63、127、255、511 和 1 023,超参数 λ 分别设

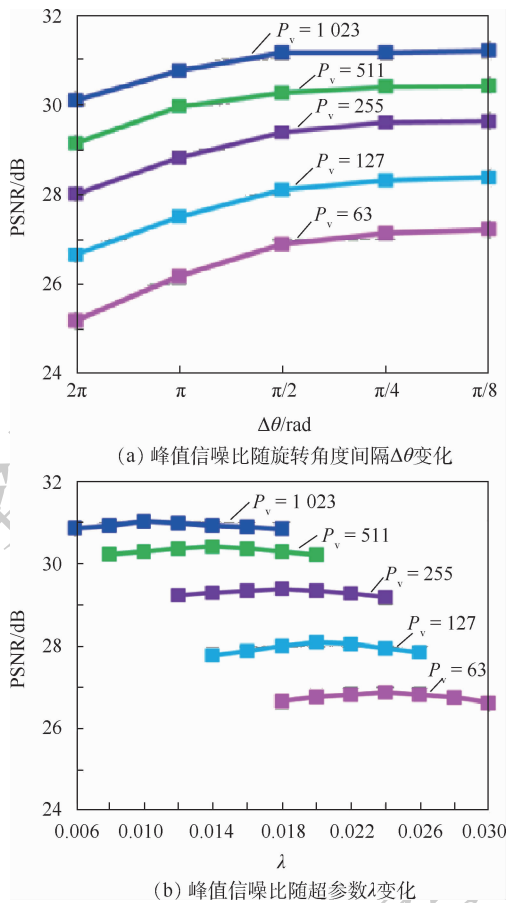


图 4 峰值信噪比随旋转角度间隔 $\Delta\theta$ 和超参数 λ 变化的曲线

Fig. 4 Variation of PSNR with rotation interval $\Delta\theta$ and hyper-parameter λ

置为 0.024、0.02、0.018、0.014 和 0.01。

为验证本文方法对模糊程度和噪声水平的鲁棒性,选用图 3 中的第 5 幅图像作为测试图像。在合成模糊图像时,参数设置如下:模糊核的尺寸 s 和标准差 σ 分别置为 $(s, \sigma) = (7, 2.5), (9, 3)$ 和 $(11, 4)$,均值都为零,泊松噪声峰值分别为 $P_v = 1023, 511$ 和 255。用本文的默认参数从合成图像中恢复出图 5 中的清晰结果。随着噪声强度的增加,复原图像略微变暗,这主要是因为较小的泊松噪声峰值 P_v (模拟实际中较暗的光照条件) 增加了图像去模糊的难度。但是,第三列结果仍然含有比较明显的边缘信息。此外,随着模糊程度的增加,复原物体的纹理边缘丢失了少量信息,但是在整体上仍然比较锐利。

4.3 合成图像上的实验

1) 非盲图像去模糊。为验证本文方法恢复透视图场景细节的性能,首先考虑高斯模糊核已知的情形,并选用图 6(a) 中的生物医学图像作为测试图像。该图像含有丰富的纹理细节,但细胞表面的细节受细胞球形结构的影响存在一定的形

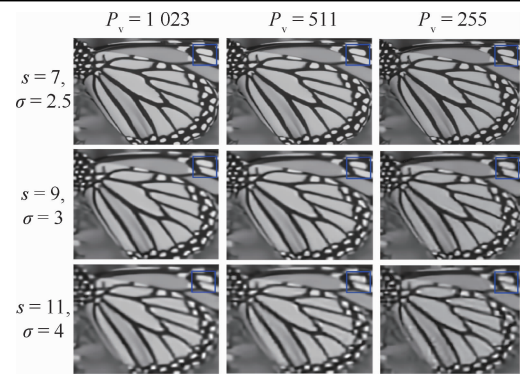


图5 本文方法对高斯模糊程度和泊松噪声水平的鲁棒性

Fig.5 Robustness of proposed method to levels of Gaussian blur and Poisson noise

变。图6(b)中的合成图像含有泊松噪声,峰值为511,不同方法的去模糊结果如图6(c)~(h)所示。通过观察放大的细节可以发现,图6(c)中的细胞表面含有明显的噪声,主要因为PURE-LET方法^[8]中的反卷积过程容易放大图像噪声。由于TGV方法^[9]中的剪切波字典与图像内容无关,图6(d)中细胞表面的一些纹理细节因过平滑而丢失。虽然文献[6-7]也采用了非局部自相似性先验,但是图6(e)、(f)中的细胞都有严重的模糊伪影,因为这2种方法在搜索相似图像块时仅仅限于正视角。本文方法如果在结构变换中使用保持原图像块的参数设置,也会产生类似的模糊效果,如图6(g)所示。相反,本文方法在式(4)中使用更多的参数量化时产生了图6(h)中锐利的纹理信息。表明本文方法在恢复透视场景细节方面优于当前最新的同类方法^[6-9]。

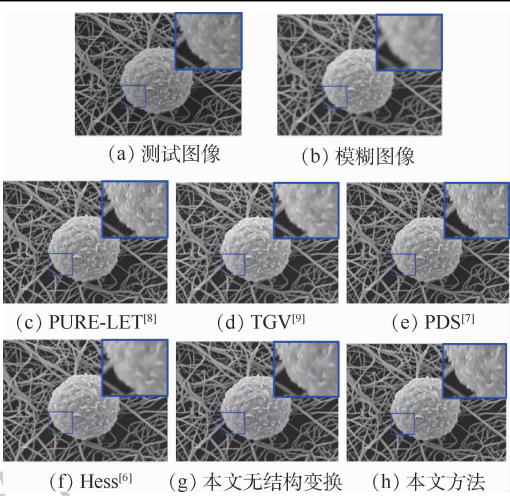


图6 非盲图像去模糊中不同方法的比较

Fig.6 Comparison of different methods in non-blind image deblurring

为进一步验证本文方法的有效性,利用图3中的20幅图像进行更多的非盲图像去模糊实验,并使用mPSNR和mSSIM作为定量评价指标,表1给出了不同方法在测试图像上的平均统计结果。考虑到噪声强度对方法性能的影响,表1也包括了3个典型的泊松噪声峰值,即63、255和1023。可以看到,在自然图像和医学图像上,本文方法、Hess^[6]、PDS^[7]和TGV^[9]的mPSNR都高于PURE-LET^[8]。这主要是因为前4种方法都基于正则化理论和迭代优化方案,能够有效地抑制泊松噪声;而PURE-LET使用端到端的线性框架,对图像中的乘性噪声比较敏感。虽然Hess^[6]、PDS^[7]和TGV^[9]产生了更高的mPSNR,但是增量比较有限。原因

表1 非盲图像去模糊中多种泊松噪声强度下不同方法恢复结果的平均峰值信噪比和结构相似性

Table 1 Mean PSNR and SSIM of results recovered by different methods on test images with various intensities of Poisson noise in non-blind image deblurring

图像	方法	mPSNR/dB			mSSIM		
		$P_v = 63$	$P_v = 255$	$P_v = 1023$	$P_v = 63$	$P_v = 255$	$P_v = 1023$
自然图像	PURE-LET ^[8]	23.09	25.13	26.02	0.8525	0.8844	0.9097
	TGV ^[9]	24.35	26.16	28.86	0.8590	0.8939	0.9188
	PDS ^[7]	23.28	25.70	26.47	0.8566	0.8921	0.9151
	Hess ^[6]	25.15	27.63	29.75	0.8734	0.9011	0.9206
	本文无结构变换	25.04	27.60	29.49	0.8601	0.8953	0.9184
	本文方法	26.32	28.65	30.61	0.8865	0.9126	0.9317
医学图像	PURE-LET ^[8]	24.05	26.09	27.06	0.8367	0.8733	0.8976
	TGV ^[9]	24.96	27.17	29.75	0.8438	0.8806	0.9043
	PDS ^[7]	24.12	26.18	27.53	0.8418	0.8797	0.9002
	Hess ^[6]	26.10	28.59	30.63	0.8575	0.8891	0.9078
	本文无结构变换	26.06	28.26	30.19	0.8442	0.8836	0.9035
	本文方法	27.14	29.70	31.26	0.8719	0.9024	0.9209

主要是: TGV^[9] 中的剪切波是固定的,不能有效地表示图像中的纹理细节,而 PDS^[7] 和 Hess^[6] 都只在正视角度下搜索相似图像块。相反,对各种强度的泊松噪声,本文方法在使用结构变换时取得的 mPSNR 和 mSSIM 都明显地高于当前最新的去模糊方法^[6-9]。这主要得益于结构变换增加了搜索空间中候选图像块的相似性,从而有利于保持图像内容中的主要信息和约减次要信息。

2) 盲图像去模糊。通常情况下,模糊核是未知的,所以在去模糊之前需要对模糊核进行估计。为验证本文方法对模糊核估计的鲁棒性,在合成数据上进行盲图像去模糊实验。为此,首先选用图 7(a)所示的模糊核作为真值,图 7(c)中含有对称结构的自然图像作为测试图像。按照 4.1 节的方法,利用这些数据生成含有峰值为 511 的泊松噪声模糊图像,如图 7(d)所示。使用文献[30]中的方法从模糊图像中估计模糊核,如图 7(b)所示。把估计的模糊核输入到不同的去模糊方法^[6,8,9]中,得到图 7(e)~(h)所示的恢复结果。

从图 7(e)可以看出,蝴蝶的左翼含有明显的噪声,因为 PURE-LET 方法^[8] 中反卷积过程使用了线性维纳滤波(Wiener Filtering),容易放大乘性泊松噪声。图 7(f)中的去模糊结果有明显的过平滑现象,主要原因是 TGV 方法^[9] 使用固定的剪切波作为基函数,以致表示细节的能力十分有限。从图 7(g)中可以看出,虽然 Hess 方法^[6] 显著地改善了纹理细节,但是恢复结果中的边缘信息仍然不够清晰。相反,本文方法恢复出了图 7(h)中更加锐利的结构信息,所以明显地优于当前最新的去模糊方法^[6,8,9]。需要注意,这些结果的复

原都基于图 7(b)中估计的模糊核。因此,实验结果表明了本文方法在盲图像去模糊中对估计的模糊核有较强的鲁棒。

4.4 结构变换的有效性

结构变换在本文方法中起着重要的作用。为进一步验证结构变换的有效性,首先,根据透视程度把图 3 中的自然图像划分成 3 组:第 1 组为轻微透视图像,包含图 3 中第 9 和第 10 幅图像;第 2 组为中等透视图像,由图 3 中第 6~第 8 幅图像组成;第 3 组为明显透视图像,包括图 3 中的第 1~第 5 幅图像。然后,按照 4.1 节的方法合成 3 组模糊图像,泊松噪声峰值均为 511。最后,使用包含和不含结构变换的本文方法对 3 组退化图像进行非盲去模糊,并计算每组结果的 mPSNR。为表明结构变换的作用,进一步计算本文方法对不含结构变换时 mPSNR 的增量。如图 8 所示,本文方法在轻微透视图像上取得了较小的 mPSNR 增量,而在明显透视图像上取得了较大的 mPSNR 增量。表明本文提出的结构变换有利于从透视结构显著的泊松图像中去除模糊。

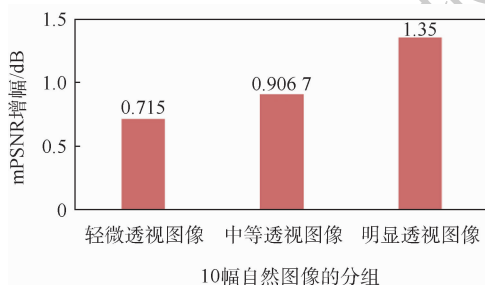


图 8 结构变换的有效性

Fig. 8 Effectiveness of structural transformation

4.5 真实图像上的实验

本文方法在合成图像上取得了较好的效果,为验证其在真实图像上的泛化性能,使用图 9(b)所示的生物学图像(<http://cellimagelibrary.org/>)进行更多的实验。对比方法选用端到端的 PURE-LET^[8] 和基于变分模型的 Hess^[6]。在实验中,首先采用文献[30]中的方法,从真实图像中估计出模糊核,如图 9(a)所示。然后,把这些估计的模糊核输入到不同的去模糊方法中,得到最终的清晰图像。从图 9(c)可以看出, PURE-LET 方法^[8] 能够有效地去除图像模糊、增强纹理细节,但也往往导致图像噪声放大,因为反卷积过程是一个端到端的线性框架,对噪声比较敏感。相反,由于 Hess 方法^[6] 中的基函数与图像内容无关,图 9(d)中的纹理细节被过度平滑。从图 9(e)可以观察到,本文方法成功地去除了高斯

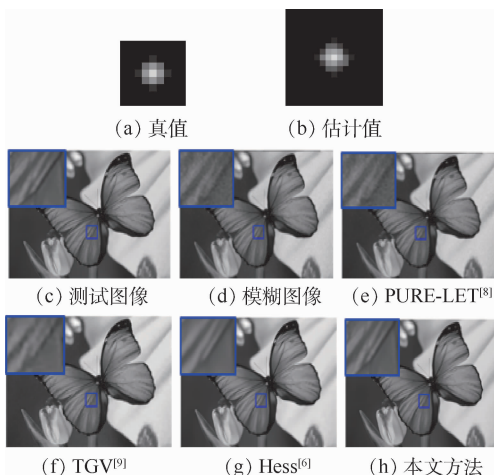


图 7 盲图像去模糊中不同方法的比较

Fig. 7 Comparison of different methods in blind image deblurring

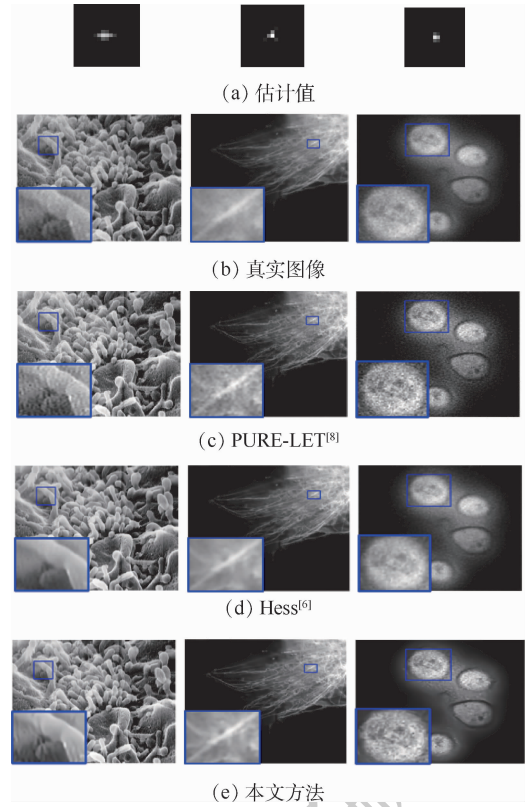


图9 真实图像去模糊中不同方法的比较
Fig.9 Comparison of different methods on real-world image deblurring

模糊和泊松噪声,同时也保留了锐利的边缘信息和纹理细节。这主要得益于结构变换提供了更相似的图像块,使得式(10)中较大奇异值减去较小的权重,相应向量能保留更多的边缘信息。

5 结 论

针对泊松噪声图像去模糊问题,提出了基于结构加权低秩近似的图像去模糊方法。该方法具有以下优势:

- 1) 充分考虑了图像内容的透视结构。引入的结构变换有利于增加搜索空间中相似图像块的相似性,从而改进了非局部自相似性在透视场景中的适用性。
- 2) 可用于无训练样本或者训练量样本量较少的观测数据,如生物医学图像,也可拓展到夜间和背光情形下的高斯模糊图像。

实验表明,本文方法明显优于当前最新的泊松图像去模糊方法,特别是在透视结构显著的图像上。在图像去模糊过程中,抑制泊松噪声放大的关键是通过从较大(小)的奇异值减去较小(大)的权重。如果结构变换的参数量化地越细,与较大奇异值对应的权重就越小,从而在去除高斯模糊和抑制泊松噪声的同时能够有效地保持边

缘信息和纹理细节。但是,这也会导致相似图像块的搜索需要更长的时间。此外,奇异值分解过程也需要时间。鉴于此,下一步工作的重点是通过探索新的搜索策略和奇异值估计方法提高本文去模糊方法的执行效率。

参考文献 (References)

[1] GUO Y F, CAO X C, ZHANG W, et al. Fake colorized image detection[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(8): 1932-1944.

[2] GUO Y F, AU O C, WANG R, et al. Half image watermarking by content aware double-sided embedding error diffusion[J]. IEEE Transaction on Image Processing, 2018, 27(7): 3387-3402.

[3] WANG R, LIANG D, ZHANG W, et al. MatchDR: Image correspondence by leveraging distance ratio constraint[C] // ACM International Conference on Multimedia. New York: ACM, 2016: 606-610.

[4] GU R L, DOGANDZIC A. Blind X-ray CT image reconstruction from polychromatic Poisson measurements[J]. IEEE Transactions on Computational Imaging, 2016, 2(2): 150-165.

[5] LI J, LUISIER F, BLU T. PURE-LET deconvolution of 3D fluorescence microscopy images[C] // International Symposium on Biomedical Imaging. Piscataway: IEEE Press, 2017: 723-727.

[6] LEFKIMMIATIS S, UNSER M. Poisson image reconstruction with Hessian Schatten-norm regularization[J]. IEEE Transactions on Image Processing, 2013, 22(11): 4314-4327.

[7] ONO S. Primal-dual plug-and-play image restoration[J]. IEEE Signal Processing Letters, 2017, 24(8): 1108-1112.

[8] LI J, LUISIER F, BLU T. PURE-LET image deconvolution[J]. IEEE Transactions on Image Processing, 2018, 27(1): 92-105.

[9] LIU X W. Total generalized variation and Shearlet transform based Poissonian image deconvolution[J]. Multimedia Tools and Applications, 2019, 78: 18855-18868.

[10] REN W Q, CAO X C, PAN J S, et al. Image deblurring via enhanced low rank prior[J]. IEEE Transactions on Image Processing, 2016, 25(7): 3426-3437.

[11] LE T, CHARTRAND R, ASAKI T J. A variational approach to reconstructing images corrupted by Poisson noise[J]. Journal of Mathematical Imaging and Vision, 2007, 27(3): 257-263.

[12] YAN Y Y, REN W Q, GUO Y F, et al. Image deblurring via extreme channels prior[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6978-6986.

[13] HARMANY Z T, MARCIA R F, WILLETT R M. This is SPIRAL-TAP: Sparse Poisson intensity reconstruction algorithms—Theory and practice[J]. IEEE Transactions on Image Processing, 2012, 21(3): 1084-1096.

[14] MA L, MOISAN L, YU J, et al. A dictionary learning approach for Poisson image deblurring[J]. IEEE Transactions on Medical Imaging, 2013, 32(7): 1277-1289.

[15] CHEN D Q. Regularized generalized inverse accelerating linearized alternating minimization algorithm for frame-based Poissonian image deblurring[J]. SIAM Journal on Imaging Sciences,

- 2014,7(2):716-739.
- [16] ROND A, GIRYES R, ELAD M. Poisson inverse problems by the plug-and-play scheme[J]. Journal of Visual Communication and Image Representation, 2016, 41: 96-108.
- [17] GETREUER P. Rudin-Osher-Fatemi total variation denoising using split Bregman[J/OL]. Image Processing On Line, 2012, 2: 74-95. (2012-05-19) [2020-03-01]. <https://doi.org/10.5201/ipol.2012.g-tvd>.
- [18] ELAD M, AHARON M. Image denoising via sparse and redundant representations over learned dictionaries[J]. IEEE Transactions on Image Processing, 2006, 15(12): 3736-3745.
- [19] BUADES A, COLL B, MOREL J M. A non-local algorithm for image denoising[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2005, 2: 60-65.
- [20] GU S, ZHANG L, ZUO W M, et al. Weighted nuclear norm minimization with application to image denoising[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014, 1: 2862-2869.
- [21] YAIR N, MICHAELI T. Multi-scale weighted nuclear norm image restoration[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018, 1: 3165-3174.
- [22] HUANG J B, SINGH A, AHUJA N. Single image super-resolution from transformed self-exemplars[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015, 1: 5197-5206.
- [23] BARNES C, SHECHTMAN E, GOLDMAN D B, et al. The generalized patchmatch correspondence algorithm[C]//European Conference on Computer Vision. Berlin: Springer, 2010, 6313(3): 29-43.
- [24] FEDOROV V, BALLESTER C. Affine non-local means image denoising[J]. IEEE Transactions on Image Processing, 2017, 26(5): 2137-2148.
- [25] GU J J, LU H N, ZUO W M, et al. Blind super-resolution with iterative kernel correction[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 1604-1613.
- [26] KAI Z, ZUO W M, ZHANG L. Learning a single convolutional super-resolution network for multiple degradations[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018, 6: 3262-3271.
- [27] LEE J S. Refined filtering of image noise using local statistics[J]. Computer Graphics and Image Processing, 1981, 15(4): 380-389.
- [28] ZHANG K, ZUO W M, CHEN Y, et al. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3142-3155.
- [29] GONG K, CATANA C, QI J Y, et al. PET image reconstruction using deep image prior[J]. IEEE Transactions on Image Processing, 2019, 38(7): 1655-1665.
- [30] KRISHNAN D, TAY T, FERGUS R. Blind deconvolution using a normalized sparsity measure[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2011, 1: 233-240.

作者简介:

吴庆波 男,博士研究生。主要研究方向:图像处理。

任文琦 男,博士,副研究员。主要研究方向:图像处理和机器学习。

Structural weighted low-rank approximation for Poisson image deblurring

WU Qingbo^{1,2}, REN Wenqi^{1,2,*}

(1. State Key Laboratory of Information Security, Institute of Information Engineering, CAS, Beijing 100193, China;
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: To solve the problem of image quality degradation caused by Gaussian blur and Poisson noise, an image deblurring method based on structural weighted low-rank approximation is proposed. First, a structural transformation is introduced by subsequently combining the four basic operations of scaling, rotation, shearing, and flipping in order to boost the similarity of candidate patches in the searching space. Then, a novel objective function is proposed by carefully designing the regularization term. To this end, we perform structural transformation on image patches and then penalize the transformed results with Weighted Nuclear Norm (WNN) based on the assumption of low-rank among non-local similar patches, suppressing Poisson noise at the same time of deblurring. Finally, an alternating optimization algorithm is presented based on the Half-Quadratic Splitting (HQS) method to solve the proposed objective function for Poisson image deblurring. Experimental results demonstrate that, under multiple intensities of Poisson noise, the proposed algorithm achieves higher Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) than the state-of-the-art deblurring methods.

Keywords: image deblurring; Poisson noise; structural transformation; weighted low-rank approximation; Half-Quadratic Splitting (HQS)

Received: 2020-03-02; **Accepted:** 2020-03-20; **Published online:** 2020-04-01 09:38
URL: kns.cnki.net/kcms/detail/11.2625.V.20200331.1722.002.html
Foundation items: National Natural Science Foundation of China (U1605252, U1803264, 61802403); National Key R & D Program of China (2019YFB1406500); Beijing Natural Science Foundation (L182057, KZ201910005007, L182057)
*** Corresponding author.** E-mail: renwenqi@iie.ac.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0064

基于收缩场学习的 Retinex 低照度图像增强



吴庆波^{1,2}, 王蕊^{1,2}, 任文琦^{1,2,*}

(1. 中国科学院信息工程研究所 信息安全国家重点实验室, 北京 100193;

2. 中国科学院大学 网络空间安全学院, 北京 100049)

摘 要: 为增强低照度图像和抑制噪声, 提出了一种通过学习收缩场(SF)改进 Retinex 分解的图像增强方法。首先, 构造新的目标函数, 在正则项中引入 2 组不同的高阶滤波器分别约束未知的反射图和照明图。高阶滤波器可以学习到多种激活模式, 有利于在恢复反射图的同时抑制噪声污染。然后, 在优化目标函数时通过求解收缩场更新隐变量, 参数化的压缩函数可以自适应地调整相应滤波器在反射图和照明图上的响应。最后, 在每个级联内更新照明图之前, 嵌入一个辅助的收缩场, 以抑制噪声和不良伪影的传播, 从而更精确地估计照明图。实验结果表明, 所提方法取得的峰值信噪比(PSNR)和结构相似性(SSIM)均高于当前最新的低照度图像增强方法。

关键词: 低照度图像; Retinex 模型; 图像增强; 图像去噪; 收缩场(SF)模型

中图分类号: TP391

文献标志码: A **文章编号:** 1001-5965(2020)09-1711-10

受光照强度的影响, 在夜间和背光条件下采集的图像往往含有较低的对比度、大面积的暗区域和明显的噪声污染。这些降质图像往往导致人们无法正确地辨识场景内容, 也常常给图像检索、多媒体信息安全等后续计算机视觉任务^[1-3]带来严峻的挑战。因此, 低照度图像增强具有重要的理论价值和现实意义, 受到学界广泛关注。

早期, 大部分工作^[4-5]通过直方图调整亮度值的分布对低照度图像进行增强。但是, 这些方法往往产生过增强或欠增强的结果, 而且忽视了较暗区域中存在的噪声。受 Retinex 理论^[6]启发, 学者提出对低照度图像 I 进行分解: $I = R \circ L$, R 为反射图 (Reflectance Image), 表示物体固有的物理属性, L 为照明图 (Illumination Map), 表示图像中物体上的光照强度, “ $\circ$ ”为 Hadamard 乘积。

基于 Retinex 分解, 大部分方法^[7-14]通过调整照明图像对低照度图像进行增强。在该框架下, Guo 等^[7]引入了 LIME (Low-light Image Enhancement) 方法, 先从低照度图像中估计初始照明图, 再利用结构先验对照明图进行改进, 最后通过除法操作恢复反射图。虽然 LIME 方法在后处理阶段使用 BM3D^[15]从反射图中去除噪声, 但该策略对噪声不够鲁棒。一方面是因为受照明图的影响, 反射图中放大的噪声在空间上分布不均匀。另一方面, BM3D 方法假设图像噪声服从高斯分布, 不适用于去除反射图中不均匀的噪声。为避免噪声放大问题, Li 等^[9]把输入图像从 RGB 颜色空间转换到 HSV 颜色空间, 并且在 V 通道上提出 SR (Structure Revealing) 变分模型, 该模型在正则项中利用输入图像的梯度对反射图的梯度进行约

收稿日期: 2020-03-02; 录用日期: 2020-03-20; 网络出版时间: 2020-04-16 14:40

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200416.1427.002.html

基金项目: 国家自然科学基金 (U1605252, U1803264, 61802403); 国家重点研发计划 (2019YFB1406500); 北京市自然科学基金 (L182057, KZ201910005007, L182057)

* 通信作者: E-mail: renwenqi@iie.ac.cn

引用格式: 吴庆波, 王蕊, 任文琦. 基于收缩场学习的 Retinex 低照度图像增强[J]. 北京航空航天大学学报, 2020, 46(9): 1711-1720. WU Q B, WANG R, REN W Q. Learning shrinkage fields for low-light image enhancement via Retinex[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1711-1720 (in Chinese).

束,以保持反射图中物体的结构信息,但该约束使得输入图像中较大的噪声不可避免地传导至反射图,以致增强图像中含有残留噪声。虽然 Retinex 分解模型在低照度图像增强任务中取得了明显的效果,但大部分图像增强方法^[8,10-14]不能有效地抑制反射图中的噪声污染。

近年来,出现了大量基于深度学习的低照度图像增强方法^[16-23],进一步提高增强图像的质量。Wei 等^[16]提出 Retinex-Net 网络对低照度图像进行增强,先使用分解网络(Decom-Net)从观测图像中得到反射图和照明图,再利用 BM3D^[15]和增强网络(Enhance-Net)分别对反射图和照明图进行去噪和增亮,最后通过 $I = R \circ L$ 得到恢复结果,该方法改进了反射图的质量,但对观测图像中较大的噪声仍然比较敏感,因为 BM3D 不适用于抑制反射图中空间变化的噪声。Shen 等^[147]注意到多尺度 Retinex(Multi-Scale Retinex, MSR)方法^[18]在结构上与前向卷积神经网络类似,于是设计了深度网络 MSR-Net,然而,该网络在后处理阶段采用了与 LIME 方法^[7]相同的去噪策略,导致增强结果中往往存在残留噪声。Park 等^[19]使用层数较少的编码-解码(Encoder-Decoder)网络去除反射图中的噪声。与此不同,端到端(End-to-End)的深度模型^[20-24]不考虑 Retinex 分解^[6],直接从低照度图像中获得增强结果。Ren 等^[20]引入混合深度网络(DHN),在内容流分支中使用编码-解码框架捕获具有全局显著性的结构信息,在边缘流分支中增强边缘信息,以弥补内容流分支中编码过程引起的细节丢失。然而,边缘流分支在增强纹理边缘的同时也会引起噪声的放大,降低图像增强结果的质量。大部分基于深度学习的方法改善了低照度图像的增强结果,但仍然容易导致增强图像中含有残留噪声。

为在增强低照度图像的同时抑制噪声,本文首先构造新的目标函数,在数据项中对观测图像进行 Retinex 分解,在正则项中使用 2 组不同的高阶滤波器分别约束未知的反射图和照明图。由于高阶滤波器可以学习到多种不同的激活模式,该设计有利于在恢复反射图像的同时抑制噪声污染。其次,在优化求解目标函数时,采用高斯径向基函数(Radical Basis Function, RBF)的线性组合对压缩函数进行参数化,并通过收缩场(Shrinkage Field, SF)求解未知量。参数化的压缩函数能够适度地调整相应滤波器在反射图和照明图上的响应。最后,在每一级联内更新照明图之前,嵌入一个辅助的 SF 对反射图进行精化,以避免反射图中

的噪声和不良伪影在优化过程中传导至照明图。改进的反射图(又称辅助的反射图)具有较高的质量,更有利于精确估计照明图。

1 传统的收缩场模型

Schmidt 和 Roth^[25]在半正定二次分裂(Half-Quadratic Splitting, HQS)方法中摒弃传统的压缩函数,通过引入 SF 求解非盲图像去模糊问题。

$$E(X|Y) = \arg \min_X \frac{1}{2} \|Y - KX\|^2 + \eta \sum_{s=1}^N \sum_{c \in \mathcal{C}} \rho_s(f_s \otimes X_{(c)}) \quad (1)$$

式中: X 为清晰图像; Y 为观测图像; K 为模糊核 k 的 BCCB 矩阵; $\eta > 0$ 为权重参数;“ $\otimes$ ”为卷积运算符; $X_{(c)}$ 为 X 中的第 c 个团(Clique); $\mathcal{C}$ 为所有团序号的集合; N 和 s 分别为滤波器的个数和序号;势函数 ρ_s 对滤波器 f_s 在 $X_{(c)}$ 上的卷积响应进行建模。

在文献[25]中,SF 定义为清晰图像 X 的解,在第 i 个级联中的数学表达式为

$$X^i = \left[K^T K + \alpha^i \sum_{s=1}^N (F_s^i)^T F_s^i \right]^{-1} \cdot \left[K^T Y + \alpha^i \sum_{s=1}^N (F_s^i)^T \varphi_s^i(F_s^i X^{i-1}) \right] \quad (2)$$

式中: K^T 为 K 的转置矩阵; $\alpha > 0$ 来自参数 η ; F_s 为滤波器 f_s 的 BCCB 矩阵;压缩函数 φ_s 表示高斯径向基函数的线性组合:

$$\varphi_s^i(l) = \sum_{r=1}^M w_{sr}^i \exp\left(-\frac{\tau}{2}(l - \mu_r)^2\right) \quad (3)$$

其中: $M, \tau > 0, w_{sr}^i > 0$ 和 μ_r 分别为径向基核函数的个数、精度、权重和均值。

文献[25]验证了 SF 模型(2)在图像去噪和非盲图像去模糊中的有效性。随后,该模型推广到其他底层计算机视觉任务,如文档图像去模糊(Document Image Deblurring)^[26]、图像去雾^[27]和图像超分辨率(Image Super-Resolution)^[28]。受这些工作的启发,本文把 SF 模型用于低照度图像增强任务。

本文和传统 SF 模型^[25]不同,主要有以下 3 个方面的区别:①传统 SF 模型仅仅复原清晰图像,而本文在恢复反射图的同时还要求解照明图,具有更大的挑战性;②文献[25]在每个级联中使用单个 SF 模型,而本文中的每个级联都由 3 个不同的 SF 模型组成;③传统 SF 模型借助快速傅里叶变换(Fast Fourier Transform, FFT)更新清晰图像,而本文采用文献[27]算法 3 的思路,

借助广义最小化残差 (GMRES) 方法^[29]更新所有未知变量。

2 本文方法

2.1 低照度图像增强模型

图像噪声不可避免地破坏反射图 R 和照明图 L 的估计^[30]。为克服该问题,本文提出一个新的目标函数对低照度图像进行增强。其中,数据项融入了 Retinex 分解模型,正则项使用 2 组不同的高阶滤波器分别约束反射图和照明图。

$$\arg \min_{R, L} \frac{1}{2} \|I - R \circ L\|^2 + \frac{1}{2} \|L - L^*\|^2 + \eta \sum_{s=1}^N \rho_s(F_s R) + \bar{\eta} \sum_{s=1}^N \bar{\rho}_s(\bar{F}_s L) \quad (4)$$

式中:第 1 项为数据项,来自于 Retinex 分解 $I = R \circ L$,该项通过欧氏范数约束恢复的反射图 R 与观测图像 I 在场景内容上保持一致;在第 2 项中, L^* 为初始估计的照明图,通常含有噪声^[30],该项用于约束照明图 L ,以缓解低照度图像增强问题的不适宜性;最后 2 个正则项分别采用高阶滤波器组 F_s 和 $\bar{F}_s$ ($s=1, 2, \dots, N$) 约束反射图 R 和照明图 L ,势函数 ρ_s 和 $\bar{\rho}_s$ 分别对高阶滤波器的响应进行建模,其重要性由权重 $\eta > 0$ 和 $\bar{\eta} > 0$ 决定。

2.2 模型的优化求解

2.2.1 优化反射图和照明图

由于反射图 R 和照明图 L 均为未知量,采用交替迭代策略求解低照度图像增强模型(4)。为此,在 RGB 颜色空间中对低照度图像的每个像素点求最大值,获得初始照明图 L^* 。

1) 更新反射图 R 。固定式(4)中的照明图 L ,保留与反射图 R 有关的项,得到如下子问题:

$$\arg \min_R \frac{1}{2} \|I - R \circ L\|^2 + \eta \sum_{s=1}^N \rho_s(F_s R) \quad (5)$$

受文献[25]的启发,式(5)在第 i 个级联中的解可以表示为 SF 模型。

$$R^i = \left[(L_D^{i-1})^T L_D^{i-1} + \alpha^i \sum_{s=1}^N (F_s^i)^T F_s^i \right]^{-1} \cdot \left[(L_D^{i-1})^T I + \alpha^i \sum_{s=1}^N (F_s^i)^T \varphi_s^i(F_s^i \tilde{R}^{i-1}) \right] \quad (6)$$

式中: L_D 为向量 L 的对角化矩阵;压缩函数 $\varphi_s^i(l)$ 定义为式(3)中高斯径向基函数的线性组合,其中参数 w_{sr}^i 为未知量。

借鉴文献[25]的做法,本文在式(3)中使用 $M=53$ 个高斯径向基函数,这些函数的均值 μ_r 均匀地分布在 $-1.1 \sim 1.1$ 的范围内。为论述方便,

SF 模型(6)简记为模型参数 $\Psi^i = (\alpha^i; \tilde{f}_s^i; w_{sr}^i)$ 。需要注意,在式(6)中用辅助的反射图 $\tilde{R}^{i-1}$ 代替了原来的变量 R^{i-1} ,原因在本节的方法部分进行讨论。

2) 更新辅助反射图 $\tilde{R}$ 和照明图 L 。由于输入图像 I 和初始的照明图 L^* 往往含有噪声,在 $i=1$ 级联中恢复的反射图 R 也通常含有残留噪声和不良特征,以致在该级联中估计的照明图不够精确。为解决该问题,在每个级联内更新照明图 L 之前引入辅助的反射图 $\tilde{R}$ 对反射图 R 进行精化。

固定照明图 L ,辅助的反射图 $\tilde{R}$ 在第 i 个级联中的解为 SF 模型。

$$\tilde{R}^i = \left[(L_D^{i-1})^T L_D^{i-1} + \tilde{\alpha}^i \sum_{s=1}^N (\tilde{F}_s^i)^T \tilde{F}_s^i \right]^{-1} \cdot \left[(L_D^{i-1})^T I + \tilde{\alpha}^i \sum_{s=1}^N (\tilde{F}_s^i)^T \tilde{\varphi}_s^i(\tilde{F}_s^i R^i) \right] \quad (7)$$

式中: $\tilde{F}_s$ 为辅助滤波器 $\tilde{f}_s$ 的 BCCB 矩阵; $\tilde{\alpha}$ 来自辅助的权重参数 $\tilde{\eta} > 0$; 压缩函数 $\tilde{\varphi}_s^i(l)$ 表示为 $\tilde{\varphi}_s^i(l) = \sum_{r=1}^M \tilde{w}_{sr}^i \exp\left(-\frac{\tau}{2}(l - \mu_r)^2\right)$, 权重参数 $\tilde{w}_{sr}^i > 0$ 与式(3)中压缩函数的 w_{sr}^i 不同。

此外,为记述方便,SF 模型(7)简记为模型参数 $\tilde{\Psi}^i = (\tilde{\alpha}^i; \tilde{f}_s^i; \tilde{w}_{sr}^i)$ 。

为更新照明图 L ,首先在目标函数中固定反射图 R ,保留与 L 有关的项。其次,用更精确的辅助反射图 $\tilde{R}$ 代替原来的反射图 R ,得到如下子问题:

$$\arg \min_L \frac{1}{2} \sum_q \|I^q - \tilde{R}_D^q \circ L\|^2 + \frac{1}{2} \|L - L^*\|^2 + \bar{\eta} \sum_{s=1}^N \bar{\rho}_s(\bar{F}_s L) \quad (8)$$

式中: $\tilde{R}^q$ 为辅助反射图 $\tilde{R}$ 的第 q 个通道 ($q \in \{R, G, B\}$), $\tilde{R}_D^q$ 为 $\tilde{R}^q$ 的对角化矩阵。

式(8)中照明图 L 在第 i 级联中的解表示为 SF 模型。

$$L^i = \left[\left(\sum_q (\tilde{R}_D^{q,i})^T \tilde{R}_D^{q,i} + E \right) + \tilde{\alpha}^i \sum_{s=1}^N (\bar{F}_s^i)^T \bar{F}_s^i \right]^{-1} \cdot \left[\left(\sum_q (\tilde{R}_D^{q,i})^T I^q + L^* \right) + \tilde{\alpha}^i \sum_{s=1}^N (\bar{F}_s^i)^T \bar{\varphi}_s^i(\bar{F}_s^i L^{i-1}) \right] \quad (9)$$

式中: E 为单位矩阵; $\tilde{\alpha} > 0$ 来自权重参数 $\tilde{\eta} > 0$; 压缩函数 $\bar{\varphi}_s^i(l)$ 为高斯径向基函数的线性组合, $\bar{\varphi}_s^i(l) = \sum_{r=1}^M \bar{w}_{sr}^i \exp\left(-\frac{\tau}{2}(l - \mu_r)^2\right)$, 权重参数 $\bar{w}_{sr}^i$ 与 SF 模型 Ψ^i 中的 w_{sr}^i 和 $\tilde{\Psi}^i$ 中的 $\tilde{w}_{sr}^i$ 均不相同。

为使用方便,将 SF 模型(9)简记为模型参数 $\bar{\Psi}^i = (\bar{\alpha}^i; \bar{f}^i; \bar{w}_{sf}^i)$ 。

综上所述,本文提出的低照度图像增强方法步骤如下所示,流程如图 1 所示。

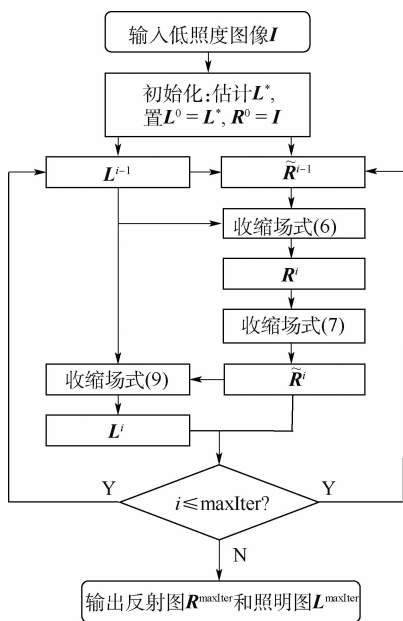


图 1 低照度图像增强方法流程

Fig. 1 Flowchart of low-light image enhancement method

步骤 1 输入低照度图像 I 。

步骤 2 按像素计算 RGB 通道的最大值作为照明图 L^* ,并用其初始化变量 L^0 。

步骤 3 通过计算式(6)更新反射图 R^i 。

步骤 4 通过计算式(7)更新辅助的反射图 $\tilde{R}^i$ 。

步骤 5 通过计算式(9)更新照明图 L^i 。

步骤 6 如果迭代次数 i 小于最大迭代次数 $\maxIter$,转到步骤 3。否则,输出反射图 $R^{\maxIter}$ 和照明图 $L^{\maxIter}$ 。

对引入的辅助反射图 $\tilde{R}$ 作进一步讨论。在式(6)、式(7)和式(9)中,参数 α^i 、 $\tilde{\alpha}^i$ 和 $\bar{\alpha}^i$ 对增强图像和抑制噪声 2 个方面的效果进行调节。以式(6)为例,当 α^i 接近于零时,式(6)更多地起到增亮低照度图像的作用,故而往往导致恢复图像中含有明显的噪声。相反,当 α^i 足够大时,式(6)更多地起到抑制噪声的作用,所以不能有效地增亮输入图像。在噪声抑制过程中,先用压缩函数 $\varphi_i(l)$ 对滤波器 F_s 在反射图 $\tilde{R}^{i-1}$ 上的响应进行调节,再对获得的结果进行反卷积操作,以恢复无噪声的反射图。类似地,上述分析也适用于 SF 模型(7)和 SF 模型(9)。

由于 α^i 在图像增强过程中固定不变,式(6)的结果 R^{i-1} 可能含有残留噪声和不良伪影,尤其

在第 $i=1$ 个级联中。同时,式(9)中 I^q 左边的矩阵本应含有极少的噪声,以利于更新照明图 L^i 。如果该矩阵含有明显的残留噪声,由式(9)估计的照明图 L^i 也往往被噪声污染。因此,为更好地保证优化结果的质量,将反射图 R 替换为更精确的辅助反射图 $\tilde{R}$ 。

2.2.2 学习收缩场模型

为方便论述,2.2.1 节潜在地假设 SF 模型 Ψ^i 、 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$ 已经给定。本节提出基于梯度的算法,从数据中学习这些模型参数。为此,定义 2 个损失函数,并计算其对模型参数的导数。

1) SF 模型 Ψ^i 。通过最小化损失函数 ℓ_1 学习该模型的参数:

$$\ell_1(R^i) = \|R^i - R^{gt}\|^2 \quad (10)$$

式中: R^{gt} 为反射图的真值。

根据式(6), R^i 为模型参数 Ψ^i 的函数,所以 $\ell_1(R^i)$ 对 Ψ^i 的导数为

$$\frac{\partial \ell_1}{\partial \Psi^i} = \frac{\partial \ell_1}{\partial R^i} \cdot \frac{\partial R^i}{\partial \Psi^i} \quad (11)$$

2) SF 模型 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$ 。通过最小化损失函数 ℓ_2 同时学习这 2 个模型的参数:

$$\ell_2(\tilde{R}^i, L^i) = \|\tilde{R}^i - R^{gt}\|^2 + \|L^i - L^{gt}\|^2 \quad (12)$$

式中: L^{gt} 为照明图的真值。

根据式(7)和式(9), $\tilde{R}^i$ 和 L^i 分别为 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$ 的函数,故而 $\ell_2(\tilde{R}^i, L^i)$ 对 $\Phi^i = (\tilde{\Psi}^i, \bar{\Psi}^i)$ 的导数为

$$\begin{aligned} \frac{\partial \ell_2}{\partial \Phi^i} &= \frac{\partial \ell_2}{\partial \tilde{R}^i} \cdot \frac{\partial \tilde{R}^i}{\partial \Phi^i} + \frac{\partial \ell_2}{\partial L^i} \cdot \frac{\partial L^i}{\partial \Phi^i} + \frac{\partial \ell_2}{\partial L^i} \cdot \frac{\partial L^i}{\partial \tilde{R}^i} \cdot \frac{\partial \tilde{R}^i}{\partial \Phi^i} = \\ &= \frac{\partial \ell_2}{\partial \tilde{R}^i} \cdot \frac{\partial \tilde{R}^i}{\partial \tilde{\Psi}^i} + \frac{\partial \ell_2}{\partial L^i} \cdot \frac{\partial L^i}{\partial \bar{\Psi}^i} + \frac{\partial \ell_2}{\partial L^i} \cdot \frac{\partial L^i}{\partial \tilde{R}^i} \cdot \frac{\partial \tilde{R}^i}{\partial \bar{\Psi}^i} \end{aligned} \quad (13)$$

把这些梯度嵌入到有限存储 BFGS 求解器^[31]中学习模型参数 Ψ^i 、 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$ 。综上,本文模型的学习算法主要包括如下步骤:

步骤 1 输入合成的低照度图像 I 、反射图真值 R^{gt} 和照明图真值 L^{gt} 。

步骤 2 对每幅低照度图像 I ,按像素计算 RGB 通道的最大值,把得到的结果作为照明图 L^* ,并用其初始化变量 L^0 。

步骤 3 通过最小化式(10)中的损失函数 ℓ_1 学习 SF 模型参数 Ψ^i 。

步骤 4 通过计算式(6)更新反射图 R^i 。

步骤 5 通过最小化式(12)中的损失函数 ℓ_2 学习 SF 模型参数 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$ 。

步骤 6 通过计算式(7)更新辅助反射图 $\tilde{R}^i$ 。

步骤 7 通过计算式(9)更新照明图 L^i 。

步骤 8 如果迭代次数 i 小于最大迭代次数 $\maxIter$,转到步骤 3。否则,输出学习到的 SF 模型参数 Ψ^i 、 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$ 。

3 实验结果与分析

3.1 实验条件

为学习 SF 模型 Ψ^i 、 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$,本文采用文献[19]的方法合成低照度图像训练集。首先,将正常照度图像作为反射图。然后,把正常照度图从 RGB 颜色空间转换到 HSV 颜色空间,并且对 V 通道进行伽马校正。这里,伽马校正因子在区间[1.6, 3.3]上随机地选取。接着,把校正的 V 通道作为照明图,通过 $I = R \cdot L$ 产生低照度图像。最后,在合成的低照度图像上添加高斯噪声,以模拟低光照条件下的噪声污染。高斯噪声的均值置为 0,方差 σ^2 在区间[0%, 5%]上随机取值。此外,通过 Google 引擎和文献[15, 19, 25, 27]获得 80 幅正常照度图像(尺寸介于 256×256 和 2000×2300 之间),并且随机选取 50 幅图像产生训练数据。图 2 给出了部分低照度图像、反射图和照明图。在训练时,把所有低照度图像裁剪成尺寸为 128×128 的图像块,以加快模型训练的速度。其余的 30 幅正常照度图像用于合成测试数据。

在对比实验中,选用当前最新的低照度图像增强方法 LIME^[7]、DHN^[20] 和 SR^[9] 作为比较对象。基于 Retinex 分解的 LIME 方法^[7] 先通过 $I = R \cdot L$ 恢复反射图,再使用 BM3D^[15] 抑制恢复结果中的噪声。基于变分模型的 SR 方法^[9] 在目标函数正则项中采用梯度约束使增强图像与输入图像保持一致,以抑制噪声放大。DHN 方法^[20] 使用



图 2 训练数据示例

Fig. 2 Examples from training data

端到端的深度网络增强低照度图像。考虑到比较的公平性,用本文的训练集对 DHN 网络进行微调。此外,基于深度学习的 DHN 方法在配置为 Titan XP GPU 的工作站上进行测试,其他所有实验在配置为 2.4 GHz Core i7 CPU 和 32 GB RAM 的计算机上完成。本文使用客观和主观 2 种评价策略比较不同方法的增强结果,并选用峰值信噪比(PSNR)和结构相似性(SSIM)作为客观评价指标。这 2 个指标的值越大,增强图像的质量越高。

3.2 合成数据上的实验

为验证本文方法的有效性,先在合成数据上进行实验。图 3(a)为按照 3.1 节步骤合成的低照度图像,高斯噪声强度为 2%,相应反射图和照明图的真值如图 3(b)和(c)所示。图 3(d)~(g)分别为不同方法的图像增强结果。从图 3(d)可以观察到,LIME 方法^[7] 增强了低照度图像的对比度,但是产生了过亮区域和块状伪影(如放大的花朵边缘)。这主要是因为 LIME 方法对照明图的估计偏低,导致通过 $I = R \cdot L$ 恢复的反射图偏亮;在后处理阶段中使用的 BM3D 方法不能有效地抑制空间变化的噪声,以致增强结果中含有块状伪影。图 3(e)显示,DHN 方法^[20] 恢复了大量的纹理信息,然而增强结果略微偏暗。相反,图 3(f)中的场景不存在过亮区域,且含边缘信息比较锐利。但是,放大的细节仍然包含残留噪声,这是因为 SR 方法^[9] 正则项采用梯度约束,容易把图像噪声传导至增强结果中。与此不同,本文方法在增亮低照度图像的同时有效地避免了过度增强和残留噪声,如图 3(g)所示。原因主要为:学习到的 SF 明显地改善了 Retinex 分解中照明图的质量。

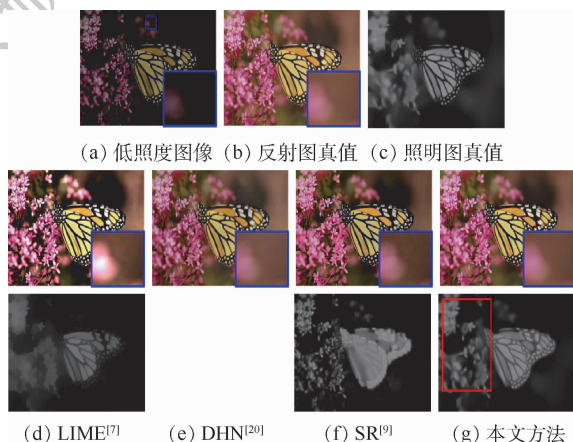


图 3 不同方法在合成数据上的比较

Fig. 3 Comparison of different methods on synthetic data

为定量验证本文方法的有效性,在 30 幅合成图像上进行了更多实验,并计算平均 PSNR(mPSNR)和平均 SSIM(mSSIM)。从表 1 看出,本文方法在多种噪声水平上均取得了最高的 mPSNR 和 mSSIM。

表 1 不同方法在含有多种噪声的测试集上增强结果的平均峰值信噪比和结构相似性

方法	mPSNR/dB			mSSIM		
	$\sigma^2 = 0\%$	$\sigma^2 = 3\%$	$\sigma^2 = 5\%$	$\sigma^2 = 0\%$	$\sigma^2 = 3\%$	$\sigma^2 = 5\%$
LIME ^[7]	19.33	18.62	17.89	0.8526	0.8249	0.7913
DHN ^[20]	19.87	19.07	18.03	0.8922	0.8556	0.8168
SR ^[9]	20.88	19.46	18.21	0.9035	0.8624	0.8304
本文方法	21.73	20.55	19.24	0.9116	0.8713	0.8520

3.3 真实数据上的实验

为评价本文方法在真实图像上的增强效果,选用图 4(a)中的低照度图像进行测试,不同方法的增强结果如图 4(b)~(e)所示。

可以看出,对比方法 LIME^[7]、DHN^[20]和 SR^[9]都有效地增亮了低照度图像。但是,LIME 方法过低估计了照明图,并且在后处理阶段使用 BM3D 抑制反射图中空间变化的噪声,导致增强结果中存在过度增强的笔记本电脑和含有块状伪影的墙面。DHN 方法在端到端深度网络中不使用 Retinex 模型,以致恢复

的场景不够锐利。SR 方法在一定程度上避免了这些问题,但是增强的场景仍然包含明显的残留噪声。究其原因,SR 方法在正则项中使用清晰图像和低照度图像的梯度约束两者在结构信息上保持一致,以致噪声等高频信息传导至增强结果中。相比之下,本文方法产生了相当或者更高质量的增强效果和边缘信息,原因主要是:学习到的高阶滤波器和压缩函数具有丰富的激活模式,能够进一步精化照明图的估计,从而能够有效地恢复反射图像和抑制噪声污染。

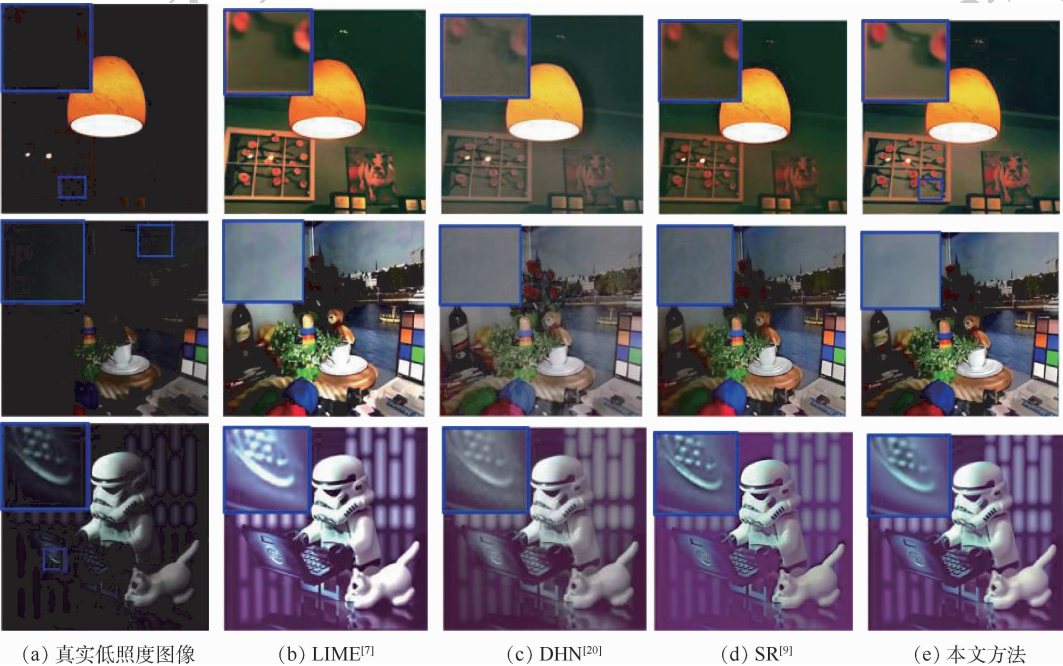


图 4 不同方法在真实数据上的比较
Fig. 4 Comparison of different methods on real-world data

4 讨论和分析

4.1 收缩场模型的有效性

第 3 节验证了本文方法在增强低照度图像的同时能够有效地抑制噪声污染。为说明 SF 模型的有效性,图 5 展示了第 $i = 1$ 级联使用的收缩场

Ψ^i 、 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$ 。每个 SF 中的高阶滤波器(上一行)和压缩函数(下一行)根据其激活程度从左到右降序排列。

在图 5(a)中,右侧 2 个滤波器的激活程度相对较小,相应压缩曲线也含有较少的波动。这与文献[25]中的去噪滤波器一致,故而在抑制噪声

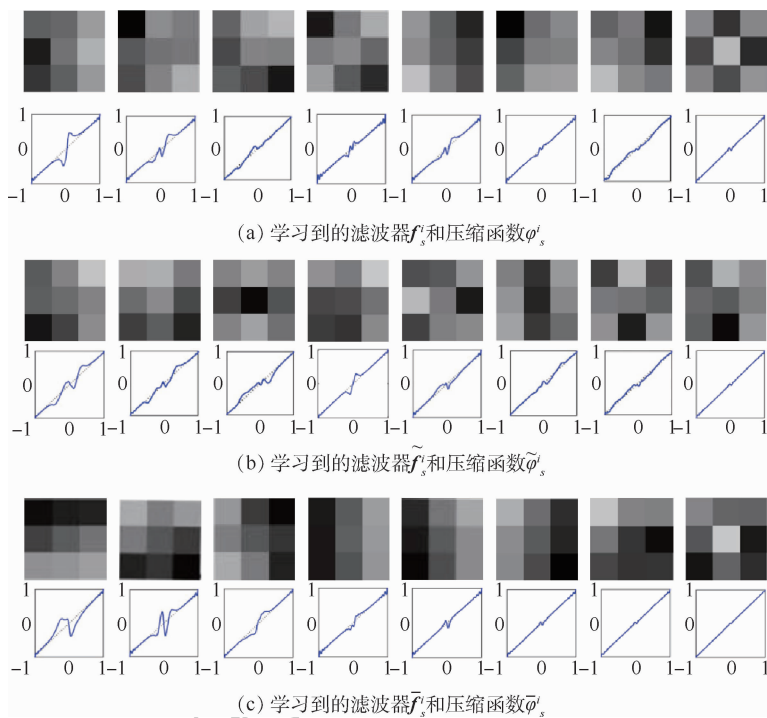


图 5 第 $i=1$ 级联中学习到的收缩场 Ψ^i 、 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$ 的滤波器和相应的压缩函数

Fig.5 Filters and corresponding shrinkage functions in SF models Ψ^i , $\tilde{\Psi}^i$ and $\bar{\Psi}^i$ at cascade $i=1$

方面更能发挥作用。相比之下,前 6 个滤波器对应的压缩曲线含有更多的波动,在图像增强过程中更有利于保持图像的结构信息。与图 5(a)相比,图 5(b)中滤波器有类似的激活程度,但包含不同的视觉模式。这主要是因为:辅助的 SF 模型 $\tilde{\Psi}$ 精化了反射图 R ,生成的辅助反射图 $\tilde{R}$ 含有较少的噪声和不良特征,所以更有助于改进照明图 L 的估计。此外,从图 5(c)可以看出,由于照明图中几乎不含纹理细节,这些滤波器接近梯度算子。综上,学习到的收缩场 Ψ^i 、 $\tilde{\Psi}^i$ 和 $\bar{\Psi}^i$ 具有多种激活模式,故能够改进 Retinex 分解中的照明图,在恢复反射图的同时有效地抑制噪声污染。

4.2 辅助收缩场的有效性

根据 3.1 节的优化步骤,本文方法先求解反射图 R^i 。虽然学习到的高阶滤波器 f_s^i 和压缩函数 ϕ_s^i 能够在恢复反射图的同时抑制噪声污染,但学习到的参数 α^i 固定不变,以致反射图 R^i 的某些区域可能因噪声放大而退化。这种现象在第 $i=1$ 级联中尤为明显。为避免噪声放大问题,引入辅助的收缩场 $\tilde{\Psi}$ 以改善反射图 R^i 的质量,进而精确地估计 Retinex 分解中的照明图 L^i 。

为验证辅助收缩场 $\tilde{\Psi}$ 的有效性,在本文测试集上对本文方法(包含和不含 $\tilde{\Psi}$ 两种情况)进行数量和质量评估。从表 2 可以看出,本文方法使用 $\tilde{\Psi}$ 时取得的 mPSNR 和 mSSIM 都明显地高于不使用时的情形,从而表明了辅助收缩场 $\tilde{\Psi}$ 的

表 2 在本文测试集上评估辅助反射图 $\tilde{R}$ 的有效性
Table 2 Effectiveness evaluation of auxiliary reflectance image $\tilde{R}$ on our test dataset

方法	mPSNR/dB	mSSIM
本文方法(不含 $\tilde{R}$)	16.38	0.7604
本文方法(包含 $\tilde{R}$)	19.24	0.8520

作用。此外,图 6(a)第 1 行为合成的低照度图像,高斯噪声的强度为 2%。从图 6(b)可以看出,本文方法不含 $\tilde{\Psi}$ 时产生了过高估计的照明图,以致增强结果 R 含有偏暗的区域。相反,本文方法包含 $\tilde{\Psi}$ 时使用图 6(a)第 2 行中的辅助反射图 $\tilde{R}$ 产生了图 6(c)中的照明图和反射图。这些改进的 Retinex 分解在视觉上更加接近图 6(d)中的真值,故而验证了辅助收缩场 $\tilde{\Psi}$ 的有效性。

4.3 最大迭代次数

在本文实验中,最大迭代次数设置为 $\text{maxIter} = 5$ 。为说明迭代次数对实验结果的影响,选用图 7(a)中合成的低照度图像展示本文方法在所有级联中的增强结果。从图 7(b)~(e)可以看出,随着级联的增加,反射图的质量逐渐改善。第 3 个级联的反射图有效地避免了过增强和噪声污染,具有高质量的视觉效果。第 5 个级联的反射图具有类似的增强效果,表明本文方法的解趋于稳定。在大量实验的基础上,本文把最大迭代次数 maxIter 置为 5。

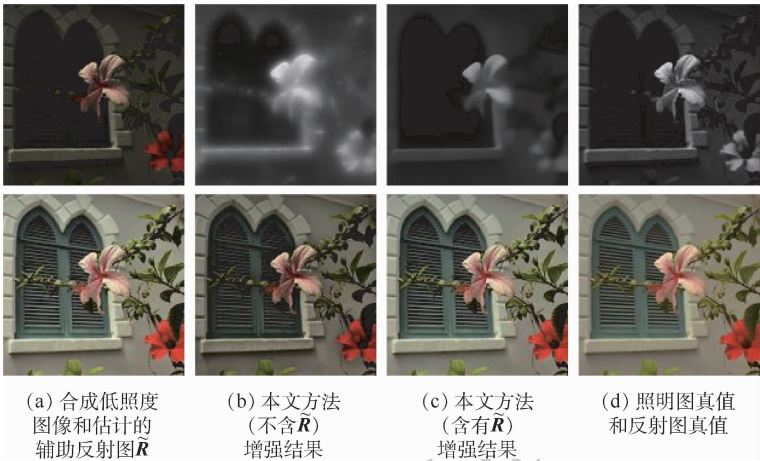


图 6 辅助反射图 $\tilde{R}$ 的有效性

Fig. 6 Effectiveness of auxiliary reflectance image $\tilde{R}$

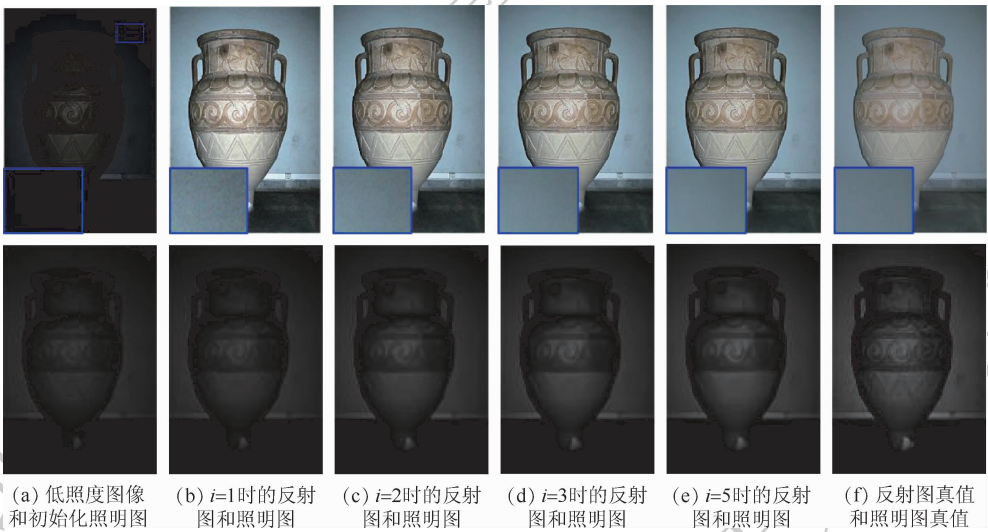


图 7 本文方法随迭代次数增加的增强结果

Fig. 7 Enhancement results of proposed method with increasing iteration times

4.4 高阶滤波器的个数

在每个 SF 中,高阶滤波器初始化为相互独立且完备的简单形式,即中心点与邻域内某点的像素值分别置为 1 和 -1。因此,高阶滤波器的个数 N 与其尺寸 $d \times d$ 存在数量关系 $N = d^2 - 1$ 。本文实验使用 3×3 的滤波器,故每个 SF 包含 $N = 8$ 个高阶滤波器。为表明滤波器的个数对最终增强结果的影响,选用图 8(a)中的合成低照度图像进行测试,该图像含有 1% 的高斯噪声。图 8(b)和(c)分别为 $N = 8$ 和 24 个滤波器的增强结果。可以看出,8 个滤波器已经产生了清晰的眼部轮廓,24 个滤波器进一步增强了眼睛中心的纹理信息。这表明滤波器的个数越大,增强结果的质量越高。然而,由于更大的滤波器减小了未知变量的更新速度,这些效果的提升以牺牲约 1.4 倍的运行时间为代价。通过权衡增强效果和运行时间,本文

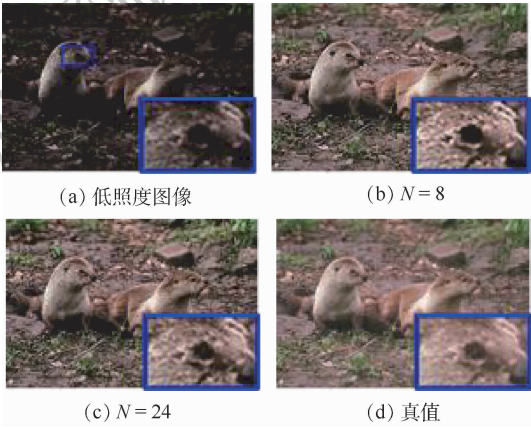


图 8 高阶滤波器个数对实验结果的影响

Fig. 8 Influence of high-order filter number on experimental results

方法在低照度图像增强实验中使用 $N = 8$ 个 3×3 的滤波器。

5 结 论

针对低照度图像增强中的噪声污染问题,本文提出通过学习 SF 模型改进 Retinex 分解,以便在增强低照度图像的同时有效地抑制噪声。合成和真实数据上的实验结果验证了本文方法的有效性。该方法具有以下优点:

1) 不需要数量巨大的训练样本,比基于深度学习的方法更适用于训练样本较少的情形。

2) 学习到的高阶滤波器和压缩函数具有丰富的激活模式,比手工设计的正则子(Regularizer)在抑制噪声污染方面更加有效。

本文方法可以有效地增强低照度图像和抑制噪声污染,但不适用于极低光照图像(Extremely Low-Light Image)的情形。下一步工作将继续改进 SF 模型,以增强极低照度图像。

参考文献 (References)

- [1] WANG R, LIANG D, ZHANG W, et al. MatchDR: Image correspondence by leveraging distance ratio constraint[C]//ACM International Conference on Multimedia. New York: ACM, 2016: 606-610.
- [2] GUO Y F, CAO X C, ZHANG W, et al. Fake colorized image detection[J]. IEEE Transactions on Information Forensics and Security, 2018, 13(8): 1932-1944.
- [3] GUO Y F, AU O C, WANG R, et al. Half image watermarking by content aware double-sided embedding error diffusion[J]. IEEE Transaction on Image Processing, 2018, 27(7): 3387-3402.
- [4] PIZER S M, JOHNSTON R E, ERICKSEN J P, et al. Contrast-limited adaptive histogram equalization: Speed and effectiveness [C]//Proceedings of 1st Conference on Visualization in Biomedical Computing. Piscataway: IEEE Press, 1990: 22-25.
- [5] ARICI T, DIKBAS S, ALTUNBASAK Y. A histogram modification framework and its application for image contrast enhancement[J]. IEEE Transactions on Image Processing, 2009, 18(9): 1921-1935.
- [6] LAND E H, MCCANN J J. Lightness and Retinex theory[J]. Journal of the Optical Society of America, 1971, 61(1): 1-11.
- [7] GUO X J, LI Y, LING H B. LIME: Low-light image enhancement via illumination map estimation[J]. IEEE Transactions on Image Processing, 2017, 26(2): 982-993.
- [8] FU X Y, LIAO Y H, ZENG D L, et al. A probabilistic method for image enhancement with simultaneous illustration and reflectance estimation[J]. IEEE Transactions on Image Processing, 2015, 24(12): 4965-4977.
- [9] LI M D, LIU J Y, YANG W H, et al. Structure-revealing low-light image enhancement via robust Retinex model[J]. IEEE Transactions on Image Processing, 2018, 27(6): 2828-2841.
- [10] YUE H J, YANG J Y, SUN X Y, et al. Contrast enhancement based on intrinsic image decomposition[J]. IEEE Transactions on Image Processing, 2017, 26(8): 3981-3994.
- [11] BANIC N, LONCARIC S. Light random sprays Retinex: Exploring the noisy illumination estimation[J]. IEEE Signal Processing Letters, 2013, 20(12): 1240-1243.
- [12] PARK S, YU S, MOON B, et al. Low-light image enhancement using variational optimization-based Retinex model[J]. IEEE Transactions on Consumer Electronics, 2017, 63(2): 178-184.
- [13] NG M K, WANG W. A total variation model for Retinex[J]. Society for Industrial and Applied Mathematics, 2011, 4(1): 345-365.
- [14] FU X Y, ZENG D, HUANG Y, et al. A weighted variational model for simultaneous reflectance and illumination estimation [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2782-2790.
- [15] DABOV V K K, FOI A, EGIAZARIAN K O. Image denoising by sparse 3-D transform-domain collaborative filtering [J]. IEEE Transactions on Image Processing, 2007, 16(8): 2080-2095.
- [16] WEI C, WANG W J, YANG W H, et al. Deep Retinex decomposition for low-light enhancement [C]//British Machine Vision Conference, 2018: 155-166.
- [17] SHEN L, YUE Z H, FENG F, et al. MSR-net: Low-light image enhancement using deep convolutional network [EB/OL]. (2017-11-07) [2020-03-01]. <https://arxiv.org/abs/1711.02488>.
- [18] JOBSON D J, RAHMAN Z U, WOODDELL G A. A multiscale Retinex for bridging the gap between color images and the human observation of scenes [J]. IEEE Transactions on Image Processing, 1997, 6(7): 965-976.
- [19] PARK S, YU S, KIM M, et al. Dual autoencoder network for Retinex-based low-light image enhancement[J]. IEEE Access, 2019, 6: 22084-22093.
- [20] REN W Q, LIU S F, MA L, et al. Low-light image enhancement via a deep hybrid network [J]. IEEE Transactions on Image Processing, 2019, 28(9): 4364-4375.
- [21] CAI J R, GU S H, ZHANG L. Learning a deep single image contrast enhancer from multi-exposure images [J]. IEEE Transactions on Image Processing, 2018, 27(4): 2049-2062.
- [22] ZHU M F, PAN P B, CHEN W, et al. EEMEFN: Low-light image enhancement via edge-enhanced multi-exposure fusion network [C]//Proceedings of the 34th AAAI Conference on Artificial Intelligence, 2020: 13106-13113.
- [23] CHEN C, CHEN Q F, XU J, et al. Learning to see in the dark [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3291-3300.
- [24] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 3431-3440.
- [25] SCHMIDT U, ROTH S. Shrinkage fields for effective image restoration [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014, 1: 2774-2781.
- [26] XIAO L, WANG J, HEIDRICH W, et al. Learning high-order filters for efficient blind deconvolution of document photographs [C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 734-749.
- [27] WU Q B, REN W Q, CAO X C. Learning interleaved cascade of shrinkage fields for joint image dehazing and denoising [J]. IEEE Transactions on Image Processing, 2020, 29: 1788-1801.
- [28] WEN B, KAMILOV U, LIU D, et al. DeepCASD: An end-to-end

approach for multi-spectral image super-resolution [C] // Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE Press, 2018: 6503-6505.

[29] SAAD Y, SCHULTZ M H. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems [J]. Society for Industrial and Applied Mathematics, 1986, 7 (3) : 856-869.

[30] WANG Y, CAO Y, ZHA Z J. Progressive Retinex: Mutually reinforced illustration-noise perception network for low light image enhancement [C] // ACM International Conference on Multimedia. New York: ACM, 2019: 2015-2023.

[31] SCHMIDT M. minFunc: Unconstrained differentiable multivariate optimization in Matlab [EB/OL]. [2020-03-01]. <https://www.cs.ubc.ca/~schmidtm/Software/minFunc.html>, 2005.

作者简介:
吴庆波 男, 博士研究生. 主要研究方向: 图像处理。
王蕊 女, 博士, 研究员, 博士生导师. 主要研究方向: 多媒体信息智能化处理、计算机视觉等。
任文琦 男, 博士, 副研究员. 主要研究方向: 图像处理和机器学习。

Learning shrinkage fields for low-light image enhancement via Retinex

WU Qingbo^{1,2}, WANG Rui^{1,2}, REN Wenqi^{1,2,*}

(1. State Key Laboratory of Information Security, Institute of Information Engineering, CAS, Beijing 100193, China;
2. School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: To enhance a low-light image and mitigate noises simultaneously, we propose an image enhancement algorithm by learning Shrinkage Field (SF) to improve the reflectance image and the illumination map in Retinex model. To this end, first, we design a novel objective function by constraining the latent reflectance image and the illumination map with two different groups of high-order filters in regularization terms. These filters can be learnt to possess various activation patterns and thus facilitate recovering the reflectance image and suppressing noises at the same time. Then, we update the latent variables in the objective function optimization by calculating SFs, where the parameterized shrinkage functions are capable to scale the responses of the corresponding high-order filters convolving the reflectance image and the illumination map. Finally, we embed an auxiliary SF model before the update of the illumination map in each cascade to suppress the propagation of noises and undesirable artifacts and further to refine the estimation of illumination map. Experimental results demonstrate that the proposed algorithm outperforms the state-of-the-art low-light image enhancement methods in terms of Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM).

Keywords: low-light image; Retinex model; image enhancement; image denoising; model of Shrinkage Field (SF)

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0068

热红外视频监控下行人目标前景区域提取

张玉贵, 沈柳青, 胡海苗*

(北京航空航天大学 计算机学院, 北京 100083)



摘 要: 在热红外视频监控环境下, 针对热红外图像因周围环境温度变化而导致热红外图像灰度值反转的问题, 提出了一种通过热红外图像的边界特征和运动特征的融合来提取行人目标前景区域的方法。首先, 利用行人目标和周围环境存在的显著性差异来提取行人目标的边界特征, 对所提取的边界特征进行边界填充, 并利用热红外行人目标分类器来排除误检目标, 从而获取最终的边界特征提取结果; 其次, 利用相邻帧之间的运动信息来获取行人目标的运动特征, 对所获取的运动特征进行形态学处理, 并利用热红外行人目标分类器来排除误检目标, 从而获取最终的运动特征提取结果; 最后, 对所获取的边界特征提取结果和运动特征提取结果进行融合来获得最终的检测结果。实验证明, 在公开的 OSU 和 LSI 热红外图像行人目标检测数据集中, 所提方法能够有效地降低环境温度变化的不利影响, 并提高行人目标前景区域提取的精度。

关 键 词: 边界特征; 运动特征; 前景区域; 行人目标检测; 灰度值

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2020)09-1721-09

视频监控在公共安全、刑侦破案、灾难救援中具有重要的作用, 尤其在公安视频监控中对行人目标的准确检测更为重要, 从而保障了公安办案人员事前预警、事中干预、事后留证^[1-3]。然而, 在实际现实世界中, 由于夜间低照度光照、雾霾、阴雨天气等的影响, 使得充足光照的时间不到全年总时间的二分之一。夜间低照度光照、雾霾、阴雨天气等光照条件, 严重影响基于可见光的视频监控环境中行人目标检测的准确性, 并且影响后续视频智能化分析处理。因此, 研究基于夜间低照度下的行人目标的准确检测, 并保障视频监控的全天准确检测, 是公安视频监控中重要的应用^[4-9]。在低照度下行人目标的运动前景区域准确提取是行人目标准确检测的基础。为此, 本文重点研究了低照度下基于热红外成像来实现行人目标前景区域的准确提取。

热红外传感器成像是根据任何物体表面的温度超过绝对零度(-273.15°C)时会辐射出不同的红外线电磁信息。这些红外线电磁信息载有物体的特征信息, 因此可用于判别不同目标物的热分布场及温度的高低。利用红外光电技术, 热成像摄像机通过测量目标与背景或目标各部分之间的辐射差异, 将物体辐射的功率信号转换成电信号, 经过放大处理形成视频图像, 也即是利用热成像摄像机的作用, 将物体发出的不可见红外能量转变为人眼可识别的热图像^[9-10]。由于热图像是基于热辐射成像, 也即红外是基于物体表面的温度来成像的, 因此不会受到低照度光照、阴雨天气、雾霾、大雪、大风等光照条件的影响。例如, 正常情况下, 人体的温度要比周边环境的温度要高, 在热图像成像上人体目标就会以高亮成像; 但在夏天, 环境温度较高的情况下, 人体温度低于周围环境温度, 人体目

收稿日期: 2020-03-02; 录用日期: 2020-04-03; 网络出版时间: 2020-05-07 10:15

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200506.1626.009.html

基金项目: 国家重点研发计划(2020AAA0130200); 国家自然科学基金(61772058)

* 通信作者。E-mail: frank0139@163.com

引用格式: 张玉贵, 沈柳青, 胡海苗. 热红外视频监控下行人目标前景区域提取[J]. 北京航空航天大学学报, 2020, 46(9): 1721-1729. ZHANG Y G, SHEN L Q, HU H M. Extraction of foreground area of pedestrian objects under thermal infrared video surveillance[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1721-1729 (in Chinese).

标的成像会不明显,甚至以暗色成像,因此在这种情况下,就会产生图像像素值反转的问题,这为行人目标的检测带来困难。同时,热红外成像技术还受到夜间的路灯、汽车的车灯、楼宇中长明灯的光亮影响,导致行人目标检测的精度降低、误检率或虚警率高,影响后续的智能化分析水平^[11]。

现有主流的基于热红外下行人目标前景区域的提取方法有以下3类。第1类方法主要使用热图像的底层视觉特征,例如目标的纹理、边缘等特征,但这类方法中的热红外图像没有颜色信息,导致不能够有效地检测到行人目标的纹理特征和颜色特征^[12]。第2类方法主要利用视频序列中的时域信息来获取运动区域,主要有帧差法^[13]、背景减法^[14]和光流法^[15]。帧差法基于图像序列中相邻帧差分来获得运动区域,该类方法对于动态环境有一定的适应性,但在目标运动过慢时,容易产生较大的“空洞”。背景减法是当前图像与背景图像进行差分以获得运动区域,该类方法能获得比较完整的运动区域,但在实际应用中难以有效地建立准确的背景模型。光流法是基于视频序列中运动目标随时间变化的光流场特性,但该方法计算复杂,容易受噪声影响,不利于实时处理。第3类方法主要利用视频序列中运动目标的时域运动信息在热红外图像中生成若干疑似包含人体目标的感兴趣区域,后续对感兴趣区域提取特征并使用分类器,来验证是否为行人目标,进而来排除误检目标^[16]。这3类方法尽管在一些特定场景和特定应用上具有较好的检测精度,但对其他场景和环境,尤其是在热红外视频监控中出现灰度值反转的情况下,其适应性与鲁棒性仍有待提升。因此,在热红外视频监控中出现灰度值反转的情况下,对行人目标的前景区域的准确检测具有重要意义。

针对上述问题,本文提出了一种基于行人目标的边界特征和运动特征来提取行人目标前景区域的方法。所提方法建立了有效结合行人目标本身的底层特征和中高层特征的机制,克服了环境中的温度变化所带来的灰度值反转,因此对各类环境的变化有一定鲁棒性。

1 相关工作

近年来,针对热红外图像中行人目标前景区域的提取,研究学者们提出了诸多算法框架与解决方案。一些研究利用图像的视觉底层信息(例如轮廓和纹理信息)来实现行人目标的前景区域的准确提取。Shaik 和 Iftikharuddin^[17]提出了基于频域相关性和贝叶斯概率技术来实现目标的检

测和跟踪,所提出的算法适合于静态或运动目标的实时检测和跟踪,同时在一定程度上能够去除杂波和背景噪声所带来的背景干扰。Davis 和 Sharma^[18]提出了一种基于轮廓的背景检测技术来检测变化很大的热图像中的行人目标,该技术首先利用背景减法识别局部感兴趣区域,其次在每个感兴趣的区域内,将前景和背景中的信息组合起来以形成轮廓显著图,并利用分水岭边界的路径约束搜索用于完成任何破碎的轮廓线段连接,最后填充轮廓图像以产生完整轮廓。Sadjadi^[19]提出了基于小波多分辨率纹理的算法,该算法使用图像的小波分解子带的概率密度函数。这些概率密度函数在聚类算法中用于从背景杂波中分割目标。该算法估计了一组多维预测分析性能模型,这些模型涉及算法内部参数、目标像素数、目标背景干扰比、目标干扰比和局部熵度量。这些模型可用于在没有可用数据的情况下预测区域的性能,并在具有已知场景和度量条件的区域中通过选择最佳参数和恒定的虚警值来优化性能。

同时,为了进一步提升红外视频监控中行人目标前景区域的提取精度,许多研究者利用了行人目标的运动特征。Lei 和 Geng^[20]提出了一种将三帧差分法与背景差分法相融合的红外人体运动目标检测方法,并采用混合高斯模型进行背景建模,结果表明,该方法在运动目标检测中比单独使用三帧差分法和背景差分法更好。付裕等^[21]提出了一种基于高斯模型的运动目标检测方法,该方法利用基本的预处理技术对红外图像做了预处理,随后引入高斯模型,建立背景图像的自适应模型,仿真结果表明,该方法可以准确地检测红外监控图像中的人体运动目标,较好地避免了人体运动速度过快或过慢所产生的拖尾或空洞现象。Viola 等^[22]提出了一种将图像强度信息与运动信息集成在一起的行人目标检测系统,其中使用了一种检测样式算法,该算法在视频序列的两个连续帧上扫描检测器,并使用 AdaBoost 对检测器进行训练,该方法可以以非常小的比例检测出行人,并且误检率非常低。Javed 等^[23]提出了一种背景减除方法,该方法包括3个不同的级别:像素级别、区域级别和帧级别,并且使用了多个线索以在背景干扰的条件下鲁棒地检测出目标,该方法能够解决实际应用中的一些常见问题,例如快速照明改变、静态背景目标的重新定位以及场景中存在运动目标的背景模型的初始化等。

因此,利用热红外图像的底层视觉特征和运动特征能够在一定程度上解决行人目标前景区域

提取的问题,但当热红外图像出现灰度值反转时,上述方法不能够有效地解决该类问题。

2 本文方法

针对在实际热红外视频监控中,由于热红外成像受环境温度的影响,可能会出现图像灰度值反转的问题。然而,不论是行人温度高于周围环境温度,还是周围环境温度高于行人温度,行人目标和背景之间总会存在明显的目标边界。基于此,本文提出了一种在热红外视频监控下提取行人目标前景区域的方法,该方法基于热红外图像行人目标的边界特征提取并融合行人目标的运动特征,包括以下步骤:

步骤 1 利用图像梯度来提取当前帧图像的边界信息,并对边界信息进行区域填充,将边界特征提取的前景区域位置映射到原始图像中的对应位置,并利用热红外行人目标分类器模型进行误检排除,从而获得最终的边界特征提取结果。

步骤 2 利用帧差法来获取相邻帧图像的运动特征,对运动特征进行形态学处理,将运动特征获取的前景区域位置映射到原始图像中的对应位置,并利用热红外行人目标分类器模型进行误检排除,从而获得最终的运动特征提取结果。

步骤 3 对所获得的边界特征提取最终结果和运动特征提取最终结果进行融合以获取行人目标前景区域结果。

其中,热红外行人目标分类器模型是在 LSI 热红外行人目标分类数据集^[24]上使用文献[24]中的 SVM 线性模型进行训练得到的,在训练时所使用的行人正负样本比例是 1:4。本文方法总体流程如图 1 所示。

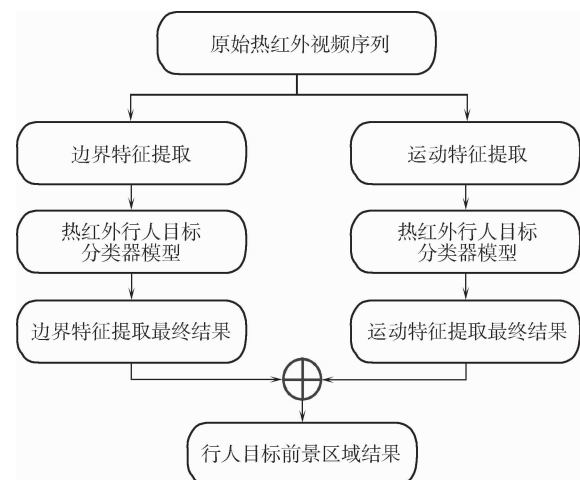


图 1 行人目标前景区域提取的方法流程

Fig. 1 Method flow for extraction of pedestrian object foreground area

2.1 边界特征提取

基于目标与背景之间的显著性差异来进行边界特征提取。Zhang 等^[25]利用运动目标光流场和背景光流场之间的差异来提取运动目标的边界。Liu 和 Xue^[26]首先通过自适应边界特征提取算法来获取可见光视频序列中的目标边界,然后利用三帧差分法的检测结果进行边界补充,并进行形态学处理获得运动目标。

图像梯度是指相邻 2 个像素在水平方向或者垂直方向上的差值。在本文中,对图像梯度分别在水平方向和垂直方向上进行绝对值运算,以提取行人目标和背景之间的边界信息。相关计算见式(1)~式(3)。

$$[dx(i,j), dy(i,j)] = \text{gradient}(I) \quad (1)$$

$$dx(i,j) = |I(i+1,j) - I(i,j)| \quad (2)$$

$$dy(i,j) = |I(i,j+1) - I(i,j)| \quad (3)$$

式中:gradient(I)为热红外图像 I 的梯度; $dx(i,j)$ 和 $dy(i,j)$ 分别为图像在位置 (i,j) 处在水平方向和垂直方向的梯度绝对值; $I(i,j)$ 为当前图像在 (i,j) 位置处的像素值。

为了去除部分干扰噪声,降低目标检测的误检率,将所获取的图像在水平方向的绝对值梯度和在垂直方向的绝对值梯度进行阈值处理,并将经阈值处理后的水平方向梯度绝对值和垂直方向梯度绝对值进行算术叠加,从而得到目标的边界信息。将所获得的目标的边界信息进行区域填充而得到边界前景特征提取结果,将边界特征提取获取的前景区域映射到原始图像中的对应位置,并利用行人目标分类器模型来进一步排除边界特征提取中可能出现的误检目标,从而获得最终的边界特征提取结果。将水平方向的梯度绝对值和垂直方向的梯度绝对值进行叠加得到的边界信息为

$$I_{\text{boundary}}(i,j) = dx_{\text{th}}(i,j) + dy_{\text{th}}(i,j) \quad (4)$$

式中: $I_{\text{boundary}}(i,j)$ 为图像在位置 (i,j) 处的边界信息; $dx_{\text{th}}(i,j)$ 和 $dy_{\text{th}}(i,j)$ 分别为在图像位置 (i,j) 处经过阈值处理的水平方向和垂直方向的梯度信息。边界特征提取及行人目标分类器模型在边界特征提取过程中的各阶段结果如图 2 所示。

图 2(a)是输入的原始热红外图像;图 2(b)是通过图 2(a)在水平方向和垂直方向上分别经过绝对值梯度运算、阈值处理和叠加而获得;图 2(c)是通过图 2(b)进行区域填充并经过行人目标分类器模型排除误检而得到的最终结果。

2.2 运动特征提取

帧差法是一种利用连续的 2 帧热红外图像的差值来判断背景像素和运动像素的检测方法。

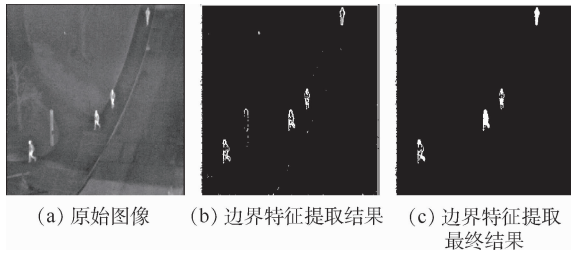


图2 边界特征提取的结果

Fig. 2 Results of boundary feature extraction

Wirayuda 等^[27]提出了一种通过在每个像素中使用运动方差来创建自适应阈值来进行运动特征提取的方法。

本文使用帧差法来获取目标的运动特征。对连续帧之间的差异,通常需要进行阈值处理以检测是否发生了运动。为了判断视频序列中的运动像素和背景像素,首先使用式(5)来计算2帧热红外图像的差值绝对值:

$$\text{diff}_i(x, y) = |I_i(x, y) - I_{i-1}(x, y)| \quad (5)$$

式中: $I_i(x, y)$ 和 $I_{i-1}(x, y)$ 分别为当前帧和上一帧的热红外图像在 (x, y) 位置处的像素值; $\text{diff}_i(x, y)$ 为2帧热红外图像的差值绝对值。

其次,基于式(5)的帧差结果,设定阈值 T 来判断背景像素或运动像素,具体如下:

$$\text{diff}_{\text{res}}(x, y) = \begin{cases} 1 & \text{diff}_i(x, y) \geq T \\ 0 & \text{diff}_i(x, y) < T \end{cases} \quad (6)$$

式中: $\text{diff}_{\text{res}}(x, y)$ 为帧差结果经过阈值判断之后的结果,其值为1表示运动点,其值为0表示背景点。

对经阈值判断的结果进行形态学处理以获取运动特征提取结果。将运动特征提取结果中的行人前景区域映射到原始图像中的对应位置,并使用行人目标分类器模型来排除运动目标检测中可能出现的误检目标,从而获取最终的运动特征提取结果。运动特征提取结果如图3所示。

图3(a)和图3(b)是视频序列中相邻2帧热红外图像数据;图3(c)是经帧差处理的结果,白色区域表示运动目标,黑色区域是背景;图3(d)为经形态学处理及行人目标分类器模型排除误检后的最终结果。

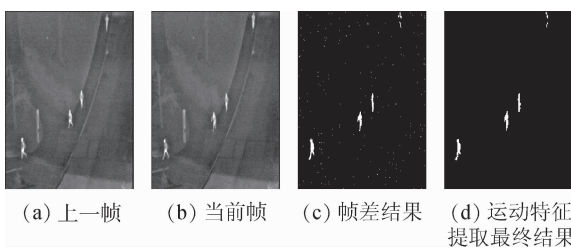


图3 运动特征提取的结果

Fig. 3 Results of motion feature extraction

3 实验结果与分析

3.1 数据集

本文采用 OSU 热红外行人目标检测数据集^[28]和 LSI 热红外行人目标检测数据集^[24]来检验所提方法。在 OSU 热红外行人目标检测数据集的采集中,使用安装有 75 mm 镜头的雷神 300D 热传感器来获得不同天气条件下(包括晴天、阴雨天)总计 284 幅热红外图像,包括 10 个固定监控视角下的热红外图像序列。LSI 热红外行人数据集^[24]是使用驾驶的车辆在室外城市场景中采集的,其中使用 Indigo Omega 成像仪获取图像。它分为行人目标分类数据集和行人目标检测数据集,其中行人目标分类数据集分为训练集和测试集,图像分辨率均为 64 像素×32 像素,宽高比均为 1/2,训练集包含 10 208 个行人正样本和 43 390 个行人负样本,而测试集包含 5 944 个行人正样本和 22 050 个行人负样本。本文的行人目标分类器模型使用 LSI 热红外行人分类数据集的训练集进行训练。LSI 热红外行人目标检测数据集是在 13 个不同的场景中获取的,每个场景中包含不同数量的图像。它总共包括 15 224 个 14 位单通道图像,图像分辨率为 164 像素×129 像素。该行人目标检测数据集的训练集包含 6 159 张图像,测试集包含 9 065 张图像。本文使用上述 2 个热红外行人目标检测数据集序列中具有运动信息的相邻帧来评估所提方法的有效性。

3.2 评价指标

本文使用如式(7)~式(10)所示的精确度(Pr)、召回率(Re)、误检率(FPR)和 F-度量函数(F-measure)指标来评估检测方法的性能^[29-30]。

$$\text{Pr} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

式中:精确度 Pr 为真目标像素量相对于所有目标像素量的比率;TP 为真目标的像素量;FP 为误检的像素量。

$$\text{Re} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (8)$$

式中:召回率 Re 为检测到的真目标像素量相对于所有的真目标像素量的比率;FN 为漏检的像素量。

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} \quad (9)$$

式中:误检率 FPR 为提取的假目标像素量相对于假目标像素量和背景像素量总和的比率;TN 为背景像素的像素量。

$$\text{F-measure} = \frac{2\text{Pr} \cdot \text{Re}}{\text{Pr} + \text{Re}} \quad (10)$$

式中:F-measure 度量函数为精确度和召回率的加权平均值。

TP、FP、TN 和 FN 是使用二值掩码和 Ground-Truth 计算得出的。

3.3 结果与分析

在 OSU 热红外行人目标检测数据集^[28]中,序列 3 存在明显的图像灰度值反转问题,序列 8 和序列 9 的图像灰度值反转不明显。所有图像的分辨率为 360 像素 × 240 像素,选择其中 51 组具有运动信息的图像序列来验证所提方法的有效性。LSI 热红外行人目标检测数据集总共有 13 个视频序列,其中 6 个序列是在摄像机静止下拍摄的,其余是在摄像机移动下拍摄的视频序列,选择在摄像机静止下拍摄的 6 个序列来验证所提方法的有效性。

3.3.1 OSU 热红外行人目标检测数据集的实验结果

在 OSU 热红外行人目标检测数据集^[28]上,将本文方法与显著性检测方法^[31]和帧差法^[13]在

视觉效果和性能指标两方面分别进行了比较。所得实验结果如图 4 所示,其中在图 4(a)和图 4(c)中,人体温度高于周围环境温度,因此不存在灰度值反转的问题,而在图 4(b)中,周围环境温度高于人体温度,存在热红外图像灰度值反转;此外,在图 4(d)中,人体温度和周围环境温度相近,使得路灯存在明显的干扰。

在图 4 中,第 1 行是原始图像,第 2 行是热红外图像所对应的 GroundTruth,第 3 行是显著性检测方法^[31]所得的实验结果,第 4 行是帧差法^[13]所得的实验结果,第 5 行是本文方法的实验结果。从视觉效果上看,相比于显著性检测方法和帧差法,本文方法能够更准确地检测行人目标前景区域,尤其是对于存在灰度值反转的情况,如图 4(b)所示。此外,本文方法还能够降低由其他不具有运动特征的路灯或不具有完整目标边界的汽车灯等所带来的噪声,如图 4(c)和 4(d)所示。显著性检测方法利用了图像的颜色信息,然而热

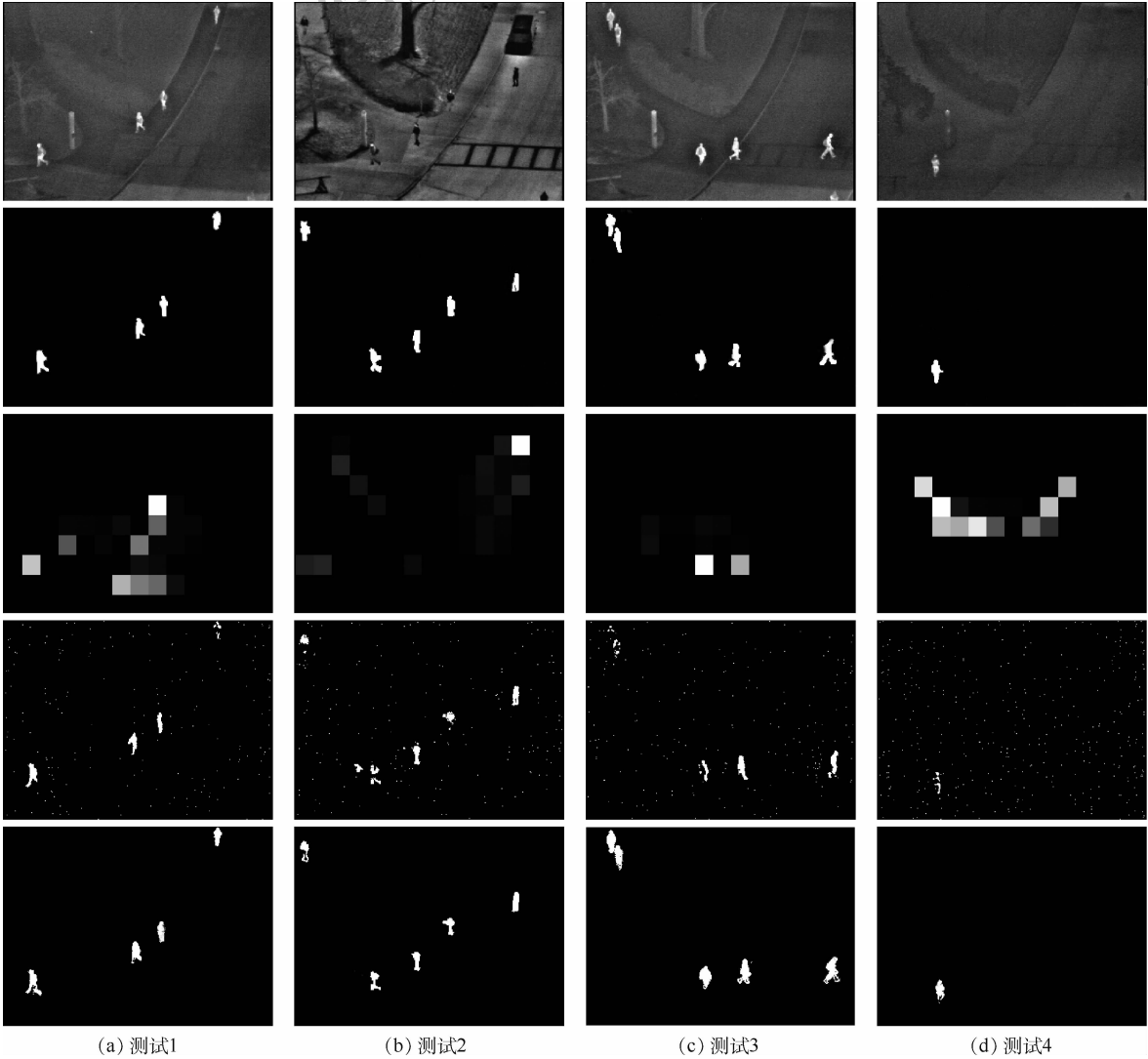


图 4 OSU 热红外行人目标检测数据集实验效果

Fig. 4 Experimental results on OSU thermal infrared pedestrian objects detection dataset

红外图像缺少颜色信息,导致检测精度低。帧差法会产生许多噪声,并存在明显的误检和漏检情况。本文方法主要利用了受图像灰度值影响较小的边界特征和运动特征,同时,使用热红外图像中的行人目标分类器模型进行误检排除,最后通过融合边界特征提取的最终结果和运动特征提取的最终结果来实现 OSU 热红外监控中行人目标前景区域的准确提取。

下面从性能上分析本文方法的有效性。性能评价指标主要是依据文献[29-30]针对前景区域提取所提供的精确度、召回率、误检率和 F-度量函数评价指标。性能评价结果如表 1 所示。

从表 1 可以看出,在 OSU 热红外行人目标检

表 1 OSU 热红外行人目标检测数据集性能评价指标

方法	精确度	召回率	误检率	F-度量函数
显著性检测方法 ^[31]	0.554	0.612	0.157	0.582
帧差法 ^[13]	0.826	0.852	0.064	0.839
本文方法	0.923	0.941	0.043	0.932

测数据集上,相比于显著性检测方法^[31]和帧差法^[13],本文方法具有更高的精确度、召回率和 F-度量函数,和更低的误检率。因此,本文方法利用边界特征和运动特征的融合能够有效解决热红外图像中由灰度值反转所带来的检测精度低的问题,同时能够排除由于实际监控中路灯和汽车灯等其他光源所引起的误检问题。因此,本文方法能够有效地在热红外视频监控下提取行人目标前景区域。

3.3.2 LSI 热红外行人目标检测数据集的实验结果

在 LSI 热红外行人目标检测数据集^[24],将本文方法与显著性检测方法^[31]和帧差法^[13]在视觉效果和性能指标两方面分别进行了比较。所得实验结果如图 5 所示,其中在图 5(a)和图 5(c)中,人体温度高于周围环境温度,因此不存在灰度值反转的问题,而在图 5(b)和图 5(d)中,周围环境温度和人体温度相近,使得图像灰度值发生反转,并且图像序列中背景相似对行人目标的前景区域的检测产生干扰,不利于区分行人目标的前景和背景。

在图 5 中,第 1 行是原始图像,第 2 行是热红

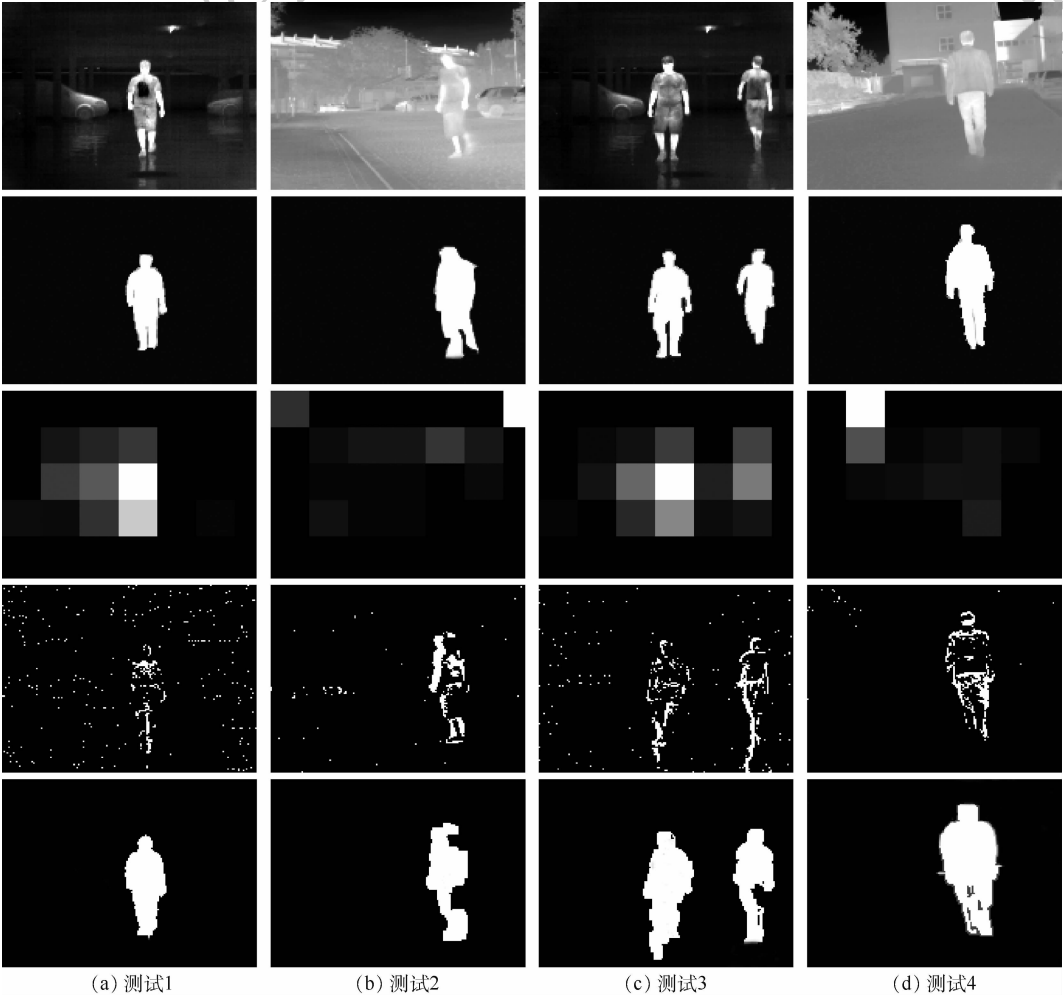


图 5 LSI 热红外行人目标检测数据集实验效果

Fig. 5 Experimental results on LSI thermal infrared pedestrian objects detection dataset

外图像所对应的 GroundTruth,第 3 行是显著性检测方法^[31]所得的实验结果,第 4 行是帧差法^[13]所得的实验结果,第 5 行是本文方法的实验结果。从视觉效果上看,相比于显著性检测方法和帧差法,本文方法能够更准确地检测行人目标前景区域,尤其是对于存在灰度值反转的情况,如图 5(b)和图 5(d)所示。此外,本文方法还能够降低由其他不具有运动特征的建筑物和路灯等所带来的噪声,如图 5(a)和图 5(c)所示。本文方法主要利用了受图像灰度值影响较小的边界特征和运动特征,同时,使用热红外图像中的行人目标分类器模型进行误检排除,最后通过融合边界特征提取的最终结果和运动特征提取的最终结果来实现 LSI 热红外监控中行人目标前景区域的准确提取。

下面从性能上分析本文方法的有效性。性能评价结果如表 2 所示。

表 2 LSI 热红外行人目标检测数据集性能评价指标
Table 2 Performance evaluation indexes of LSI thermal infrared pedestrian objects detection dataset

方法	精确度	召回率	误检率	F-度量函数
显著性检测方法 ^[31]	0.601	0.653	0.124	0.625
帧差法 ^[13]	0.742	0.726	0.087	0.734
本文方法	0.895	0.917	0.059	0.906

从表 2 可以看出,在 LSI 热红外行人目标检测数据集上,相比于显著性检测方法^[31]和帧差法^[13],本文方法具有更高的精确度、召回率和 F-度量函数,和更低的误检率。因此,本文方法利用边界特征和运动特征的融合能够有效解决热红外图像中由灰度值反转所带来的检测精度低的问题,从而有效提取在热红外视频监控下行人目标前景区域。

4 结 论

本文提出了一种提取行人目标前景区域的方法,并在公开数据集上对所提方法进行了实验验证。实验结果表明:

- 1) 所提方法能够有效降低外界环境温度改变对热红外图像灰度值的不利影响。
- 2) 所提方法能够有效地降低行人目标前景区域提取的误检率。

参考文献 (References)

[1] MA Y L, WU X K, YU G Z, et al. Pedestrian detection and tracking from low-resolution unmanned aerial vehicle thermal imagery[J]. Sensors,2016,16(4):446.

[2] 宫文靖. 基于视频的红外运动目标检测研究与应用[D]. 成都:电子科技大学,2016.

[3] GONG W J. Research and application of video-based infrared moving target detection[D]. Chengdu: University of Electronic Science and Technology of China,2016(in Chinese).

[4] LEE J H, CHOI J S, JEON E S, et al. Robust pedestrian detection by combining visible and thermal infrared cameras[J]. Sensors,2015,15(5):10580-10615.

[5] ZHANG L, WU B, NEVATIA R. Pedestrian detection in infrared images based on local shape features[C] // Proceeding of 2007 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2007:1-8.

[6] 李娟. 基于智能视觉技术的公安视频侦查研究[J]. 法制与社会,2019(33):206-207.

[7] LI J. Research on public security video investigation based on intelligent vision technology[J]. Legal System and Society, 2019(33):206-207(in Chinese).

[8] 郭丽静. 视频监控在公安工作中的应用[J]. 电子技术与软件工程,2018(10):140.

[9] GUO L J. Application of video surveillance in public security work[J]. Electronic Technology and Software Engineering, 2018(10):140(in Chinese).

[10] 夏宇. 浅述公安视频监控系统现状与发展要求[J]. 中国安防,2014(20):86-89.

[11] XIA Y. On the status and development requirements of public security video surveillance system[J]. China Security, 2014(20):86-89(in Chinese).

[12] 姜正财,周川. 智能视频监控系统中引入红外热成像技术的探究[J]. 电子制作,2017(16):27-28.

[13] JIANG Z C, ZHOU C. Exploration of infrared thermal imaging technology in intelligent video surveillance system[J]. Electronic Production,2017(16):27-28(in Chinese).

[14] BISWAS S K, MILANFAR P. Linear support tensor machine with LSK channels: Pedestrian detection in thermal infrared images[J]. IEEE Transactions on Image Processing, 2017, 26(9):4229-4242.

[15] WANG Y, ZHU L Q, YU Z J. Foreground detection for infrared videos with multiscale 3-D fully convolutional network[J]. IEEE Geoscience and Remote Sensing Letters, 2019, 16(5):712-716.

[16] 许茗,于晓升,陈东岳,等. 复杂热红外监控场景下行人检测[J]. 中国图象图形学报,2018,23(12):1829-1837.

[17] XU M, YU X S, CHEN D Y, et al. Downward human detection in complex thermal infrared surveillance scenes[J]. Journal of Image and Graphics,2018,23(12):1829-1837(in Chinese).

[18] THOMAS O, FRANK B. Moving object detection and event recognition algorithms for smart cameras[C] // Proceedings of DARPA Image Understanding Workshop,1997:205-208.

[19] TANG Q, DAI S G, YANG J. Object tracking algorithm based on camshift combining background subtraction with three frame difference[J]. Applied Mechanics and Materials, 2013, 373-375:1116-1119.

[20] APARNA A, NIDHI K, RIPUL G, et al. Adaptive contour-based statistical background subtraction method for moving target detection in infrared video sequences[J]. Infrared Physics &

- Technology, 2013, 63:103-109.
- [15] SUN D Q, ROTH S, BLACK M J. A quantitative analysis of current practices in optical flow estimation and the principles behind them[J]. International Journal of Computer Vision, 2014, 106(2):115-137.
- [16] ALSULAMI F, AHMADI K, SALARI E. Detection and tracking of dim objects in infrared (IR) images using support vector machine[C]//2016 IEEE International Conference on Electro Information Technology (EIT). Piscataway: IEEE Press, 2016: 357-361.
- [17] SHAIK J, IFTEKHARUDDIN K M. Detection and tracking of targets in infrared images using Bayesian techniques[J]. Optics and Laser Technology, 2009, 41(6):832-842.
- [18] DAVIS J W, SHARMA V. Robust background-subtraction for person detection in thermal imagery[C]//IEEE International Workshop on Object Tracking & Classification Beyond the Visible Spectrum. Piscataway: IEEE Press, 2004:128.
- [19] SADJADI F A. Infrared target detection with probability density functions of wavelet transform subbands[J]. Applied Optics, 2004, 43(2):315-323.
- [20] LEI M Y, GENG J P. Fusion of three-frame difference method and background difference method to achieve infrared human target detection[C]//2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCA-SIT). Piscataway: IEEE Press, 2019:381-384.
- [21] 付裕, 王志高, 陈文, 等. 基于高斯背景模型的红外人体运动目标检测技术[J]. 自动化与仪器仪表, 2020(1):63-65.
- FU Y, WANG Z G, CHEN W, et al. Infrared human moving target detection technology based on Gaussian background model[J]. Automation and Instrumentation, 2020(1):63-65 (in Chinese).
- [22] VIOLA P, JONES M J, SNOW D. Detecting pedestrians using patterns of motion and appearance[C]//IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2003: 734-741.
- [23] JAVED O, SHAFIQUE K, SHAH M. A hierarchical approach to robust background subtraction using color and gradient information[C]//Proceedings of Workshop on Motion and Video Computing. New York: ACM, 2002:22-27.
- [24] OLMEDA D, PREMEBIDA C, NUNES U, et al. Pedestrian detection in far infrared images[J]. Integrated Computer-Aided Engineering, 2013, 20(4):347-360.
- [25] ZHANG Y G, ZHENG J, ZHANG C, et al. An effective motion object detection method using optical flow estimation under a moving camera[J]. Journal of Visual Communication & Image Representation, 2018, 55:215-228.
- [26] LIU X J, XUE F. Moving target detection based on adaptive edge extraction algorithm[C]//2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA). Piscataway: IEEE Press, 2018:1206-1211.
- [27] WIRAYUDA T A B, LAKSITOWENING K A, STEVANIE F, et al. Development methods for hybrid motion detection (frame difference-automatic threshold)[C]//2013 International Conference of Information and Communication Technology (ICoICT). Piscataway: IEEE Press, 2013:218-222.
- [28] DAVIS J W, KECK M A. A two-stage template approach to person detection in thermal imagery[C]//IEEE Workshops on Application of Computer Vision. Piscataway: IEEE Press, 2005: 364-369.
- [29] LARA S L, SUN D Q, JAMPANI V, et al. Optical flow with semantic segmentation and localized layers[C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016:3889-3898.
- [30] WANG Y, JODOIN P M, PORIKLI F, et al. CDnet 2014: An expanded change detection benchmark dataset[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014:393-400.
- [31] ZHU W J, LIANG S, WEI Y C, et al. Saliency optimization from robust background detection[C]//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014:2814-2821.

作者简介:

张玉贵 男, 博士研究生。主要研究方向: 图像处理。

胡海苗 男, 博士, 副教授, 博士生导师。主要研究方向: 图像处理。

Extraction of foreground area of pedestrian objects under thermal infrared video surveillance

ZHANG Yugui, SHEN Liuqing, HU Haimiao *

(School of Computer Science and Engineering, Beihang University, Beijing 100083, China)

Abstract: Under the thermal infrared video surveillance, in order to solve the problem of the gray value inversion in thermal infrared image caused by the changes in ambient temperature, this paper proposes a method of extracting the foreground area of pedestrian objects by the fusion of boundary feature and motion feature in the thermal infrared images. First, the boundary feature of the pedestrian objects are extracted by using the significant differences existing between the pedestrian objects and the surrounding environment, and then the extracted boundary feature is subjected to area filling, which is followed by the elimination of the false detection objects by using thermal infrared pedestrian objects classifier, thus obtaining the final boundary feature extraction results. Second, the motion feature of the pedestrian objects are obtained by using the motion information between adjacent frames, and then the resulting motion feature is subjected to the morphological processing, which is followed by the elimination of the false detection objects by using thermal infrared pedestrian objects classifier, thus obtaining the final motion feature extraction results. Finally, the final boundary feature extraction results and the final motion feature extraction results are fused to obtain the final detection results. Our experiments show that the proposed method can effectively reduce the adverse effects brought about by the changes in ambient temperature, and further improve the extraction accuracy of foreground area of the pedestrian objects on the OSU and LSI thermal infrared image pedestrian objects detection dataset.

Keywords: boundary feature; motion feature; foreground area; pedestrian objects detection; gray value

Received: 2020-03-02; Accepted: 2020-04-03; Published online: 2020-05-07 10:15

URL: kns.cnki.net/kcms/detail/11.2625.V.20200506.1626.009.html

Foundation items: National Key R & D Program of China (2020AAA0130200); National Natural Science Foundation of China (61772058)

* Corresponding author. E-mail: frank0139@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0062

联合足迹识别与监控视频分析的智能刑侦系统



陶一宁¹, 苏峰¹, 袁培江^{1,*}, 王田苗¹, 钟涛², 郝静如²

(1. 北京航空航天大学 机械工程及自动化学院, 北京 100083;

2. 北京市公安局刑事侦查总队, 北京 100006)

摘 要: 刑侦破案是打击违法犯罪和确保国家长治久安的基本要求。刑侦破案的一大关键是如何有效地利用采集到的信息。为了更好地配合刑侦工作,提出了联合足迹识别与监控视频分析的智能刑侦系统。该系统基于深度学习方法,首先根据足迹信息,包括鞋印长度、宽度、受力分布情况、步长、步幅等,利用卷积神经网络技术实现嫌疑人个人特征的预测;其次联合周边监控视频大数据进行智能分析比较,利用大数据技术快速处理信息,分析视频中行人的个人特点;最后运用虚拟现实仿真技术构建足部压力和鞋底受力分析有限元模型,利用模型获得各种复杂场景下的仿真足迹。三者相互印证,有机结合,快速筛选刑侦对象。实验结果表明,该系统可以高效准确地根据足迹特征实现身高预测,并且与视频监控大数据相结合,可以迅速缩小排查范围并锁定凶手。

关 键 词: 深度学习; 大数据; 刑侦破案; 监控视频分析; 足迹识别

中图分类号: TP391.4

文献标志码: A

文章编号: 1001-5965(2020)09-1730-09

随着科技的发展,犯罪活动智能化、系列化、团伙化的趋势逐渐出现,作案人员反侦察能力越来越强,刻意破坏现场的情况时有发生,被称为“证据之王”的指纹在犯罪现场中提取获得的难度显著提高,而在犯罪现场提取到的作案人员所留足迹越来越多,世界各国的刑侦机关都非常重视对犯罪现场遗留足迹的研究与利用^[1]。

足迹是在犯罪现场上遗留率最高的一类痕迹,所以足迹检验技术是现代刑事技术领域中的重要组成部分。因为每个人的身高、体重、职业、习惯都不一样,所以每个人的步态特征也不相同;步态特征是经过长期习惯后形成的,所以由步态特征形成的足迹具有相对的稳定性。此外,足迹具有一定的指向性,通过对足迹的分析检验可以推测一个人的身高、体重、年龄、走路姿态等,也可

以分析遗留时间、作案路线等,充分利用足迹信息可以实现快速高效地定位嫌疑人。

除犯罪现场以外,周边的监控视频对于刑侦破案也起着至关重要的作用。视频监控是公共安全的重要基础设施,基于深度学习的监控视频分析技术在维护社会安定和打击违法犯罪等方面已经取得了长足的进步,并在很多地区得到了落地应用。在刑侦破案方面,将犯罪现场的足迹信息与周边监控视频中的信息进行联合分析,能够大幅提升破案的效率,形成高效准确的侦查体系,对未来基层民警在相关犯罪现场的刑侦鉴定和破案追凶中起到积极作用。目前在刑侦破案过程中使用的足迹特征识别以及视频图像处理技术存在以下问题:①足迹识别主要依赖经验,缺乏稳定的可推广到基层的智能分析系统;②缺乏根据犯罪现

收稿日期: 2020-03-02; 录用日期: 2020-04-03; 网络出版时间: 2020-05-07 16:13
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200506.1854.011.html
基金项目: 国家自然科学基金(61603015)
* 通信作者: E-mail: itr@buaa.edu.cn

引用格式: 陶一宁, 苏峰, 袁培江, 等. 联合足迹识别与监控视频分析的智能刑侦系统[J]. 北京航空航天大学学报, 2020, 46(9): 1730-1738. TAO Y N, SU F, YUAN P J, et al. Intelligent criminal investigation system based on both footprint recognition and surveillance video analysis[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1730-1738 (in Chinese).

场足迹信息和周边监控视频信息进行目标关联的方法和研究。

为解决这些问题,本文提出了一种联合足迹识别与监控视频分析的智能刑侦系统,该系统基于深度学习方法根据足迹信息准确预测嫌疑人个人特征,同时联合周边监控视频大数据进行智能筛选,迅速缩减刑侦排查范围,提高系统可靠性和实用程度。

本文创新点如下:

1) 通过足迹信息预测嫌疑人个人特点,主要是身高信息。

2) 从监控视频中识别行人的个人特点,并与足迹识别结果匹配,实现刑侦筛查。

1 相关工作

1.1 国内外研究现状

由于足迹特征与人体行为特征的关系十分密切,各国都非常重视对于足迹的研究和利用。西澳大利亚大学的 Ishak^[2-3]、Hemy^[4]等提出根据足迹和人体测量学推断性别的理论方法,并基于澳洲西部的人口进行了测试。英国学者 Reel 等^[5]对二维足迹测量可靠性的方法进行了研究。印度学者 Krishan^[6]针对北印度地区内的人口足迹,进行了足迹个人特征方面的探索。印度学者 Moorthy 等^[7]针对印度泰米尔地区人口足迹,对足迹测量推测身高的可行性进行了验证^[7]。比利时 RSscan International 公司研制了 Footscan 足迹压力测量分析系统^[8],主要针对立体足迹的压力面测量和分析。

而中国是从 1960 年左右开始对足迹信息预测个人特征的相关技术进行研究,当时公安部组织进行相关方向研究的专家学者^[9]对内蒙古自治区步法跟踪专家马玉林的丰富足迹鉴定经验进行了讨论和学习,总结了步法跟踪理论,扩大了足迹检验的应用范围。在此之后,苗春青、葛有忠根据马玉林的研究基础,进一步提出了“单足迹步法检验”^[10]。此方法使用角度描述足迹起落脚特征,然后进行定量分析,以单足迹的起落脚角度为基础进行计算并推断另一对应足迹的起落脚角度,从而认识成趟足迹。最后结合足迹反映出的特征,分析犯罪嫌疑人的个人习惯和关键信息,并对个体的身份做同一认定。河北省雄县公安局刘树权等^[11]提出了一种足迹动力形态检验方法,该方法使用正交设计将影响人体行走特征的主要因素提取出来,由此确认相应的特征量,然后进行分析处理,获取关键参数,实现足迹检验。河南周口

市公安局与解放军信息工程大学等单位联合研制了“立体足迹计算机自动识别系统”^[12],利用足迹高程测量仪采集足迹特征数据,通过计算机进行三维曲面拟合与造型,对比现场足迹和嫌疑人足迹的相似度来综合鉴定足迹同一性。山东科技大学的闫旭琴^[13]设计的足迹对比系统采用二维图像对比足迹图像中的纹理特征,但实际操作时采集的足迹样本信息往往并不完整。王永栋等^[14]研发的“基于重压面形状的立体足迹造型与识别系统”,该系统对立体足迹的基准面选取和变化比较敏感。蒋晓玲^[15]、董菲菲^[16]研发和设计了“足迹三维比对系统”,通过对足迹纹理立体信息分析,进行足迹相似度比对,该方法虽然能够更客观地反映足迹的信息,但是对足迹特征参数的提取不够明显。

1.2 卷积神经网络

卷积神经网络是一种高使用率的深度学习结构,其组成包括输入层、卷积层、池化层、全连接层以及输出层。

卷积层使用多个卷积核对输入图像进行特征提取,获得特征图。卷积层为了表达复杂特征,还包含激励函数,通常选择线性整流函数,可以在不改变卷积层的前提下,增强神经网络整体的非线性特性。池化层位于卷积层之后,输入图像通过卷积层提取出特征图,然后需要通过池化层进行特征选择和信息过滤。目前使用较多的池化函数是最大池化函数,有效缩减参数的数量和计算量,控制过拟合的出现,还可以降低卷积层对边缘的敏感性。全连接层一般位于输出层之前,其作用是对卷积层和池化层提取出的特征进行非线性组合,从而得到输出结果。

卷积神经网络相比于其他前馈神经网络,需要考量更少的参数,且输出结果表现良好,所以广泛应用于图像处理和语音识别领域,是深度学习的代表算法之一。

1.3 大数据分析

当前时代是一个信息爆炸的时代,为了处理海量的信息数据,大数据分析技术在各个行业的使用频率不断增加。

大数据是指大量的复杂的巨型数据集,远远超出传统数据处理技术的能力范围,为了处理这些数据集,大数据分析方法应运而生。大数据分析包括可视化分析、数据挖掘算法、预测性分析能力以及数据质量和数据管理等。可视化分析是将数据可视化,更加直观地展示数据,确保用户更高效地获取数据信息。数据挖掘算法,包括集

群、分割、孤立点分析等算法,可以更细致地分析数据,挖掘更多有效信息。预测性分析能力是在现有数据分析结果的基础上进行预测性判断。数据质量和数据处理是指为了得到高质量分析结果,需要在数据处理时使用高质量的标准化流程和工具。

2 基于足迹识别和监控视频的刑侦系统

2.1 系统技术路线

本文系统以足迹识别为基础,结合监控视频大数据分析和虚拟现实仿真技术,在刑侦过程中实现智能快速破案,过程如图 1 所示。

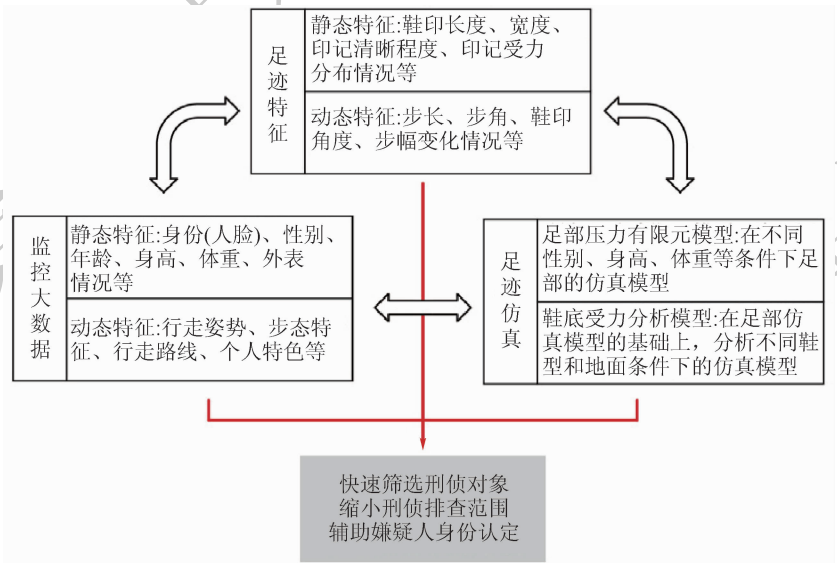
足迹特征中包含静态特征和动态特征,由此提取关键信息,分析嫌疑人个人特点;运用大数据分析技术处理周边监控视频,案发现场周边视频中含有大量数据资料,利用大数据技术可以快速

处理其中的信息,分析预测视频中行人的个人特点;虚拟现实仿真技术运用于足部压力和鞋底受力分析有限元模型,利用模型获得各种复杂场景下的仿真足迹。

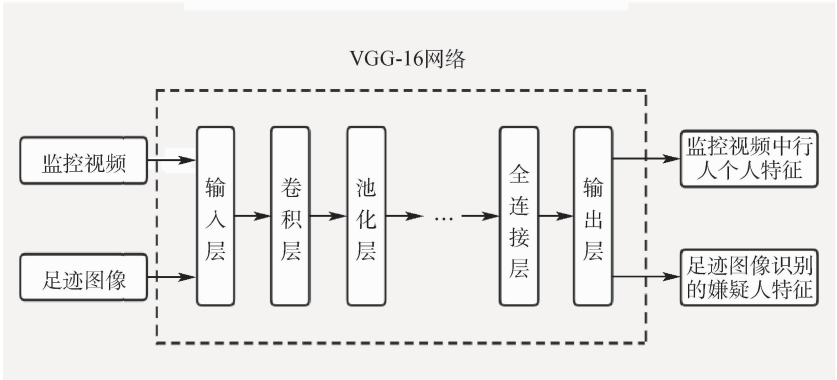
运用大规模针对复杂对象的哈希方法,将足迹特征、监控大数据和足迹仿真有机结合起来。三大模块相互交融,相互印证,有利于快速筛选刑侦对象,缩小排查范围,并辅助进行嫌疑人身份认定。本文主要研究了足迹特征模块和监控大数据模块。

2.2 足迹识别预测个人特征

足迹特征是地面、鞋和嫌疑人三者中多个因素相互作用的结果,犯罪现场以及实验环境下采样的足迹图像如图 2 所示。该系统使用深度学习模型确定足迹特征与嫌疑人个人特点之间的映射关系。因为足迹特征的波动往往较大,利用深度学习技术训练的卷积神经网络进行预测,具有很



(a) 系统整体框架



(b) 深度学习框架

图 1 基于深度学习的足迹识别技术路线

Fig. 1 Footprint recognition technology route based on deep learning



图 2 足迹图像采样

Fig. 2 Sampling of shoeprints

强的容错性,可以有效提高预测结果准确性;而且与传统预测方法相比,使用深度学习方法实现足迹识别能够完善足迹学理论。现有的足迹特征分析方法只是计算人为选择的一些特征值,如静态的鞋印长度、宽度和动态的步长、步角等,这些特征往往难以捕捉到完备的足迹信息;深度学习是特征无关的分析方法,能够更好地捕捉到足迹动态特征。神经网络通过将低维的足迹特征转化成更高维的状态向量,更容易完成足迹与个人特点之间的映射。由此建立基于深度卷积神经网络的身高预测模型,配合图像增广技术,在保证模型准确率和泛化能力的前提下进行预测。本文选择使用的卷积神经网络是 VGG-16 图像分类网络。VGG-16 结构去除池化层与回归层以后包含 16 个隐藏层,即 13 个卷积层和 3 个全连接层。VGG-16 采用连续的多个 3×3 的卷积核代替较大的卷积核,增加网络深度,提高神经网络的效果,而且代价小,参数更少。整个网络使用的都是 3×3 的卷积核和 2×2 的最大池化尺寸,结构非常简洁。

此外,通过虚拟现实仿真技术,建立足部压力和鞋底受力分析有限元模型;在犯罪现场采集地面相关参数,如地面硬度、弹性、时间周期等,并设置嫌疑人的相关参数,如体重、身高、脚型等;结合有限元模型,得到仿真足迹;对比仿真足迹和实际采样足迹之间的差异,优化嫌疑人相关预测参数,得到与采样足迹最匹配的仿真足迹,从而锁定嫌疑人的身高、体重等个人特点。

2.3 监控视频预测个人特征

提取足迹特征之后,使用深度学习技术训练卷积神经网络,使其利用计算机视觉从周边监控视频中提取行人的个人特点,预测个人特征,并建立人像图片库。

利用监控视频预测行人特征,首先需要进行

图像预处理,使用高斯滤波进行图像去噪,提高目标提取效果。其次使用前景提取算法进行运动行人的提取,减少视频中其他物体产生的干扰。目前使用较多的前景提取方法包括帧差法、平均法、高斯建模法等,考虑到视频图像处理速度快,算法操作简洁的特点,本文选择使用帧差法,该方法是将相邻帧图像做差运算,得到运动物体的大致轮廓,可以快速获得计算结果。最后进行行人的个人特征预测,本文系统着重于行人身高信息的预测。现有的通过监控视频获取人体身高信息的测量方法主要包括 2 种:未标定测量和标定测量^[17]。标定测量获得的结果精度很高,但是测量过程需要高精度的特殊标定块,以此获得监控拍摄的畸变系数,过程十分繁琐,需要耗费大量的时间。未标定测量精度稍差,但更加灵活,成本更低,不需要借助标定块,会主动根据拍摄场景中存在的实际固定物体(比如垃圾桶、消防栓等)或运动物体(比如车辆等)本身的结构化信息进行测量。本文系统选择使用未标定测量方法,根据透视学原理,以下测量公式成立:

$$\frac{h_1}{h_2} = \frac{l_1}{l_2}$$

式中: h_1 为实际行人身高; h_2 为视频图像中行人身高; l_1 为实际物体尺寸信息; l_2 为视频图像中物体尺寸信息。

利用数据集训练卷积神经网络,使其能够识别常见物体的长度、高度信息,比如各种车辆、消防栓、楼层高度、垃圾桶等。而后利用上述原理进行行人身高预测。

将监控视频中提取出的行人的个人特点与从犯罪现场足迹信息中获得的嫌疑人个人特点进行联合分析,进一步确定嫌疑人关键信息,从而快速筛选刑侦对象,缩小刑侦排查的范围。

3 实验过程与结果分析

3.1 足迹图像数据集建立

对每一例数据集,收集鞋印信息和对应的身高信息。如图 3 所示,将筛选出的 2350 张清晰图片,从 75 pixel/cm 下采样至 5 pixel/cm,降低数据维度。然后进行归一化处理(见图 4),去除标度尺,将鞋印居中,统一图像尺寸为 256×128 dpi,并增广 30 倍(见图 5)。对数据中的所有身高信息做统计分析,身高分布结果如图 6 所示。从所有足迹图像中,随机选择 261 张图像作为测试数据集,测试数据集约占数据总量的 1/9,测试数据集中鞋印长度和身高信息的对应关系如图 7 所示。

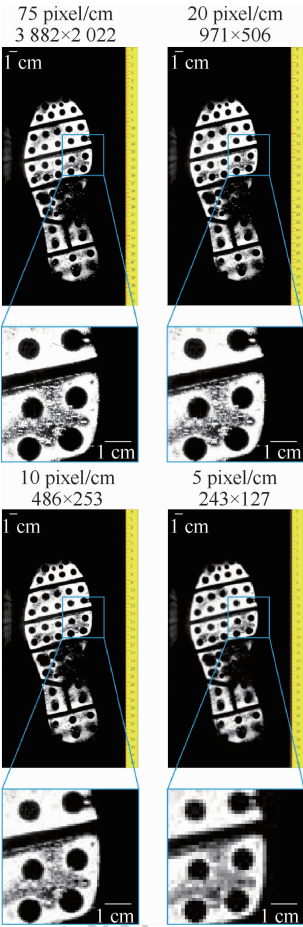


图 3 降低数据维度过程

Fig. 3 Data dimension reduction process

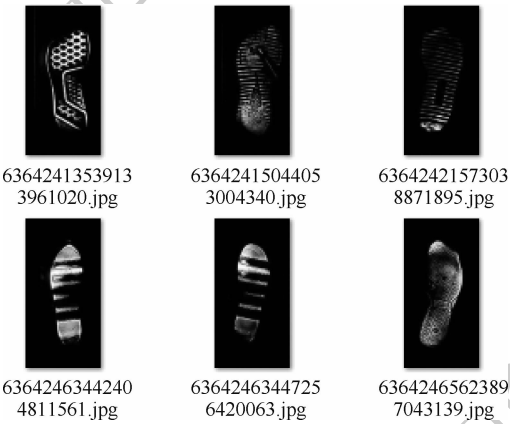


图 4 归一化处理结果

Fig. 4 Results after normalization

3.2 足迹识别预测身高

目前刑侦机关利用足迹长度预测身高主要依靠经验公式,预测结果与实际身高差值集中于 10 ~ 20 cm。该方法误差大,准确性低,在嫌疑人身份确认的主要特征数据提取方面尚存缺陷,对于实际刑侦工作帮助十分有限。

为解决上述问题,本文采用卷积神经网络技术建立足迹识别系统。该系统考虑到同一待

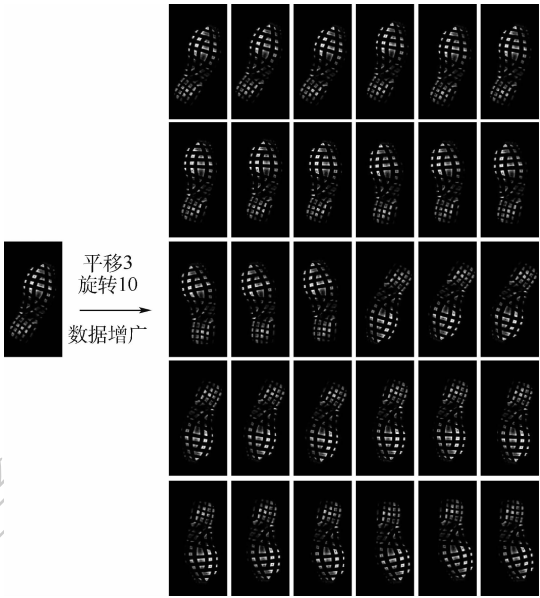


图 5 增广 30 倍结果

Fig. 5 Results after augmented 30 times

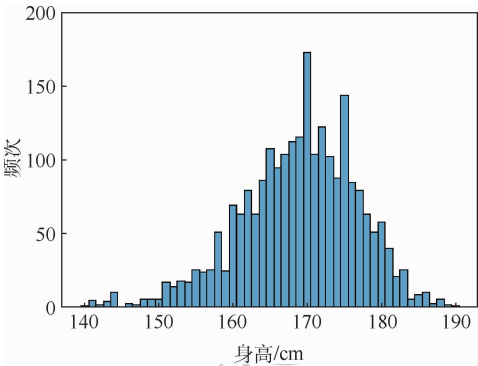


图 6 所有数据中身高的分布

Fig. 6 Distribution of height in dataset

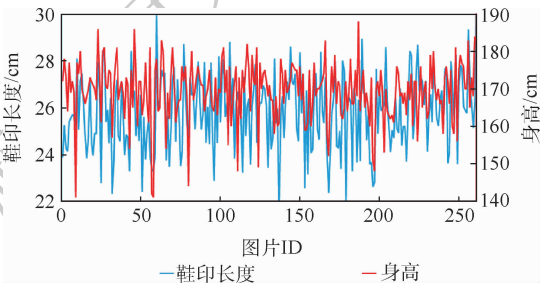


图 7 测试数据中身高与鞋印长度的关系

Fig. 7 Correlation between height and length of shoeprint in test data

测人员穿戴不同鞋型时,鞋的结构变化会使得足迹图像发生变动,改变图像参数。所以加入不同鞋型的选项,减小误差。为了提高预测结果的准确性和泛用性,加入数据增广选项,开启后会对输入的足迹图像进行平移和旋转的变换实现增广 30 倍。

系统界面如图 8 所示,打开足迹图像后,原始图像显示于左侧窗口;设置分辨率,进行图像预处

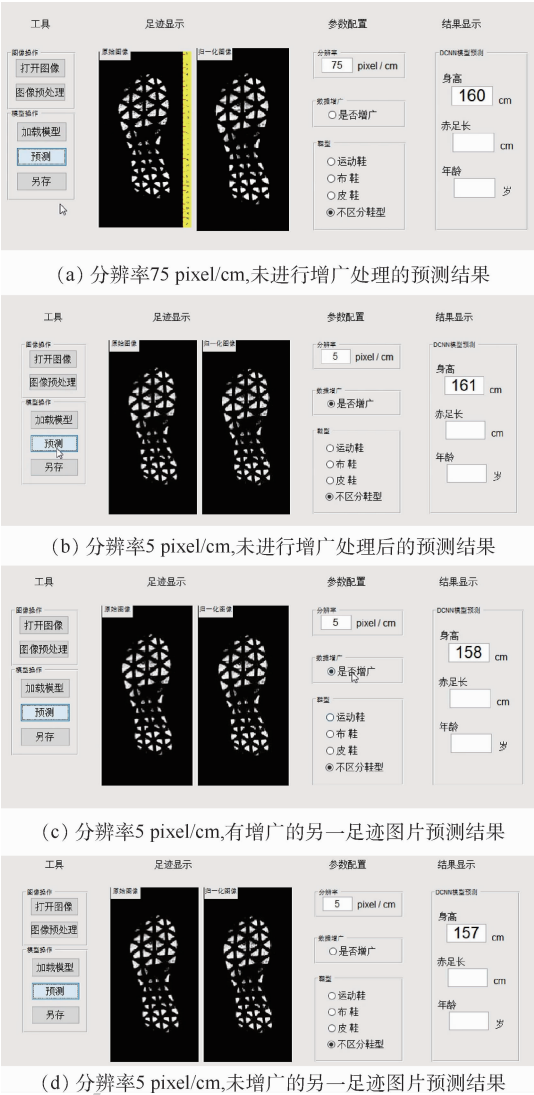


图 8 各种处理条件下的预测结果

Fig.8 Prediction results under various processing conditions

理,处理后的足迹图像在右侧窗口显示;而后选择是否区分鞋型和是否增广,点击“预测”按钮即可输出身高预测结果。

由图 8 所示结果可以发现,图像分辨率的改变对身高预测影响为 0~1 cm,是否增广与鞋型变动对身高预测值的改变约为 1~3 cm。

3.3 足迹识别性能分析

本文设计的足迹识别系统在不同条件下的误差范围如图 9 所示。可以发现随着误差阈值的提高,模型准确率增长趋于平缓,在误差阈值为 8 cm 时可以满足准确率大于 80%。模型对于身高 160~180 cm 之间的数据预测准确性高,可以保证 90% 的预测结果误差值在 10 cm 以下,且对身高 170 cm 左右的数据预测结果最稳定,误差值更小;而身高低于 160 cm 和高于 180 cm 的数据预测性一般,误差值集中在 10~20 cm。

比较训练误差和验证误差可得,相同误差阈

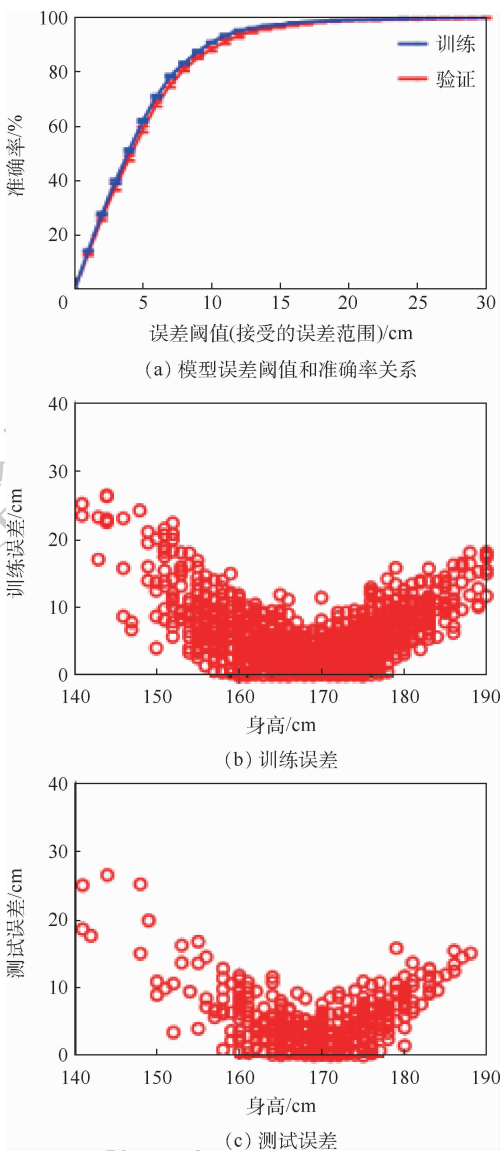


图 9 模型拟合结果、准确率和误差

Fig.9 Model fitting results, accuracy and deviation

值下,训练准确性稍高于测试准确性;测试误差值相比于训练误差值更加集中于 0~8 cm,误差值为 10~20 cm 的测试结果明显减少。

经验公式和最小二乘法对于数据的预测结果如图 10 所示。两者预测准确率随误差阈值上升而增加;预测结果误差值集中于 0~10 cm。

最小二乘法拟合得到的直线 ($y = 2.14x + 114$) 斜率为 2.14,相比于直线 ($y = 7x - 11.33$) 斜率为 7 的经验公式,更加贴合数据分布,误差有所降低。

比较足迹识别模型、经验公式和最小二乘法预测结果的准确率,得到如表 1 所示结果。从表 1 中可以看出,本文提出的足迹识别系统对于身高的预测准确率在 3 种方法中位居首位,与传统经验公式预测方法相比,可信度大幅提升。

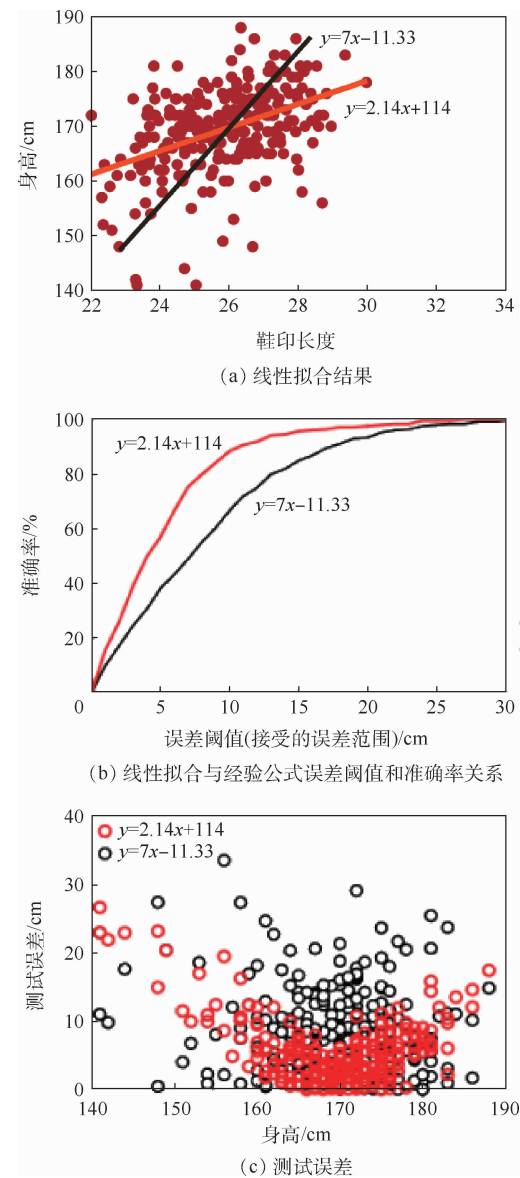


图 10 线性拟合结果、准确率和误差

Fig. 10 Linear fitting results, accuracy and deviation

表 1 三种方法预测结果准确率

Table 1 Accuracy of three methods for predicting results

方法	误差阈值/cm	训练准确率/%	测试准确率/%
本文	5	62.11 ± 0.97	58.98 ± 2.24
	10	90.97 ± 0.50	88.38 ± 1.60
经验公式	5		38.31
	10		66.67
最小二乘法	5		53.09
	10		84.51

3.4 足迹和监控身高预测联合筛查

利用卷积神经网络进行足迹识别预测,得到嫌疑人的身高特点,首先以此为基础,比对周边视频监控分析处理结果,筛查出具有相似身高特点的行人;其次以监控中筛查出的重点嫌疑人的个人特点为基础,最后确定足迹预测身高特点的准确性和可靠性,并进行修正(见图 11)。经过实验

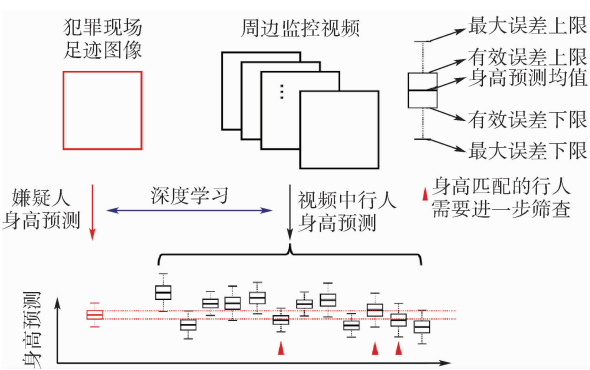


图 11 足迹识别系统和监控视频分析的有机结合

Fig. 11 Dynamic combination of footprint recognition system and surveillance video analysis

验证,确定了两者结合进行刑侦筛查的可行性。在将来实际使用中,两者可以相互印证,缩小可疑目标范围,协助刑侦工作高效准确进行。

未来可以在在数据收集上做进一步改进,并扩充数据集,分析在不同场景下运行的可靠性。

4 结 论

为有效配合刑侦工作的进行,形成高效准确的刑侦鉴定体系。本文提出了一种联合足迹识别与监控视频分析的智能刑侦系统。实验结果和分析表明,该系统存在以下两点优势:

- 1) 高效准确地提取足迹信息,并使用深度卷积神经网络技术预测嫌疑人身高,与传统经验公式相比,准确度提高 20% ~ 23%。
- 2) 将根据足迹特征预测的嫌疑人身高特点与周边视频监控数据分析结果有机结合,缩小嫌疑人排查范围,提高刑侦工作效率。

为了更好的提取足迹信息,下一步将考虑进行包含深度信息的 3D 足迹特征提取,分析足迹区域的压力分布,大幅提升后期分析和推理的准确性和可靠性。

参考文献 (References)

[1] 孙传森. 浅议犯罪现场足迹检验[J]. 经济研究导刊, 2009 (8) :253-254.

SUN C S. On the footprint inspection of crime scene[J]. Economic Research Guide, 2009 (8) :253-254 (in Chinese).

[2] ISHAK N I, HEMY N, FRANKLIN D. Estimation of stature from hand and handprint dimensions in a Western Australian population[J]. Forensic Science International, 2012, 216 (1-3) :199. e1-199. e7.

[3] ISHAK N I, HEMY N, FRANKLIN D. Estimation of sex from hand and handprint dimensions in a Western Australian population[J]. Forensic Science International, 2012, 221 (1-3) :154. e1-154. e6.

- [4] HEMY N, FLAVEL A, AISHAK N I, et al. Sex estimation using anthropometry of feet and footprints in a Western Australian population[J]. *Forensic Science International*, 2013, 231(1-3):402.e1-402.e6.
- [5] REEL S, ROUSE S, VERNON W, et al. Reliability of a two-dimensional footprint measurement approach[J]. *Science & Justice*, 2010, 50(3):113-118.
- [6] KRISHAN K. Individualizing characteristics of footprints in Gijjars of North India—Forensic aspects[J]. *Forensic Science International*, 2006, 169(2):137-144.
- [7] MOORTHY T N, MOSTAPA A, BOOMINATHAN R, et al. Stature estimation from footprint measurements in Indian Tamils by regression analysis[J]. *Egyptian Journal of Forensic Sciences*, 2014, 4(1):7-16.
- [8] 高毅. 基于Footscan足底压力步态分析系统的足迹动力形态特征研究[J]. *中国人民公安大学学报(自然科学版)*, 2011, 17(2):9-13.
GAO Y. Research on the dynamic morphology of footprints on the bases of gait analysis system of plantar pressure[J]. *Journal of Chinese People's Public Security University(Science and Technology)*, 2011, 17(2):9-13(in Chinese).
- [9] 振强. 马玉林学术研究会 in 内蒙古赤峰成立[J]. *刑事技术*, 1993(1):30.
ZHEN Q. Ma Yulin academic research association was established in Chifeng, Inner Mongolia[J]. *Criminal Technology*, 1993(1):30(in Chinese).
- [10] 吴大有. 足迹检验的历史与现状[J]. *中国刑事警察*, 1995(2):48-49.
WU D Y. History and current status of footprint inspection[J]. *Chinese Criminal Police*, 1995(2):48-49(in Chinese).
- [11] 刘树权, 叶剑平, 温媛, 等. 足迹个体识别技术自动化——足迹计算机自动鉴定系统[J]. *刑事技术*, 1998(1):44-46.
LIU S Q, YE J P, WEN Y, et al. Automation of individual footprint identification technology—Automatic footprint computer identification system[J]. *Criminal Technology*, 1998(1):44-46(in Chinese).
- [12] 王清举, 平西建, 王永栋. 立体足迹计算机自动识别系统的应用[J]. *刑事技术*, 2003(5):41-43.
WANG Q J, PING X J, WANG Y D. Application of automatic stereo footprint computer recognition system[J]. *Criminal Technology*, 2003(5):41-43(in Chinese).
- [13] 闫旭琴. 自动鞋印分析与比对系统中的关键技术的研究与实现[D]. 青岛: 山东科技大学, 2007.
YAN X Q. Research and implementation of key technologies in automatic shoe print analysis and comparison system[D]. Qingdao: Shandong University of Science and Technology, 2007(in Chinese).
- [14] 王永栋, 顾士清, 党素琴, 等. 基于重压面形状分析的足迹识别系统的研究[C]//中国图象图形学学会. 第十四届全国图象图形学学术会议论文集. 北京: 中国图象图形学学会, 2008:794-797.
WANG Y D, GU S Q, DANG S Q, et al. Research on footprint recognition system based on shape analysis of pressure surfaces[C]//Chinese Society of Image Graphics. Proceedings of the 14th National Conference on Image Graphics. Beijing: Chinese Society of Image Graphics, 2008:794-797(in Chinese).
- [15] 蒋晓玲. 基于图像处理技术的足迹检验系统的设计与实现[D]. 南京: 南京理工大学, 2011.
JIANG X L. Design and implementation of footprint inspection system based on image processing technology[D]. Nanjing: Nanjing University of Science and Technology, 2011(in Chinese).
- [16] 董菲菲. 基于三维鞋印的鞋底花纹分类及模块设计[D]. 南昌: 南昌大学, 2010.
DONG F F. Sole pattern classification and module design based on three-dimensional shoe prints[D]. Nanchang: Nanchang University, 2010(in Chinese).
- [17] 张彩霞, 付焕利. 未标定视频下的动态人体身高测量[J]. *计算机工程与应用*, 2017, 53(21):162-166.
ZHANG C X, FU H L. Dynamic human height measurement under uncalibrated video[J]. *Computer Engineering and Applications*, 2017, 53(21):162-166(in Chinese).

作者简介:

陶一宁 男, 硕士研究生。主要研究方向: 计算机视觉、机器学习。

袁培江 男, 博士, 副教授, 硕士生导师。主要研究方向: 计算机视觉、机器人。

Intelligent criminal investigation system based on both footprint recognition and surveillance video analysis

TAO Yining¹, SU Feng¹, YUAN Peijiang^{1,*}, WANG Tianmiao¹, ZHONG Tao², HAO Jingru²

(1. School of Mechanical Engineering and Automation, Beihang University, Beijing 100083, China;

2. Criminal Investigation Corps, Beijing Public Security Bureau, Beijing 100191, China)

Abstract: Criminal investigation is the basic requirement for cracking down on crimes and ensuring the long-term security of the country. One of the key points in criminal investigation is how to effectively use the collected information. In order to support criminal investigation better, an intelligent criminal investigation system based on both footprint recognition and surveillance video analysis is proposed. The system is powered by deep learning methods. First, using convolutional neural network technique, it predicts the characteristics of suspected individuals based on footprint information, including the footprint length, width, distribution of force, step length, stride, etc. Second, it uses the surrounding surveillance video big data for intelligent comparison to quickly screen criminal investigation objects and analyze the personal characteristics of pedestrians. Finally, the virtual reality simulation technology is used to construct a finite element model of foot pressure and sole stress analysis, and the model is used to obtain simulation footprints in various complex scenarios. The three confirmed each other and organically combined to quickly select criminal investigation targets. Experimental results show that the system can efficiently and accurately achieve people height prediction based on footprint characteristics. Combined with video surveillance big data, the proposed system can quickly narrow down the investigation range and find out the killer.

Keywords: deep learning; big data; criminal investigations; surveillance video analysis; footprint recognition

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0063

双光谱智能体温检测与健康大数据管理系统



张洁茹, 苏峰, 袁培江*, 王田苗, 陶一宁, 丁东

(北京航空航天大学 机械工程及自动化学院, 北京 100083)

摘 要: 公共安全视频监控在新型冠状病毒肺炎防治攻坚战中发挥了重要作用。针对中国人口密度高、人流量大、新型冠状病毒肺炎易传播的特点,建立了融合可见光和红外光双光谱成像监控的智能体温检测与健康大数据管理系统,实现了无接触快速体温检测与佩戴口罩情况下的人脸识别,快速完成人员信息登记。系统已在多地完成部署,通过了有效性和可靠性验证,测量速度快,响应时间在 30 ms 以内;测量精度高,测量温度误差在 $\pm 0.3^{\circ}\text{C}$ 以内;测量范围广,可监控距离 0.1 ~ 10 m;人脸抓拍率 99% 以上,识别率 95% 以上;健康大数据管理系统能实时监控和追踪回溯人员流动,在多维度上对人员信息和疫情发展大数据进行统计分析,并对疫情发展动态进行建模和预测,根据分析结果完善疫情防控策略,开展精准高效的疫情防控。

关 键 词: 深度学习; 视频监控; 双光谱; 体温检测; 疫情

中图分类号: TP391.4

文献标志码: A **文章编号:** 1001-5965(2020)09-1739-08

公共安全视频监控是中国一项重要的基础设施建设,具有易于部署、信息捕获充分、接口和标准统一、可全时段运行等优点,对于建立社会治安防控体系具有重要意义,是维护国家和社会安定的重要手段^[1]。近年来,面向图像分析和视频监控分析的深度学习技术和大数据挖掘技术取得了很多突破性进展,同时也推动了公共安全视频监控在治安防范、社会治理、服务民生等方面的应用^[2]。如在建立公共安全视频监控集成管理系统的过程中,通过规范联网、提高集成程度、统一数据接口、提升动态数据分析能力等方式,将人脸识别、车牌识别、智能预警、地理位置等众多信息有机融合在一起,形成全方位的公共安全监管体系,对于提高公安视频图像信息的综合应用水平具有重要意义^[3]。在新型冠状病毒肺炎(简称新冠肺炎)防治攻坚战中,将红外热成像技术和人脸识别技术融入到监控视频体系,从而实现无

接触式体温快速检测和人员信息登记,对于迅速遏制新冠肺炎的传播起到了重要作用,同时也印证了视频监控公共安全监管体系在治安防范和保障民生等方面的强效作用。

新冠肺炎的传播时值春节返乡之际,人员流动促使新冠肺炎迅速肆虐全国,对人民健康带来了严重威胁,对社会安稳带来了潜在的破坏因素,同时也对国家经济的发展带来严重影响^[4]。新冠肺炎潜伏期长、传染性强,甚至存在无症状感染的病例,能在人与人之间传播,在人员密集的公共场所极易引起爆发式的感染,故而打好新冠肺炎战役的一个重要举措就是加强实时检测和防控力度。

车站、机场、地铁站、商圈以及复工之后的办公楼等,都是人员密集、人流量大的场所,在人群中快速找出病毒潜伏者以及感染者,并进行隔离观察和治疗,对于个人方面的加速治愈以及社会

收稿日期: 2020-03-02; 录用日期: 2020-04-03; 网络出版时间: 2020-05-06 17:15

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200506.1423.005.html

基金项目: 国家自然科学基金(61603015)

* 通信作者: E-mail: itr@buaa.edu.cn

引用格式: 张洁茹, 苏峰, 袁培江, 等. 双光谱智能体温检测与健康大数据管理系统[J]. 北京航空航天大学学报, 2020, 46(9): 1739-1746. ZHANG J R, SU F, YUAN P J, et al. Dual-spectrum intelligent temperature detection and health big data management system[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1739-1746 (in Chinese).

方面的遏制病毒传播、保障民生都具有重要意义。钟南山院士团队对大量新冠肺炎患者的症状进行了深入分析,发现发热是新冠肺炎患者的显著临床症状,1099 名患者中发热症状高达 87.9%^[5]。因此,在客流量较大的公共场所,采用无接触体温快速检测方法,实现病毒潜伏者和感染者的筛查和预警,是最为快速有效的方式。

正常生理状态下,人体中枢神经系统会进行体温调节,使人体能够在环境温度变化的情况下保持体温的相对稳定。当体温下降时,则产热增加散热减少,使体温维持在正常水平(36.2 ~ 37.3℃);当体温升高时,通过减少产热和增加散热来维持体温恒定^[6]。病毒感染会激活人体的免疫系统,使吞噬细胞和嗜中性粒细胞的杀菌活性增强,导致体温上升至 37.3℃ 以上,从而产生发热的症状,发热是新冠肺炎患者的常见临床症状和检测指标。传统的体温测量方法分为接触式和非接触式两大类,而在新冠肺炎患者的体温检测中,玻璃温度计和电子温度计等接触式方法有很大的传染隐患^[7]。为防止交叉传染,目前常见的是手持式额温度仪为代表的非接触式测温方法。然而手持式额温度仪测量误差大,测量效率低,此外检测需要一名专门人员操作,仍然存在交叉传染的风险。因此,一种非接触、高精度、速度快的智能体温检测系统对于公共场所体温异常人员的识别具有重要的意义,是疫情防控的重要一环。此外,将人脸识别技术和智能体温检测相结合,并实时上传数据库,对发热人员身份识别并快速追溯人员近期踪迹,有效提高疫情防控水平。

为了提高人脸识别和体温检测的精度,本文基于深度学习、非接触红外热成像测温、智能 AI 终端识别和大数据技术,将红外热图像可解决人脸受光照影响的优势和可见光图像的高清晰度优势相结合,采用可见光和红外光双摄像头目标检测,对可见光图像与红外热图像进行融合,并通过卷积神经网络对人脸特征及人员是否佩戴口罩进行识别,将识别到的人员信息和体温信息进行整合,建立体温检测与健康大数据管理系统,对客流量较大的场所进行视频监控,快速准确地确定人员身份及人体温度,建立疫情登记档案,监控人员流动,实现无人值守条件下的全自动高精度数据分析。同时优化大数据系统,定制数据分析 APP 平台,实时向相关部门上报健康数据,保证信息数据、测温数据电子化、结构化,并给予大数据对疫情进行统计、分析和预测,做到准确定位,细致分

析,完善疫情防控策略,为疫情下的监控视频防控工作带来有效的帮助。

1 视频监控大数据

1.1 基于深度学习的目标检测

目标识别是人工智能领域的热门应用方向,是指用计算机实现人的视觉功能,它的研究目标是使计算机具有图像或者视频中认知目标和环境的能力^[8]。传统目标检测算法通常可以分为 3 个阶段:首先在待测的图像上提取一些感兴趣区域,其次从这些感兴趣区域中提取特征,最后使用训练分类器实现分类。然而这类算法识别的准确率低,计算复杂度大。近年来随着人工智能的发展,目标检测算法也逐渐从传统的人为设计特征转向了基于深度神经网络的检测技术;基于深度学习的目标检测算法可分为 2 个关键的子任务:目标定位和目标分类。目标定位即对待测图像中的感兴趣区域所在位置进行检测,输出物体的包围框或物体中心等参数表示物体的位置信息^[9]。目标分类任务负责判断前一步定位得到的图像区域中是否包含目标类别的物体,最终得到目标类别的物体出现在所选择图像区域中的概率。与传统目标检测算法比较,深度学习算法精度高,速度快,特别适用于复杂条件下的目标识别。目前常用的目标检测算法基于卷积神经网络,包括 Faster R-CNN、SSD、YOLO 等算法,其有助于从图像中识别出所需要的目标^[10]。根据目标的识别过程,又可以将上述算法分为 2 类,即 Two-Stage 和 One-Stage 目标检测算法。Two-Stage 是先由算法生成一系列作为样本的候选框,然后对候选区域进行分类和位置精修,再通过卷积神经网络进行样本分类;One-Stage 则没有生成候选框这一步骤,直接将目标框定位的问题转化为回归问题,在处理过程中一步到位^[11]。这 2 种算法原理上的差异直接导致其在性能上的优劣,Two-Stage 在检测准确率 and 定位精度上具有很大优势,One-Stage 在算法速度上具有优势。随着技术的不断发展,2 类算法都在不断地优化和改进,在识别精度和识别速度上均能达到可观的效果。

对人体进行快速非接触式体温检测,主要通过对人体额头表面温度进行测量,因此首先需要人脸在视频中的位置进行检测,这同样也是人脸识别的第一步,基于 Two-Stage 的高精度识别优势,本文采用 Mask R-CNN 对人脸区域进行检测,Mask R-CNN 算法在 Faster R-CNN 算法的基础上进行了优化,可以在每一个像素上表示目标的所

属类别,达到像素级别的特征点定位,可以更加精准地检测人脸目标区域,同时划分温度检测的感兴趣区域,将感兴趣区域映射到红外热图像上,在红外热图像上获取特定区域的温度,进而得到待测人员的体温。

人脸区域检测完成后,需对人脸特征进行识别,由于人脸图像类内差异大,类间差异小,本文采用可以较好提取人脸特征值的 VGG-Net 神经网络对人员开展基于面部特征的人脸识别,并与现有的人脸库进行比对,从而识别出人员身份^[12];此外,通过对人脸特征的识别还能推断出目标人员的性别、年龄等信息,将信息整合上传到和安防系统连接的大数据分析系统追溯行踪,准确发现疫情高发地区人员。

1.2 红外热像仪体温检测

当物体受热时,其内部分子和原子的运动随着温度的升高而加剧,而分子和原子受热从低能态跃迁至高能态并辐射出电磁波,称为热辐射,且该辐射与物体本身的温度成比例。红外热辐射的波长在 0.76 ~ 100 μm 之间,在电磁波连续频谱中的位置处于无线电波与可见光之间的区域^[13]。用红外热像仪对辐射信号进行采集、放大,得到模拟物体表面温度空间分布的信息,经过后期处理形成热图像视频信号,即红外热成像^[14]。红外热成像技术可以获得图像中每个区域的温度,通过红外光与可见光双摄像头同时对目标拍摄,并进行图像的配准,这样就能够避免图像中的背景干扰,准确划分感兴趣区域并获得目标温度。

利用高精度数字式医用红外热像仪 MTI-Economy-C,搭配高清可见光摄像头,实现双摄像头数据融合。医用红外热像仪温度检测精度高于普通商用红外热像仪,满足检测误差在 ±0.1℃ 范围内的要求,搭配高清可见光摄像头,实现高清图像和高精度热感应图像的映射。通过优化智能 AI 识别终端,利用双光谱融合技术实现异源图像的融合。采用 Mask R-CNN 作为目标识别和分割模型,实现像素级的图像分割并映射至热感应图像,精确查找目标测温部位。

先对测温系统进行标定,将黑体辐射源放置在红外热像仪视野范围内,对测量温度进行实时校正,将视频图像和待测人员体温对应显示,拟合标定曲线;当系统运行时,红外热图像将热辐射转化为灰度值,通过黑体辐射源标定得到的测温算法模型建立灰度值与温度之间的对应关系,从而实现测温功能。对人体温度进行检测时,遍历目标检测提取的感兴趣区域的每个像素点,为了保

证疫情期间对体温测量的不遗漏、不错判,严格对异常温度进行把控,将感兴趣区域的最高温度作为系统输出温度,可以有效避免外界遮挡面部区域,造成局部温度较低的情况。

2 双光谱体温检测与人员识别

2.1 系统框架

利用红外热成像及可见光成像获得图像信息,基于深度学习图像分析技术,建立以神经网络算法为核心的人体高精度实时测温系统,完成热成像体温筛查和人员身份识别;对超过正常体温人群,通过红外热像仪近距离对手臂成像,对人体体温进行复查,获得准确体温信息,并结合应急大数据一体化平台,实现实时追踪报警、情报轨迹跟踪,配合疫情管控,协助公安机关进行疑似病例排查。系统框架设计如图 1 所示。

系统主要通过以下方法提高检测精度:缩小传感器的测温范围,结合热成像的测温特性对传感器进行优化,采用深度学习的算法结合彩色图像信息和热成像图像信息分析图像,缩小成像范围,对可疑人群手臂红外成像检测,利用黑体对检测设备实时校准,实现高精度测量人体体温。从而实现非接触、高精度、高速度的测量。

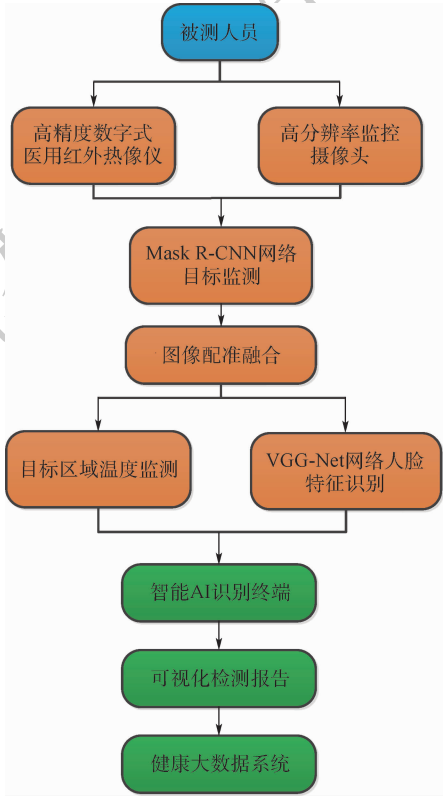


图 1 双光谱智能体温检测与健康大数据管理系统框架 Fig. 1 Architecture of dual-spectrum intelligent temperature detection and health big data management system

通过红外热成像系统检测人体热辐射的红外
线特定波段信号,将该信号转换成可视化、可分辨
的图像,并利用可见光成像系统获得 RGB 图像;
将得到的信息进行光照归一化、拍摄角度矫正及
模糊图像复原处理;将红外热像仪得到的红外图
像与可见光成像系统得到的 RGB 图像相结合,生
成四通道图像^[15],联动对测温区域进行判断,对
得到的图像基于神经网络的深度学习算法进行分
析;并利用测温算法模型补偿环境参数对测温结
果的影响,得到人体温度信息并识别人脸;对超过
正常体温的人员,采用红外热像仪近距离成像的
方法,增大图像目标识别区域,并通过对手臂成像
降低环境对人体表面体温的影响,从而提高测量
准确率。

2.2 可见光与红外光双光谱成像融合

红外热像仪采集待测物体表面产生的红外辐
射能量,经过能量收集系统汇聚后,将红外光波映
射至红外焦平面探测器上^[16],生成与待测物体热
辐射分布相应的热成像图像。可见光成像利用待
测物体的光谱反射特性形成二维图像,包含丰富
的纹理细节信息。对红外热成像与可见光成像生
成的图像进行融合,可以实现信息互补,使得物体
的热辐射分布特征与纹理细节外观特征有机结
合,识别性能不易受光照变化的影响,形成包含信
息更加丰富,更有利于识别处理的图像^[17]。

如图 2 所示,首先对前端视频进行预处理,包
括光照归一化、拍摄角度矫正、模糊图像复原、图
像比对等;将红外热图像和可见光图像配准;采用
Mask R-CNN 网络提取人体头部目标区域并映射

到红外热图像进行测温,然后利用图像融合算法
将可见光图像和红外热图像融合得到双光谱图
像,同时将融合后的图像送入卷积神经网络 VGG-
Net 中,结合彩色图像信息和热成像图像信息,对
人脸区域进行深度特征的提取,在检测目标区域
温度的同时识别人脸特征。

2.2.1 图像配准

为了在红外热图像中获取准确部位的人体温
度,需要对可见光图像与红外热图像进行配准。
其中可见光摄像头分辨率为 1920 像素 × 1080 像
素,红外摄像头分辨率为 324 像素 × 256 像素。
在图像配准之前,首先基于 Opencv 对双目相机进
行标定,获取相机内外参数,从而对图像进行畸变
校正,然后统一两幅图像的分辨率和大小,在不改
变图像清晰度的情况下对红外图像进行重采样,
将红外图像的水平分辨率和竖直分辨率分别重采
样为 1920 像素和 1080 像素,此操作只增加了数
据量,不影响图像清晰度。

由于可见光成像和红外成像的原理不同,且
拍摄时间不能完全同步、拍摄角度存在不同,因此
需要获得两幅图像的空间变换关系实现图像配
准。首先利用 Canny 算子对图像进行边缘提取,
当边缘线出现断点时,利用插值的方法使其联通,
由于可见光与红外热图像的特征点相对位置基本
不变,因此选用仿射变换作为图像空间变换方法,
仿射变换的模型可表示为

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

(1)

式中: (x_1, y_1) 和 (x_2, y_2) 分别为红外热图像和可
见光图像对应特征点的坐标; k_{11} 、 k_{12} 、 k_{21} 、 k_{22} 、 t_1 、 t_2
为 2 幅图像坐标之间的转换关系矩阵参数。为了
达到拟合最优解的目的,选取多个匹配点,利用最
小二乘法求取最佳变换参数,将可见光图像作为
基准图像,红外热图像向基准图像进行配准变换,
实现图像配准。

2.2.2 图像融合

为了消除光照变化对人脸区域产生的影响,
更好地识别人脸特征,将可见光和红外热图像进
行融合。图像融合算法直接影响融合后的图像质
量,由于人脸局部特征是由多个像素表征的,因此
选择基于区域的小波变换算法对图像进行融合,
融合基本流程如图 3 所示。

为了保留图像的彩色信息,首先将配准后的
可见光图像从 RGB 变换到 HIS 空间,并提取 I 分
量,对配准后的红外热图像进行灰度变换增强图
像效果,然后对可见光的 I 分量及变换后的红外热

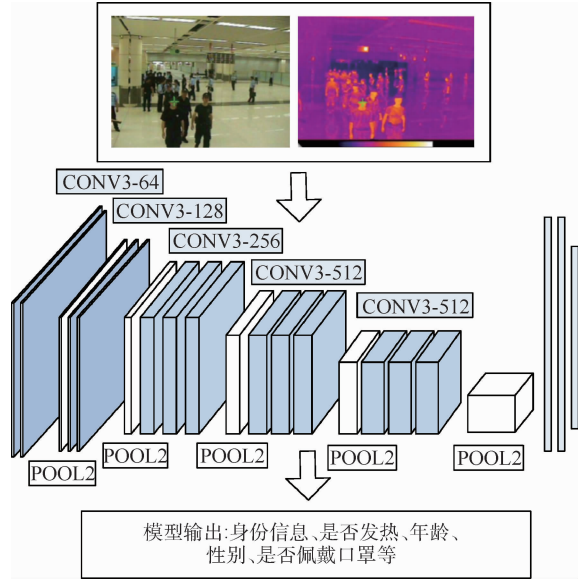


图 2 基于深度学习的双光谱体温检测与人脸识别
Fig. 2 Dual-spectrum temperature detection and
face recognition based on deep learning

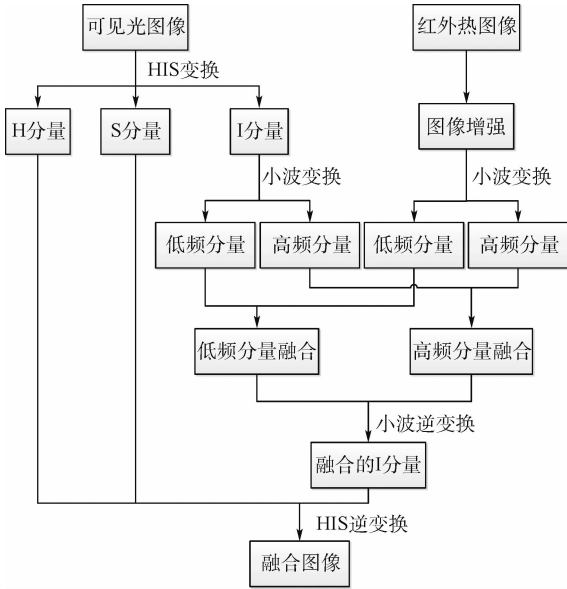


图 3 图像融合流程图
Fig.3 Image fusion flowchart

图像进行小波分解,得到各自的高频子图和低频子图,对高频和低频子图分别处理,其中低频子图包含了图像的轮廓信息,高频子图包含了图像的细节信息,对低频子图采用以下策略进行融合:当红外热图像和可见光图像的低频子图系数值相近时,取两者平均值作为融合后的低频系数值,反之,取其中较大值。对高频子图采用局域能量最大法进行融合,首先对各自的高频子图区域方差计算,得出区域能量,然后计算出红外热图像和可见光图像在各自高频子图中所对应的局部区域内的特征相似度,并设定一个匹配阈值,根据特征相似度和匹配阈值之间的关系,确定融合后的高频系数值。最后使用融合后的小波系数矩阵逆变换得到新的 I 分量,经过 HIS 逆变换得到融合图像。

2.3 监控视频效果分析

如图 4 所示,系统对视频中出现的多张人脸同时进行体温检测,并采集多张角度人脸照片,图中 2 名待测人员均佩戴口罩,其中有一名人员额头被遮挡,此时系统输出为该人员人脸检测区域未遮挡部分的最高温度,而另一名人员的检测输出温度为额头区域最高温度。此外,系统会对人员属性进行检测采集,可对人体年龄及性别进行判断,并支持视频回溯,可对视频中的人像以图搜图,有效提高疫情防控;获得数据可实时上传到 AI 应急防疫一体化平台,根据前端采集的多维数据,通过大数据平台分析后,提供搭建三维可视化地图。大数据记录事件发生事前、事中、事后的全过程数据,以便分析原因、持续优化、针对性改善以及模拟事件推演。

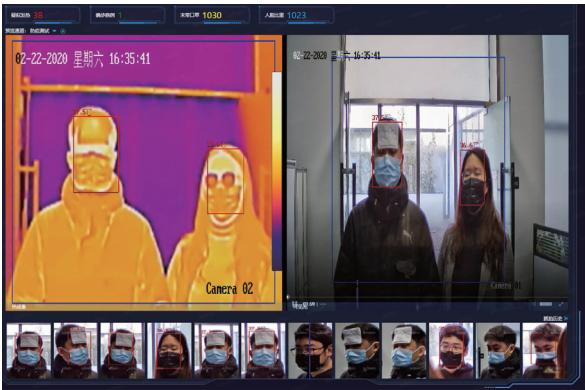


图 4 监控视频分析
Fig.4 Analysis of surveillance video

2.4 智能体温检测系统实施案例和可靠性验证

双光谱智能体温检测系统在监控视频下的快速体温检测 and 人脸识别实施案例如图 5 所示,系统对监控视频下的人员进行抓拍,将抓拍到的可见光图像和热成像图像进行配准融合,并对人脸区域检测划分,通过卷积神经网络进行人脸识别、体温检测及是否佩戴口罩预警,并将人脸检测、人脸识别、体温测量等结果反映在软件界面上;图中人员体温检测结果为 36.2℃,与人脸底库的特征相似度为 93%。系统支持在断网模式下单机运行,断网模式下储存的数据可以后期上传至平台数据库,也可以在联网模式下将监控数据、设备比对结果及抓拍照片实时上传给平台保存。设计实验对不同人员在不同场景(场景 1:大型商场入口,场景 2:校园食堂入口,场景 3:工业园区入口)下进行多次体温检测,实验数据如表 1 所示。通过实验验证可得:

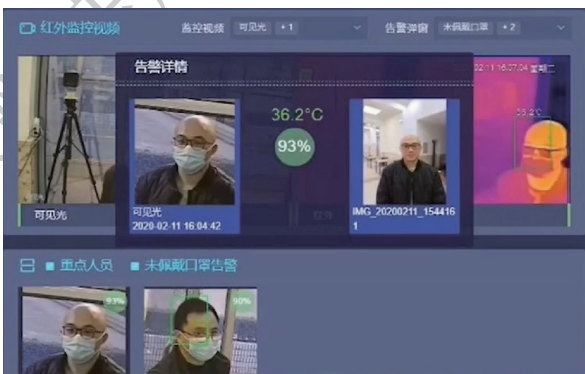


图 5 双光谱智能体温检测系统快速体温检测与人脸识别
Fig.5 Fast temperature detection and face recognition by dual-spectrum intelligent temperature detection system

1) 多目标体温检测的响应时间在 30 ms 以内,满足人员流动区域的快速测温要求;实现了对经过检测区域的被检测人流非接触式动态检测,解决了接触式测温效率低且存在感染风险的问

表 1 不同场景下的体温检测情况

Table 1 Temperature detection in different scenes

人员	系统响应 时间/ms	体温测量误差/℃			戴口罩人脸 识别情况
		场景 1	场景 2	场景 3	
A	30	-0.1	0	-0.3	成功
B	20	0.1	-0.2	-0.1	成功
C	20	0.2	-0.1	-0.2	成功
D	25	0.2	-0.1	-0.1	成功
E	30	0.2	0.3	-0.2	成功
F	25	-0.2	0.2	-0.3	成功
G	20	-0.3	0	-0.2	成功

题;其测量温度误差在 $\pm 0.3^{\circ}\text{C}$ 以内,适应 $1.4 \sim 2.0\text{ m}$ 的身高范围。

2) 可实现全域范围内的各类疫情卡点、单位、公共场所、工作场所等的证件、视频、人像、手机、温度信息的全面感知和统一汇聚。建立城市出入卡口、市内关键路段与场所等的无死角、无遗漏、闭环式证件、人像、车轨、视频、手机、温度信息的采集,形成全市统一的人像视频资料库和泛感知信息库等。

3) 该系统基于深度学习算法进行人脸识别,可同时对人体是否戴口罩、画面中的人体性别进行识别;并结合大数据实施未防护预警、发热预警、未防护人员数据分析、疫区人员布控、人员轨迹分析、人员身份鉴别等。识别条件要求最小瞳距 30 像素,面部无大面积遮挡,露出鼻梁及以上区域,上下俯仰角小于 45° ,左右侧脸角度小于 45° 。正常使用情况下人脸抓拍率 99% 以上,对于佩戴口罩情况下的人脸识别,当人脸底库为最近采集的现场照片时,识别率可达到 95%;若人脸底库为身份证底库照片,识别率可达到 90% 以上。

4) 环境温度会对人体温度造成影响,如在冬天的室外环境下,由于人体额头暴露在外,表面温度会变低,因此,在此环境下进行室外量测体温,测出的温度可能会偏低。即由于人体在不同环境下的体表温度发生变化,导致测出的温度与人体实际温度有偏差,但这并不是测温系统本身在不同温度下的测量准确性造成的。

2.5 健康大数据管理系统

结合双光谱智能体温检测结果,建立健康大数据管理系统,系统框架如图 6 所示。健康大数据管理系统能实时监控和追踪回溯人员流动,在多维度的对人员信息和疫情发展大数据进行统计分析,实现地域、年龄、性别、时间分布、人员流动等多维度上的疫情状态可视化,并对疫情发展动

态进行建模和预测,根据分析结果完善疫情防控策略,开展精准高效的疫情防控。

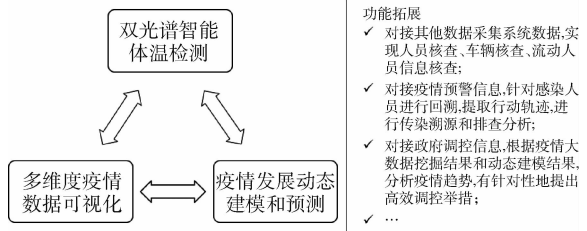


图 6 健康大数据管理系统框架

Fig. 6 Architecture of health big data management system

3 结 论

1) 本文建立了融合可见光和红外光双光谱智能体温检测与健康大数据管理系统,既提高了温度检测区域的精准性,又避免了光照变化对人脸识别的影响。利用红外热成像技术实现非接触式快速体温检测,针对目前疫情需求,无需摘下口罩即可进行人脸识别;同时对人员属性及是否佩戴口罩等进行识别并实时上传至数据库,快速完成人员信息登记。

2) 系统已在多地完成部署,通过了有效性和可靠性验证,测量速度快,响应时间为 30 ms 以内;测量精度高,测量温度误差在 $\pm 0.3^{\circ}\text{C}$ 以内;测量范围广,可监控距离 $0.1 \sim 10\text{ m}$;人脸抓拍率 99% 以上,识别率 95% 以上。

3) 将检测结果融入到健康大数据管理系统,实时跟进疫情的动态趋势,方便快捷地实现疫情发展的宏观把控,对于疫情防治具有指导意义。

参考文献 (References)

[1] 颜培超. 浅析公共安全视频监控大数据融合应用[J]. 数字技术与应用, 2019, 37(9): 30-31.
YAN P C. Analysis on the application of large data fusion in public security video surveillance [J]. Digital Technology and Application, 2019, 37(9): 30-31 (in Chinese).
[2] 黄凯奇, 陈晓棠, 康运锋, 等. 智能视频监控技术综述[J]. 计算机学报, 2015, 38(6): 1093-1118.
HUNG K Q, CHEN X T, KANG Y F, et al. Intelligent visual surveillance: A review [J]. Chinese Journal of Computers, 2015, 38(6): 1093-1118 (in Chinese).
[3] 马培才. 视频交换共享平台方案及应用场景研究[J]. 数字技术与应用, 2018, 36(6): 78-79.
MA P C. Research on public security video sharing platform and application scenarios [J]. Digital Technology and Application, 2018, 36(6): 78-79 (in Chinese).
[4] 赵路. 加强我国公共卫生管理的若干建议[J]. 中国科学院院刊, 2020, 35(2): 190-194.
ZHAO L. Suggestions on strengthening public health manage-

- ment in China[J]. Bulletin of the Chinese Academy of Sciences, 2020, 35(2): 190-194 (in Chinese).
- [5] GUAN W J, NI Z Y, HU Y, et al. Clinical characteristics of 2019 novel coronavirus infection in China[J]. The New England Journal of Medicine, 2020, 382: 1708-1720.
- [6] 冯云, 张旭光, 胡开强. 人工髋关节与膝关节置换术后人体体温差异性的研究分析[J]. 西北国防医学杂志, 2016, 37(6): 409-411.
- FENG Y, ZHANG X G, HU K Q. Analysis of the difference in human body temperature after artificial hip and knee replacements[J]. Northwest Journal of National Defence Medical Science, 2016, 37(6): 409-411 (in Chinese).
- [7] 王刚, 高晓丁, 唐明飞, 等. 基于无线传感器网络的人体体温监测系统[J]. 纺织高校基础科学学报, 2012, 25(1): 111-113.
- WANG G, GAO X D, TANG M F, et al. The infrared temperature detection system based on wireless net[J]. Journal of Basic Sciences of Textile Universities, 2012, 25(1): 111-113 (in Chinese).
- [8] 高淑雅, 高跃清. 一种基于卷积神经网络的雷达目标分类方法[J]. 信息技术, 2020, 44(1): 91-94.
- GAO S Y, GAO Y Q. Radar target classification based on convolutional neural network[J]. Information Technology, 2020, 44(1): 91-94 (in Chinese).
- [9] 张全, 盛妍, 吴佐平, 等. 公共区域监控视频数据目标特征跟踪定位方法[J]. 自动化与仪器仪表, 2020(1): 51-54.
- ZHANG Q, SHENG Y, WU Z P, et al. Target tracking and location method for common area surveillance video data[J]. Automation and Instrumentation, 2020(1): 51-54 (in Chinese).
- [10] 寇大磊, 权冀川, 张仲伟. 基于深度学习的目标检测框架进展研究[J]. 计算机工程与应用, 2019, 55(11): 25-34.
- KOU D L, QUAN J C, ZHANG Z W. Research on progress of object detection framework based on deep learning[J]. Computer Engineering and Application, 2019, 55(11): 25-34 (in Chinese).
- [11] 秦川. 基于卷积神经网络的图像识别[J]. 电子技术与软件工程, 2020(1): 98-99.
- QIN C. Image recognition based on convolutional neural network[J]. Electronic Technology and Software Engineering, 2020(1): 98-99 (in Chinese).
- [12] 徐晓艳. 人脸识别技术综述[J]. 电子测试, 2015(10): 30-35.
- XU X Y. Survey of face recognition technology[J]. Electronic Test, 2015(10): 30-35 (in Chinese).
- [13] 师立国. 浅议红外测温仪工作原理及应用[J]. 东方企业文化, 2013(19): 213.
- SHI L G. Discussion on the working principle and application of infrared thermometer[J]. Oriental Enterprise Culture, 2013(19): 213 (in Chinese).
- [14] 李晓梅, 耿晓光. 红外热成像技术及其在水务视频监控上的应用[J]. 工业控制计算机, 2014, 27(4): 76-77.
- LI X M, GENG X G. Application of infrared thermal imaging technology in video monitoring on the water[J]. Industrial Control Computer, 2014, 27(4): 76-77 (in Chinese).
- [15] 唐传佳, 周波, 王克杰, 等. 一体化红外测温监测技术研究与应用[J]. 安徽电气工程职业技术学院学报, 2017, 22(1): 118-121.
- TANG C J, ZHOU B, WANG K J, et al. Research and application of integrated infrared temperature measurement monitoring technology[J]. Journal of Anhui Vocational College of Electrical Engineering, 2017, 22(1): 118-121 (in Chinese).
- [16] 曹彦鹏, 许宝杯, 何泽威, 等. 红外热成像信号处理技术的研究进展[J]. 振动、测试与诊断, 2018, 38(2): 219-227.
- CAO Y P, XU B B, HE Z W, et al. Research advances in infrared thermography signal processing technology[J]. Vibration, Test and Diagnosis, 2018, 38(2): 219-227 (in Chinese).
- [17] 李霄. 可见光图像与红外图像融合技术研究[D]. 成都: 电子科技大学, 2014.
- LI X. Technical research of image fusion about visible and infrared image[D]. Chengdu: University of Electronic Science and Technology of China, 2014 (in Chinese).

作者简介:

张洁茹 女, 硕士研究生。主要研究方向: 计算机视觉、机器学习。

袁培江 男, 博士, 副教授, 硕士生导师。主要研究方向: 计算机视觉、机器学习、机器人。

Dual-spectrum intelligent temperature detection and health big data management system

ZHANG Jieru, SU Feng, YUAN Peijiang*, WANG Tianmiao, TAO Yining, DING Dong

(School of Mechanical Engineering and Automation, Beihang University, Beijing 100083, China)

Abstract: Public safety video surveillance has played an important role in the battle against Corona virus disease 2019. Aimed at the characteristics of high population density, large flow of people, and the easy spread of Corona virus disease 2019 in China, an intelligent temperature detection and health big data management system combining visible and infrared dual-spectral imaging monitoring is established to achieve contactless rapid temperature detection and face recognition while wearing a mask, and to quickly complete the registration of personal information. The system has been deployed in multiple places, and has passed the verification of effectiveness and reliability. The measurement speed is fast and the response time is within 30 ms. The measurement accuracy is high and the measurement temperature error is within $\pm 0.3^{\circ}\text{C}$. The measurement range is wide and the monitoring distance is 0.1–10 m. The face capture rate is over 99%, and the recognition rate is over 95%. The health big data management system can monitor and track back-to-back personnel movements in real time, perform statistical analysis on personnel information and epidemic development big data in multiple dimensions, conduct epidemic development trends modeling and prediction, improve epidemic prevention and control strategies based on the analysis results, and carry out accurate and efficient epidemic prevention and control.

Keywords: deep learning; video surveillance; dual-spectrum; temperature detection; epidemic situation

Received: 2020-03-02; Accepted: 2020-04-03; Published online: 2020-05-06 17:15

URL: kns.cnki.net/kcms/detail/11.2625.V.20200506.1423.005.html

Foundation item: National Natural Science Foundation of China (61603015)

* Corresponding author. E-mail: itr@buaa.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0075

基于空间注意力机制的行人再识别方法



张子昊^{1,2}, 周千里^{1,2,3}, 王蓉^{1,2,*}

(1. 中国人民公安大学 警务信息工程与网络安全学院, 北京 100038;

2. 安全防范技术与风险评估公安部重点实验室, 北京 100038; 3. 北京市公安局, 北京 100740)

摘 要: 行人再识别是图像检索领域的一个重要部分,但是由于行人姿态各异、背景复杂等因素,导致提取到的行人特征鲁棒性和代表性不强,进而影响行人再识别的精度。在 AlignedReID++ 算法基础上,提出了基于空间注意力机制的行人特征提取方法,应用在行人再识别中取得了很好的效果。首先,在特征提取部分,引入空间注意力机制来增强特征表达,同时抑制可能的噪声;其次,通过在卷积层中引入实例正则化层(IN)来辅助批正则化层(BN)对特征进行归一化处理,解决单一 BN 层对特征色调变化以及光照变化的不敏感性,提高特征提取对亮度、色调变化的鲁棒性;最后,在 Market1501、DukeMTMC 和 CUHK03 3 个行人再识别通用数据集上对所提改进模型进行测试评价。实验结果显示:改进后的模型在 3 个数据集上识别精度分别提升了 2%、2.9% 和 5.1%,表明改进后的模型相较于改进前的模型,在精度以及鲁棒性上都有显著提高。

关 键 词: 深度学习; 空间注意力机制; 行人特征; 特征增强; 卷积神经网络

中图分类号: O235; TP183

文献标志码: A

文章编号: 1001-5965(2020)09-1747-09

利用视频监控进行案件侦破时,多数场合只能看清行人的身体部分,同时需要进行跨视域的连续跟踪查找,因此行人再识别对于查获犯罪嫌疑人及可疑人员,寻找发现侦查破案所需的线索与信息至关重要。行人再识别核心思想是利用行人的整体特征对没有视野重叠覆盖的跨摄像机行人图像间建立对应关系。

依据提取特征方式不同,行人再识别可分为基于特征表达、基于度量学习以及基于深度学习等方法。基于特征表达的方法主要是通过提取行人图像的颜色、纹理、形状和属性特征来对行人进行再识别,但这些浅层的人工特征易受到行人姿态以及行人图片光照、分辨率的影响。基于度量学习的方法是对特征进行相似度度量,主要依靠

的是欧氏距离^[1]以及马氏距离^[2],然而,由于受行人图像复杂背景的影响,使得基于度量学习在行人匹配上遇到了识别精度不高的瓶颈。基于深度学习的方法将特征表达与度量学习结合在一起,实现了端到端的行人再识别,此类方法通过海量的数据来训练模型,使提取到的行人特征更具泛化性。Yi 等^[3]利用一个孪生网络^[4],将输入图片分割为 3 个重叠的水平条纹,并使用卷积层分别提取 3 个条纹的特征,最终融合成为一个特征向量作为输出,然后再进行相似度比较。随着全局特征遇到瓶颈,学者开始将目标转到局部特征。Varior 等^[5]采用图像切块的方式,将被分割好的若干块图像块按照顺序送到一个 LSTM 网络,网络最后输出的全局特征融合了所有图像块的局部

收稿日期: 2020-03-03; 录用日期: 2020-03-27; 网络出版时间: 2020-04-02 17:07

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200402.1658.001.html

基金项目: 国家重点研发计划(A19808); 中央高校基本科研业务费专项资金(2019JKF111)

* 通信作者。E-mail: dbdxwangrong@163.com

引用格式: 张子昊, 周千里, 王蓉. 基于空间注意力机制的行人再识别方法[J]. 北京航空航天大学学报, 2020, 46(9): 1747-1755.
ZHANG Z H, ZHOU Q L, WANG R. Pedestrian re-identification method based on spatial attention mechanism[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1747-1755 (in Chinese).

特征。为了解决局部特征对齐的问题,Zhang 等^[6]提出了一种端到端的方法 AlignedReID,让网络自动学习人体对齐,在提取局部信息时采用动态匹配的方法选取最短路径从而进行行人对齐。为了解决行人再识别数据不足的问题,Deng 等^[7]通过引入 CycleGAN 来实现不同摄像头下图片风格的转化,并将生成的图片用于训练。Luo 等^[8]在 AlignedReID 基础上提出了动态匹配局部信息(DMLI)的新方法,即 AlignedReID++,该方法可以动态调整局部信息,无需额外的监督。

本文在 AlignedReID++基础上,提出了一种改进的行人特征提取方法。该方法通过引入空间注意力机制,利用特征的空间特性动态调整特征表达,使提取到的特征更具有全局性和泛化性。同时,又引入了实例正则化(Instance-Normalization,IN)模块来辅助批正则化(Batch-Normalization,BN)层对特征进行归一化处理,提高网络对颜色、纹理特征变化的鲁棒性。

1 AlignedReID++ 算法

AlignedReID++算法主要包括特征提取和相似度度量 2 个部分,如图 1 所示。在特征提取部分,将 256×128 大小的行人图片通过 ResNet50 网络进行特征提取,并对提取到的特征分别输入给全局分支和局部分支。在全局分支中使用全局池化层来进一步提取全局特征。在局部分支中使用水平池化层,将行人图片从水平方向进行划分,并对每一条划分后的图片条纹进行局部特征提取。在相似度度量部分,分别对提取到的局部特征和全局特征进行特征间的距离计算。对于全局分支提取到的全局特征,使用 L2 距离来计算全局距离,其表达式为

$$d_g(A,B)=\|f_A-f_B\|_2 \tag{1}$$

式中: f_A 和 f_B 分别为 2 幅行人图片的特征向量,通过计算 2 幅图像特征向量之间的 L2 距离来计算 2 幅图像的相似度。对于局部特征部分,采用动态匹配局部特征的方法解决行人图片姿态各异无法对齐的问题,对于一幅行人图像的多个水平局部特征,动态地从上至下匹配局部部分,寻找拥有最短距离的对齐。对于 2 幅图像的局部特征 $I_A=\{l_{A1},l_{A2},\cdots,l_{AH}\}$ 和 $I_B=\{l_{B1},l_{B2},\cdots,l_{BH}\}$,首先使用 Element-wise 的转化,将距离归一化到[0,1)之间,如下:

$$d_{i,j}=\frac{e^{\|l_A^i-l_B^j\|_2^{-1}}}{e^{\|l_A^i-l_B^j\|_2^{-1}}+1} \quad i,j\in 1,2,\cdots,H \tag{2}$$

式中: $d_{i,j}$ 为第一张图片第 i 个数值条与第 2 张图片的第 j 个数值条的距离,这些距离组成了距离矩阵 D 。动态的局部距离则定义为在距离矩阵 D 中,从(1,1)到(H,H)总距离的最短路径。局部距离动态规划计算过程如下:

$$S_{i,j}=\begin{cases} d_{i,j} & i=1,j=1 \\ S_{i-1,j}+d_{i,j} & i\neq 1,j=1 \\ S_{i,j-1}+d_{i,j} & i=1,j\neq 1 \\ \min(S_{i-1,j},S_{i,j-1})+d_{i,j} & i\neq 1,j\neq 1 \end{cases} \tag{3}$$

式中: $S_{i,j}$ 为距离矩阵 D 从(1,1)到(i,j)的总距离,而 $S_{H,H}$ 则表示 2 幅图片最短路径的总距离,其计算公式为

$$d_l(A,B)=S_{H,H} \tag{4}$$

所以,2 幅图像之间总的距离可以表示为局部距离与全局距离之和:

$$d(A,B)=d_g(A,B)+\lambda d_l(A,B) \tag{5}$$

式中: $d_l(A,B)$ 为局部距离; λ 为平衡全局距离以及局部距离的权重系数,这里取值为 1。训练过程中使用 Triplet 损失^[9]作为度量学习的损失,同

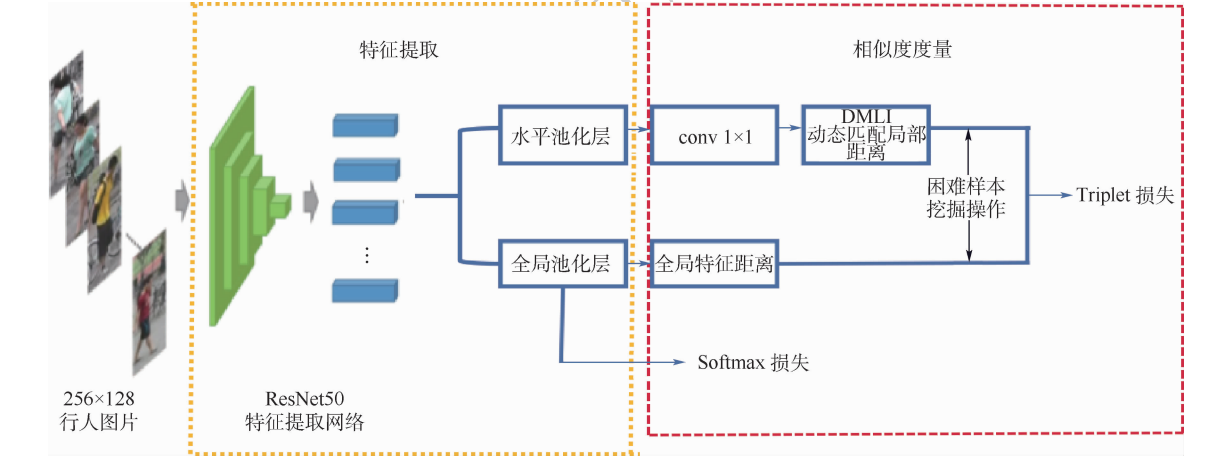


图 1 AlignedReID++ 算法流程图

Fig.1 AlignedReID++ algorithm flowchart

时在全局分支中,使用 Softmax 损失来辅助分类, AlignedreID++ 的整体损失函数为

$$L = L_{ID} + L_T^g + L_T^l \tag{6}$$

式中: L_T^l 、 L_{ID} 和 L_T^g 分别为局部分支的 Triplet 损失、全局分支的 Softmax 损失和 Triplet 损失,将 3 个损失的和作为整体的损失函数,使用随机梯度下降优化。

2 基于空间注意力机制的行人特征提取

通常来说,在 AlignedReID++ 算法中提取到的行人特征是由许多子特征组成,这些子特征以组的形式分布在卷积层中的每一层,而子特征易受背景噪声的干扰,会导致错误的识别和定位。BN 层^[10]是将一个 batch 的数据进行归一化,增加了不同特征语义之间的差异性,但是当训练数据集和测试数据集差异较大时,单一的 BN 层会影响测试的结果。

针对上述问题,本文提出一种改进的行人特征提取方法,其框架如图 2 所示。图中: x 、 $\hat{x}$ 、 C 、 H 、 W 分别为空间注意力模块处理前每一个组的特征向量、经过空间注意力模块处理后的每一个组的特征向量、特征图的通道数量、特征图的高度以及特征图的宽度; g 为经过全局平均池化(Global Average Pooling, GAP)后的特征向量; a

为经过归一化后得到的特征向量矩阵; $F_{gp}(\cdot)$ 为特征向量经过全局平均池化的过程。整体框架与 AlignedReID++ 类似,包括特征提取和相似度度量 2 个部分。在特征提取网络 ResNet50 的基础上,引入空间注意力模块 SGE(Spatial Group-wise Enhance)^[11],根据每个特征在空间上的位置来生成空间因子,动态分配每个特征权重,增强有用特征,抑制可能的噪声。同时,为了进一步增强卷积网络对于不同风格行人图片的泛化性与鲁棒性,在 BN 层^[10]的基础上引入 IN 层,组成 IBN(Instance-batch Normalization)模块,使得网络性能得到进一步的优化。

2.1 空间注意力模块

SGE 模块通过在每个组里生成注意力向量,根据每个子特征在空间位置上的重要性,动态地分配权重,同时有针对性地学习和抑制噪声。

SGE 模块首先将特征图按通道维度分为 G 个组,对于每一个组进行全局平均池化得到池化后的特征向量 g ,计算过程为

$$g = \frac{1}{m} \sum_{i=1}^m X_i \tag{7}$$

式中: m 为特征向量的维度; X_i 为每一个组中的特征向量。之后对于特征向量中的每一个元素,将其与 g 进行逐元素点乘运算,得到每一个位置对应的注意力权重为

$$c_i = gx_i \tag{8}$$

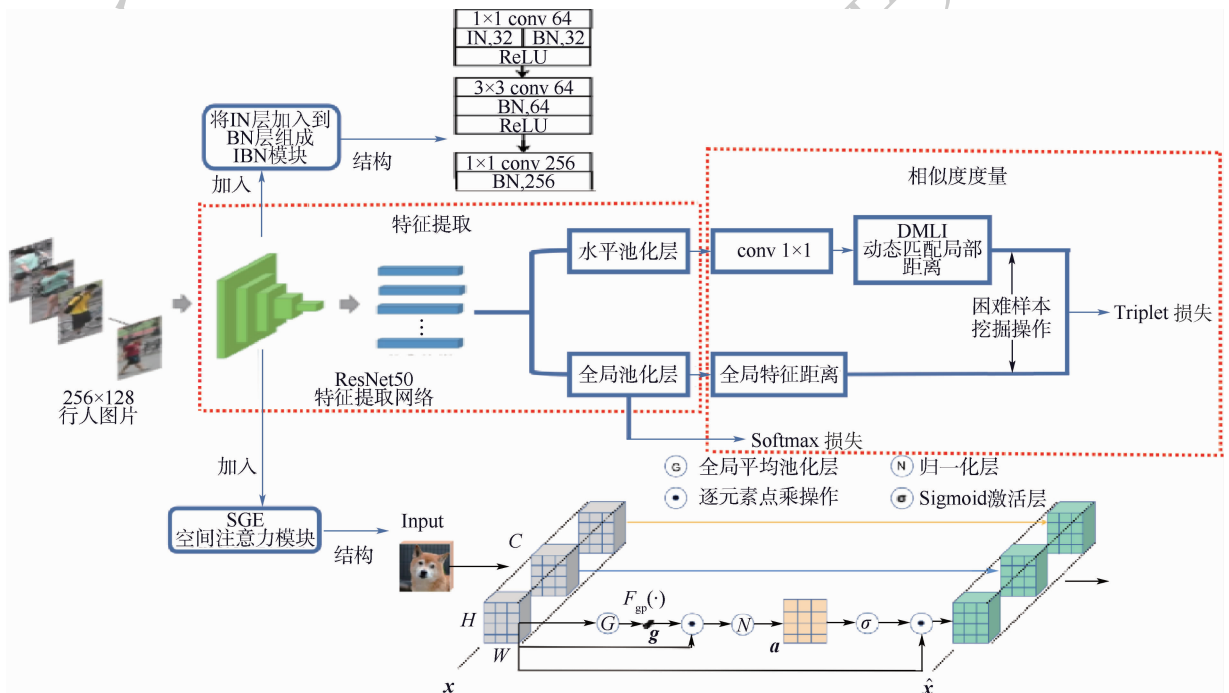


图 2 改进后 AlignedReID++ 算法流程图
Fig. 2 Improved AlignedReID++ algorithm flowchart

之后,对逐元素点乘后的结果进行归一化处理,使特征向量中数值分布在(0,1)之间,并将处理过后的结果经过 Sigmoid 激活层,最后,将经过激活层的结果与原组中的特征进行逐元素点乘的运算,得到对于一个组生成的空间注意力初始值,计算过程为

$$\hat{x}_i = x_i \sigma(a_i) \quad (9)$$

式中: $\hat{x}_i$ 和 x_i 分别为空间注意力模块处理后的每一个组的特征向量和处理前的每一个组的特征向量; $\sigma(a_i)$ 代表 Sigmoid 激活层, a_i 为经过归一化处理后的特征矩阵。

将空间注意力模块加入到 ResNet50 特征提取网络中,使得改进后的特征提取网络,提取到的特征更具代表性和泛化性,加入空间注意力机制 SGE 模块后的模型具体结构如图 3 所示。图中: x 为未经过残差模块处理的特征向量, $\tilde{x}$ 为经过残差模块处理后的特征向量。

在 ResNet50 特征提取网络的每一个瓶颈层的最后一个 BN 层之后,加入 SGE 模块,使得特征提取网络可以根据特征的空间分布来增强有用的特征表达,同时抑制可能的噪声,提高了特征提取的鲁棒性和代表性。这一操作与 SEnet^[11] 的网络架构相似。不同的是,SEnet 通过学习的方式来动态获得每一个特征通道的重要性,根据重要性来动态分配权重。而 SGEnet 是根据每个组中子特征的空间位置来生成注意力向量,根据注意力向量来再学习过程中动态的分配权重,提高有用特征的权重,同时减少可能的噪声。

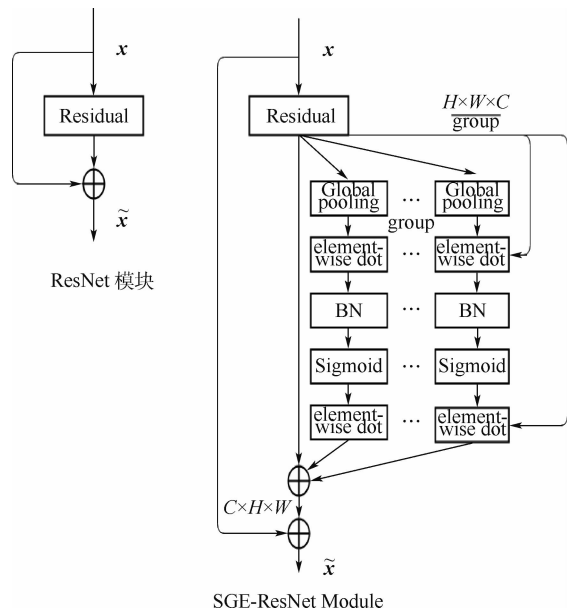


图 3 SGE-ResNet 模块结构

Fig. 3 SGE-ResNet module structure

2.2 IBN 归一化层

IN 层^[12]的作用是将不同风格的图片中的风格信息过滤掉,而只保留内容信息,具有很好的外观特征不变性。因此,为了提高特征提取网络对色调、色差的鲁棒性,使得提取到的特征更具代表性和全局性,本文在 BN 层的基础上引入 IN 层,组成 IBN 模块,将 IBN 模块加入到 ResNet50^[13]后的具体结构如图 4 所示。

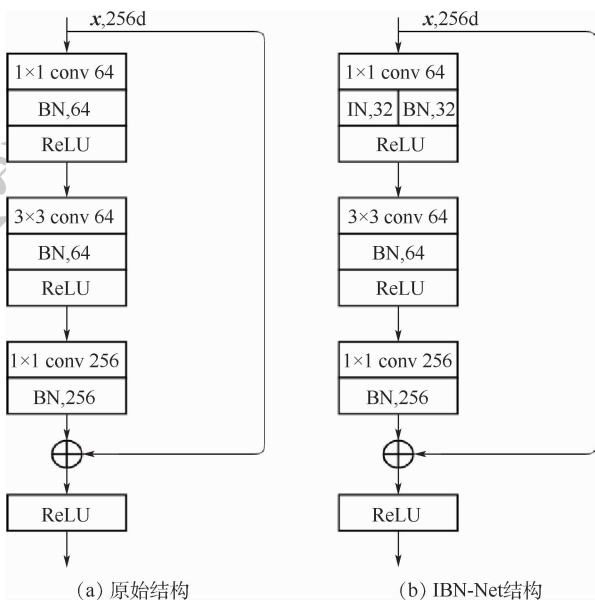


图 4 IBN 模块结构

Fig. 4 IBN module structure

为了防止网络在高层的语义判别性被破坏,IN 层只加在网络低层中,同时为了保留低层中的语义信息,网络低层中也保留 BN 层。在第一个归一化层,IN 层和 BN 层各占一半的通道数,一方面保证了不增加网络参数和计算量,另一方面也是为了在 IN 层过滤掉反映外观变化的信息的同时用 BN 层保留语义信息。

3 仿真实验与结果分析

3.1 数据集

本文选用 Market1501、DukeMTMC、CUHK03 3 个行人再识别通用数据集作为训练集以及测试集。Market1501 数据集包括由 6 个摄像头拍摄到的 1501 个行人,训练集有 751 人,包含 12 936 张图像,测试集有 750 人,包含 19 732 张图像;CUHK03 数据集包含 1 467 个行人,包含 767 个行人的训练集和包含 700 个行人的测试集;DukeMTMC 数据集包含了 1 404 个行人,其中来自 702 个人的 16 522 张图片作为训练集,另外的 702 个人的 2 228 个查询图像作为测试集,还有 17 661 张图像作为搜索库 (Gallery)。

3.2 实验设置

实验中所有算法都是在 Pytorch 框架下通过 Python 语言实现的。训练和测试步骤在 NVIDIA GTX1080Ti GPU 上运行。训练总轮数 epoch 设置为 300,训练学习率初始设置为 0.0002,训练每个 batchsize 设置为 32,迭代到 150 个 epoch 时学习率除以 10。总的权重衰减参数设置为 5×10^{-4} 。困难三元组损失中参数 margin 设置为 0.3。空间注意力机制 SGE 模块中的分组数设置为 64。

3.3 仿真实验

实验 1 将空间注意力机制 SGE 模块引入 ResNet50 网络,分别在 Market1501、DukeMTMC、CUHK03 三个行人再识别通用数据集评价改进模型性能。性能评价均采用全局距离加局部距离的结果(Global + DMLI),以及进行再排序(Re-ranking,RK)后的结果。

Rank1 是指搜索结果中置信度最高的图片是正确结果的概率;Rank1 (RK)是将结果再排序后,结果中置信度最高的图片是正确结果的概率;CMC(Cumulative Match Characteristic)^[14]曲线是一种 top-k 的击中概率,主要用来评估闭集中 rank 的正确率,这里取 Rank1 击中的概率;mAP(mean Average Precision)^[15]是模型的平均精度均值,衡量的是模型在所有类别上的好坏;mAP(RK)是将结果进行再排序操作后,模型的平均精度均值;再排序是获得图像特征后,选择 L2 欧氏距离来计算排序或检索任务的相似度得分。

由表 1 可以看出,在加入 SGE 模块后,在 Market1501^[16]数据集上性能相差不多,但是在进行 Re-ranking 操作后在 Rank1 击中率以及 mAP 上提升了 0.4%、0.3%;在 DukeMTMC^[17]数据集上,Rank1 击中率提升了 0.8%,mAP 提升了 0.2%,进行 Re-ranking 操作后,Rank1 击中概率

提升 0.5%,mAP 提升了 1.5%;在 CUHK03^[18]数据集上,Rank1 击中率提升了 1%,mAP 提升了 0.2%,在进行 Re-ranking 操作后,Rank1 提升了 3.5%,mAP 提升了 2.8%。可以看出,在加入空间注意力机制后,相较于原始模型效果有明显提升。

实验 2 将 IN 层加入到归一化层^[19]中,辅助 BN 层,组成 IBN 层共同加入到 ResNet50 网络中,分别在 Market1501、DukeMTMC、CUHK03 三个行人再识别通用数据集评价改进模型性能。性能评价均采用全局距离加局部距离的结果(Global + DMLI),以及进行 Re-ranking 操作后的结果。

由表 2 可以看出,在引入 IBN 归一化层后,在 Market1501 数据集上性能相差不多,Rank1 提升了 0.2%,mAP 提升了 1.9%。在进行 Re-ranking 操作后在 Rank1 击中率提升了 0.5%,mAP 提升了 1.2%;在 DukeMTMC 数据集上,Rank1 击中率提升了 2.9%,mAP 提升了 2.4%,进行 Re-ranking 操作后,Rank1 击中概率提升 1.9%,mAP 提升了 2.7%;在 CUHK03 数据集上,Rank1 击中率提升了 4.2%,mAP 提升了 3%,在进行 Re-ranking 操作后,Rank1 提升了 5.5%,mAP 提升了 4.5%。可以看出,在引入 IBN 归一化层后,相较于原始模型效果有明显提升。

实验 3 将 SGE 模块和 IBN 归一化层加入到 ResNet50 网络中,在 Market1501、DukeMTMC、CUHK03 三个行人再识别通用数据集评价改进模型性能。性能评价均采用全局距离加局部距离的结果(Global + DMLI),以及进行 Re-ranking 操作后的结果。

由表 3 可以看出,将空间注意力机制 SGE 模块以及 IBN 归一化层加入到 ResNet50 网络后,在 3 个数据集上效果均有显著提升,在 Market1501 数

表 1 基于 SGE 模块改进的模型性能在 Market1501、DukeMTMC、CUHK03 数据集上的评价

Table 1 Evaluation of improved model performance based on SGE module on Market1501, DukeMTMC and CUHK03 datasets							%
网络结构	Market1501		DukeMTMC		CUHK03		
	Rank1/Rank1 (RK)	mAP/mAP (RK)	Rank1/Rank1 (RK)	mAP/mAP (RK)	Rank1/Rank1 (RK)	mAP/mAP (RK)	
ResNet50	91.0/92.0	77.6/88.5	80.7/85.2	68.0/81.2	60.9/67.6	59.7/70.7	
SGE-ResNet50	90.5/92.4	76.8/88.8	81.5/85.7	68.2/82.7	61.9/71.1	59.9/73.5	

表 2 基于 IBN 层改进的模型性能在 Market1501、DukeMTMC、CUHK03 数据集上的评价

Table 2 Evaluation of improved model performance based on IBN layer on Market1501, DukeMTMC and CUHK03 datasets							%
网络结构	Market1501		DukeMTMC		CUHK03		
	Rank1/Rank1 (RK)	mAP/mAP (RK)	Rank1/Rank1 (RK)	mAP/mAP (RK)	Rank1/Rank1 (RK)	mAP/mAP (RK)	
ResNet50	91.0/92.0	77.6/88.5	80.7/85.2	68.0/81.2	60.9/67.6	59.7/70.7	
IBN-ResNet50	91.2/92.5	79.5/89.7	83.6/87.1	70.4/83.9	65.1/73.1	62.7/75.2	

数据集上,Rank1 击中率提升了 0.8%,mAP 提升了 2.7%,Re-ranking 操作后 Rank1 击中率提升了 0.9%,mAP 提升了 2%;在 DukeMTMC 数据集上,Rank1 击中率提升了 2.9%,mAP 提升了 3%,Re-ranking 操作后 Rank1 击中率提升了 2.2%,mAP 提升了 2.9%;在 CUHK03 数据集上,Rank1 击中率提升了 4.7%,mAP 提升了 3.2%,Re-ranking 操作后 Rank1 击中率提升了 5.4%,mAP 提升了 5.1%。

图 5 给出了原模型以及加入 SGE 模块和 IBN 归一化层后的性能曲线。可以看出,改进后的算法在性能上相较于原算法有显著的提升,证明了改进后模型的有效性。

图 6 是在不同网络下,针对待查询图库 query

中的同一幅较为模糊的行人图片,从查询库中输出与其相似度最高的 10 幅行人图片。绿色标签表示查询结果与待查询图片是同一标签,即是同一个人;红色标签表示查询结果与待查询图片不是同一标签,即不是同一个人。可以看出,在引入 SGE 空间注意力机制以及 IBN 归一化层后,查询结果有了很大的改善,表明了改进模型的有效性。

表 4 为改进后的方法与行人再识别最新方法的比较。可以看出,与改进前的方法相比较,改进后方法在识别精度上,在 Market1501、DukeMTMC 和 CUHK03 3 个数据集上均有提升。与表中其他行人再识别方法相比,改进后的方法在识别准确率上也有较为明显的提升,由此可以得出改进后方法的有效性。

表 3 基于 IBN 层以及 SGE 模块改进的模型性能在 Market1501、DukeMTMC、CUHK03 数据集上的评价

Table 3 Evaluation of improved model performance based on IBN layer and SGE module on Market1501, DukeMTMC and CUHK03 datasets

网络结构	Market1501		DukeMTMC		CUHK03	
	Rank1/Rank1 (RK)	mAP/mAP (RK)	Rank1/Rank1 (RK)	mAP/mAP (RK)	Rank1/Rank1 (RK)	mAP/mAP (RK)
ResNet50	91.0/92.0	77.6/88.5	80.7/85.2	68.0/81.2	60.9/67.6	59.7/70.7
SGE-IBN-ResNet50	91.8/92.9	80.3/90.5	83.6/87.4	71.0/84.1	65.6/73.0	62.9/75.8

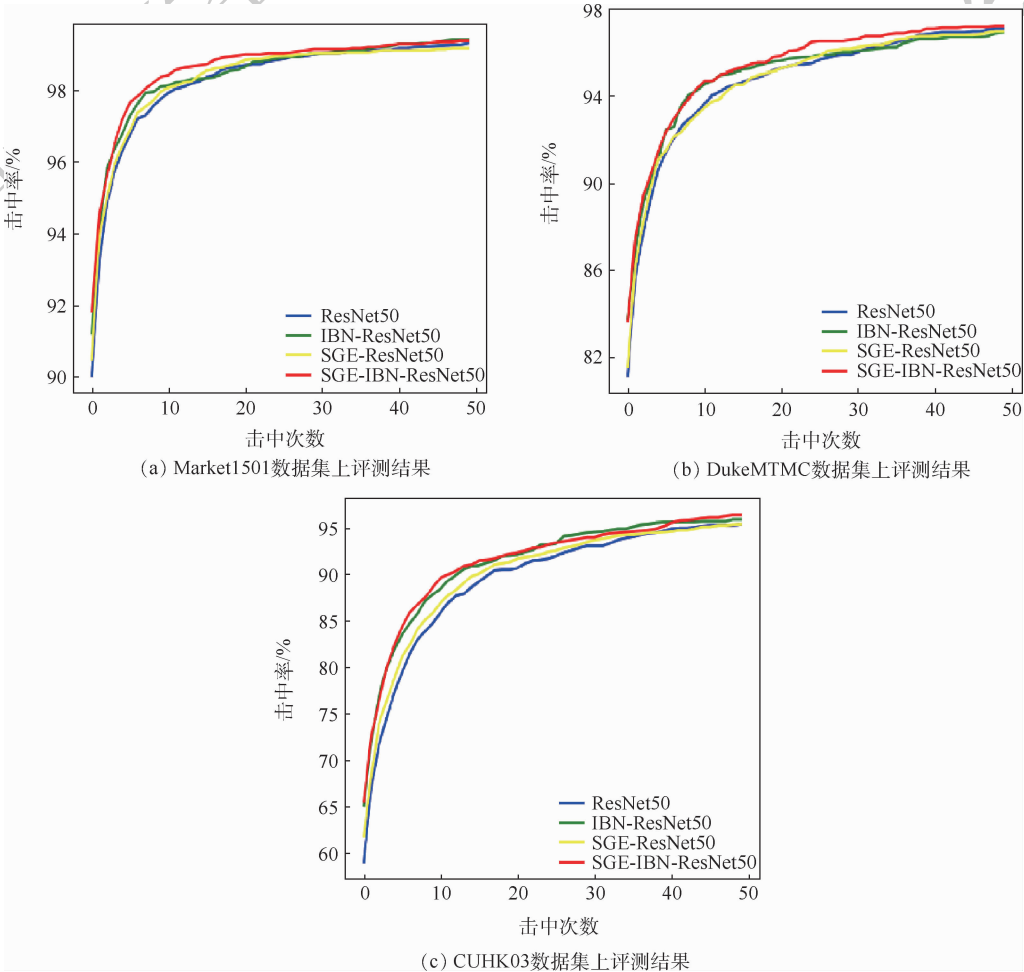


图 5 4 种网络在 Market1501、DukeMTMC、CUHK03 数据集上的 CMC 曲线

Fig. 5 CMC curves of four networks on Market1501, DukeMTMC and CUHK03 datasets

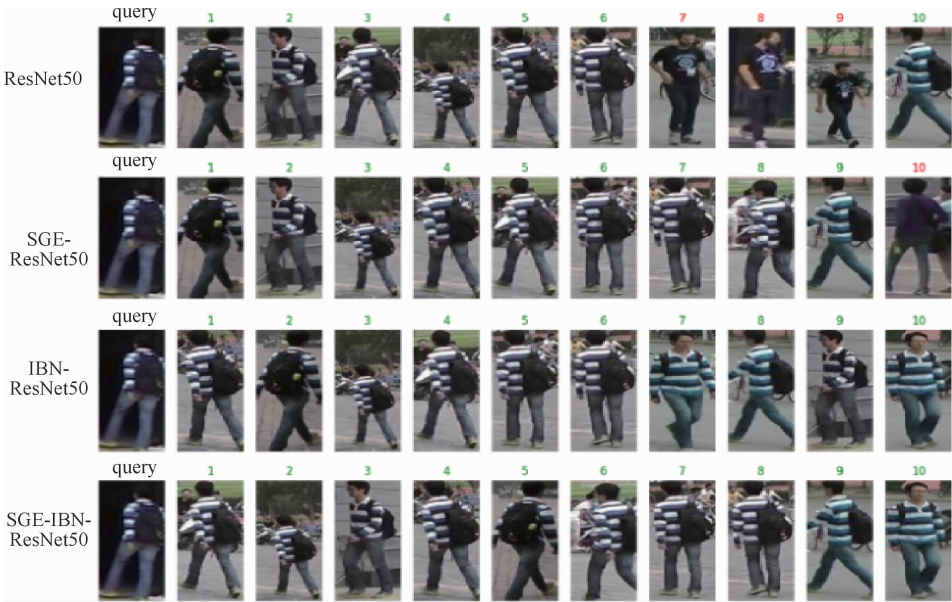


图 6 不同网络下相似度最高的 10 张行人图片
Fig. 6 Ten most similar pedestrian images on different networks

表 4 改进后的方法与最新行人再识别方法比较

Table 4 Comparison between improved method and latest pedestrian re-identification methods %

方法	Market1501		DukeMTMC		CUHK03	
	Rank1	mAP	Rank1	mAP	Rank1	mAP
SVD ^[20]	82.3	62.1	76.7	56.8	41.5	37.3
PCE&ECN ^[21]	87.0	69.0	79.8	62.0	30.2	27.3
MLFN ^[22]	90.0	74.3	81.0	62.8	52.8	47.8
HA-CNN ^[23]	91.2	75.7	80.5	63.8	41.7	38.3
AlignedReID + +	91.0	77.6	80.7	68.0	60.9	59.7
AlignedReID + + (RK)	92.0	88.5	85.2	81.2	67.6	70.7
AlignedReID + + (SGE + IBN)	91.8	80.3	83.6	71.0	65.6	62.9
AlignedReID + + (SGE + IBN) (RK)	92.9	90.5	87.4	84.1	73.0	75.8

4 结 论

1) 本文提出了一种基于注意力机制的行人特征提取方法。在 AlignedReID + + 基础上,对特征提取部分进行了改进。通过引入空间注意力模块,根据特征的空间位置,动态地增强有用特征,抑制可能的噪声。同时通过引入新的归一化层 IBN 层,进一步提高了提取特征的鲁棒性和全局性。

2) 相较于改进前 AlignedReID + + 中的 Res-Net50 特征提取网络,本文方法可以提取更为丰富的行人特征,提取到的特征更具准确性和鲁棒性。在 Market1501、DukeMTMC 以及 CUHK03 三个数据集上进行的算法测试实验结果表明,改进后的 AlignedReID + + 算法的准确性更高,具有更强的鲁棒性。

参考文献 (References)

[1] LIAO S C, HU Y, ZHU X Y, et al. Person re-identification by local maximal occurrence representation and metric learning [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 2197-2206.

[2] DE MAESSCHALCK R, JOUAN-RIMBAUD D, MASSART D L. The Mahalanobis distance [J]. Chemometrics and Intelligent Laboratory Systems, 2000, 50 (1) : 1-18.

[3] YI D, LEI Z, LIAO S C, et al. Deep metric learning for person re-identification [C] // 2014 22nd International Conference on Pattern Recognition. Piscataway: IEEE Press, 2014: 34-39.

[4] LI W, ZHAO R, XIAO T, et al. DeepReID: Deep filter pairing neural network for person re-identification [C] // 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 152-159.

[5] VARIOR R R, SHUAI B, LU J, et al. A Siamese long short-term memory architecture for human re-identification [C] // European Conference on Computer Vision. Berlin: Springer, 2016: 135-153.

- [6] ZHANG X, LUO H, FAN X, et al. AlignedReID: Surpassing human-level performance in person re-identification [EB/OL]. (2018-01-31) [2020-03-02]. <https://arxiv.org/abs/1711.08184>.
- [7] DENG W, ZHENG L, YE Q, et al. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification [C] // 2018 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 994-1003.
- [8] LUO H, JIANG W, ZHANG X, et al. AlignedReID + + : Dynamically matching local information for person re-identification [J]. Pattern Recognition, 2019, 94: 53-61.
- [9] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification [EB/OL]. (2017-11-21) [2020-03-02]. <https://arxiv.org/abs/1703.07737>.
- [10] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [EB/OL]. (2015-03-02) [2020-03-02]. <https://arxiv.org/abs/1502.03167>.
- [11] LI X, HU X, YANG J. Spatial group-wise enhance: Improving semantic feature learning in convolutional networks [EB/OL]. (2019-05-25) [2020-03-02]. <https://arxiv.org/abs/1905.09646>.
- [12] PAN X, LUO P, SHI J, et al. Two at once: Enhancing learning and generalization capacities via IBN-Net [EB/OL]. (2018-07-27) [2020-03-02]. <https://arxiv.org/abs/1807.09441>.
- [13] XIAO Q, LUO H, ZHANG C. Margin sample mining loss: A deep learning based method for person re-identification [EB/OL]. (2017-10-07) [2020-03-02]. <https://arxiv.org/abs/1710.00478>.
- [14] ZHENG L, YANG Y, HAUPTMANN A G. Person re-identification: Past, present and future [EB/OL]. (2016-10-10) [2020-03-02]. <https://arxiv.org/abs/1610.02984>.
- [15] LIU H, FENG J S, QI M B, et al. End-to-end comparative attention networks for person re-identification [J]. IEEE Transactions on Image Processing, 2017, 26(7): 3492-3506.
- [16] CHENG D, GONG Y H, ZHOU S P, et al. Person re-identification by multi-channel parts-based CNN with improved triplet loss function [C] // 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1335-1344.
- [17] 杜鹏, 宋永红, 张鑫瑶. 基于自注意力模态融合网络的跨模态行人再识别方法研究 [J/OL]. 自动化学报: 1-12 (2019-10-16) [2020-01-06]. <https://kns.cnki.net/kcms/detail/detail.aspx?doi=10.16383/j.aas.c190340>.
- [18] 张丽红, 孙志琳. 基于多层深度特征融合的行人再识别研究 [J]. 测试技术学报, 2018, 32(4): 48-52.
- ZHANG L H, SUN Z L. Person re-identification based on multi-layer deep feature fusion [J]. Journal of Test and Measurement Technology, 2018, 32(4): 48-52 (in Chinese).
- [19] 李鹏, 王德勇, 师文喜, 等. 大数据环境下基于深度学习的行人再识别 [J]. 北京邮电大学学报, 2019, 42(6): 29-34.
- LI P, WANG D Y, SHI W X, et al. Research on person re-identification based on deep learning under big data environment [J]. Journal of Beijing University of Posts and Telecommunications, 2019, 42(6): 29-34 (in Chinese).
- [20] SUN Y, ZHENG L, DENG W, et al. SVDNet for pedestrian retrieval [C] // 2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 3820-3828.
- [21] SARFRAZ M S, SCHUMANN A, EBERLE A, et al. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking [EB/OL]. (2018-04-05) [2020-03-02]. <https://arxiv.org/abs/1711.10378>.
- [22] AN L, QIN Z, CHEN X J, et al. Multi-level common space learning for person re-identification [J]. IEEE Transactions on Circuits & Systems for Video Technology, 2018, 28(8): 1777-1787.
- [23] LI W, ZHU X, GONG S. Harmonious attention network for person re-identification [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2285-2294.

作者简介:

张子昊 男, 硕士研究生。主要研究方向: 模式识别、人工智能。

周千里 男, 博士研究生。主要研究方向: 模式识别、人工智能。

王蓉 女, 博士, 教授, 博士生导师。主要研究方向: 模式识别、人工智能。

Pedestrian re-identification method based on spatial attention mechanism

ZHANG Zihao^{1,2}, ZHOU Qianli^{1,2,3}, WANG Rong^{1,2,*}

(1. School of Police Information Engineering and Network Security, People's Public Security University of China, Beijing 100038, China;

2. Key Laboratory of Security Technology & Risk Assessment, Beijing 100038, China;

3. Beijing Public Security Bureau, Beijing 100740, China)

Abstract: Pedestrian re-identification has always been an important part of image retrieval. However, due to different pedestrian poses and complex backgrounds, the extracted pedestrian features are not robust and representative, which in turn affects the accuracy of pedestrian re-recognition. In this paper, based on AlignedReID++ algorithm, we propose a pedestrian re-identification method based on spatial attention mechanism. First, in the feature extraction part, a spatial attention mechanism is introduced to enhance feature expression while suppressing possible noise. Second, the Instance-Normalization (IN) layer is introduced in the convolution layer to assist the Batch-Normalization (BN) layer to normalize the features and to solve the problem of single BN layer insensitivity to feature tonal and illumination changes, which enhances the robustness of feature extraction to tonal and illumination changes. Finally, to validate the proposed method, extensive experiment has been carried out on the Market1501, DukeMTMC, and CUHK03 pedestrian re-identification datasets. The experimental results show that the recognition accuracy of the improved model on the three datasets has been improved by 2%, 2.9%, and 5.1%, respectively, compared with model before modification, which indicates that the proposed method achieves higher accuracy and more robustness.

Keywords: deep learning; spatial attention mechanism; pedestrian characteristics; feature enhancement; convolutional neural network

Received: 2020-03-03; **Accepted:** 2020-03-27; **Published online:** 2020-04-02 17:07

URL: kns.cnki.net/kcms/detail/11.2625.V.20200402.1658.001.html

Foundation items: National Key R & D Program of China (A19808); the Fundamental Research Funds for the Central Universities (2019JKF111)

* **Corresponding author.** E-mail: dbdxwangrong@163.com

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0049

基于 MobileFaceNet 网络改进的人脸识别方法



张子昊^{1,2}, 王蓉^{1,2,*}

(1. 中国人民公安大学 警务信息工程与网络安全学院, 北京 100038;

2. 安全防范技术与风险评估公安部重点实验室, 北京 100038)

摘 要: 为了解决训练过程中卷积模型参数较多、收敛速度较慢的问题,提出了一种基于 MobileFaceNet 网络改进的人脸识别方法。首先,使用 MobileFaceNet 网络提取人脸特征,在提取特征的过程中,通过引入可分离卷积减少模型中卷积层参数的数量;其次,通过在 MobileFaceNet 网络中引入风格注意力机制来增强特征的表达,同时使用 AdaCos 人脸损失函数来训练模型,利用 AdaCos 损失函数中的自适应缩放系数,来动态地调整超参数,避免了人为设置超参数对模型的影响;最后,分别在 LFW、AgeDB 和 CFP-FF 测试数据集上对训练模型进行评估。实验结果显示:改进后的模型在 LFW、AgeDB 和 CFP-FF 测试数据集上的识别精度分别提升了 0.25%、0.16% 和 0.3%,表明改进后的模型相较于改进前的模型在精度和鲁棒性上有所提高。

关 键 词: 人脸识别;深度学习;MobileFaceNet;AdaCos;卷积神经网络

中图分类号: O235; TP183

文献标志码: A

文章编号: 1001-5965(2020)09-1756-07

人脸特征具有易获取、易捕捉、易处理以及非接触式等特性,已经受到学者们的广泛关注,并在公共安全等领域应用日趋广泛。但是,在实际的应用场景下,人脸识别精度依然受到不同的姿态、光照、遮挡等因素的影响。因此,如何提取更具鲁棒性的特征来有效地辨识人脸便成为解决问题的关键。传统的人脸识别方法主要有主成分分析法(PCA)^[1]以及局部二值模式(LBP)^[2]。主成分分析法就是将高维人脸信息通过正交变化投影到低维子空间中,形成特征脸,然后通过分类器对低维的特征脸进行分类;局部二值法是将检测窗口划分为许多小区域,对每一个像素将其与周围相邻像素比较,并重新赋值,通过计算每个区域的直方图,并使用机器学习的方法对其进行分类。近年来,随着深度学习应用的日趋广泛,使用深度学

习的方法来进行人脸识别成为计算机视觉领域的热门方向。

与传统方法相比,深度学习的方法通过海量的数据来训练模型,使模型提取到的人脸特征更具有泛化性。最具代表性的是香港中文大学汤晓鸥团队提出的 DeepID^[3]系列。该团队中的 Sun 等提出 DeepID1^[4]通过使用单一的卷积网络来进行特征提取并进行分类识别,之后,又在文献[4]的基础上提出了 DeepID2^[5]模型,通过将人脸认证信号和人脸验证信号引入网络,以达到增大类间距缩小类内距的要求。随后,为了进一步提高模型提取特征的代表性和鲁棒性,Google 提出了 FaceNet^[6]模型,将模型映射到欧几里德空间,并使用三元组损失 Triplet loss^[7]来增大类间距,从而提高模型性能;为了解决不同特征向量大小对

收稿日期: 2020-02-25; 录用日期: 2020-04-18; 网络出版时间: 2020-04-21 11:39

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200421.0911.001.html

基金项目: 公安部技术研究计划项目(2017JSYJB01); 中央高校基本科研业务费专项资金(2019JKF111)

* 通信作者。E-mail: dbdxwangrong@163.com

引用格式: 张子昊, 王蓉. 基于 MobileFaceNet 网络改进的人脸识别方法[J]. 北京航空航天大学学报, 2020, 46(9): 1756-1762.
ZHANG Z H, WANG R. Improved face recognition method based on MobileFaceNet network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1756-1762 (in Chinese).

分类效果的影响,Deng 等提出 Arcface^[8] 损失,将特征向量归一化到超球面上,使模型在角度空间对分类的边界进行最大化;为了解决训练过程中超参数的选择问题,Zhang 等提出 AdaCos^[9] 人脸损失函数,引入自适应动态缩放参数来解决超参数的选择问题。但是以上介绍的方法中主干网络大多比较复杂,模型参数较多,增加了计算量,影响计算效率以及模型的收敛速度。

本文在 MobileFaceNet^[10] 网络的基础上加入风格注意力机制,将改进后的模型作为特征提取网络,使提取到的特征更具代表性和鲁棒性。同时引入 AdaCos 人脸损失函数来对模型进行训练,通过动态自适应缩放系数的调节优化训练,减少了人为调节超参数对训练的影响,提高了训练效率。

1 基于 MobileFaceNet 网络的人脸识别方法

1.1 MobileNetV1 和 MobileNetV2 网络

MobileNetV1^[11] 是 Google 发布的网络架构,提出深度可分离卷积的概念。其核心就是把卷积拆分为 Depthwise + Pointwise 两部分。其中 Depthwise 卷积是指不跨通道的卷积,即 Feature Map 的每个通道有一个独立的卷积核,并且这个卷积核作用且仅作用在这个通道之上,输出 Feature Map 的通道数等于输入 Feature Map 的通道数,因此它并没有升维或者降维的功能。Pointwise 卷积用于特征合并以及升维或者降维,一般使用 1×1 的卷积来实现功能。由此,可以将一个普通的卷积拆分成如图 1 所示的结构,图中: N 、 M 和 D_k 分别为卷积的个数、卷积通道数和卷积的宽、高。

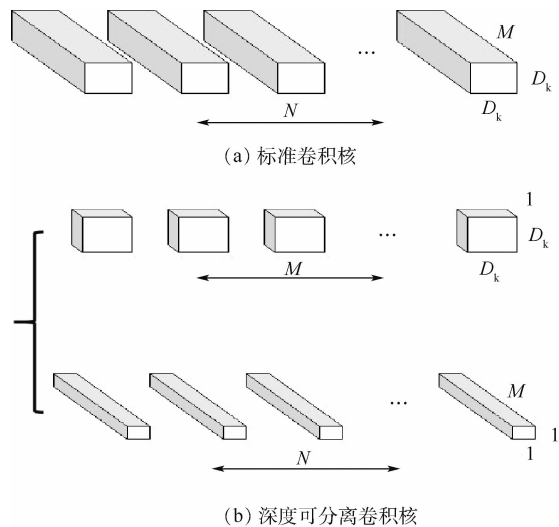


图 1 可分离卷积示意图

标准卷积的参数数量是 $D_k \times D_k \times M \times N$;卷积核的尺寸是 $D_k \times D_k \times M$,一共有 N 个,每一个都要进行 $D_w \times D_h$ 次运算,所以标准卷积的计算量 P_1 为

$$P_1 = D_k \times D_k \times M \times N \times D_w \times D_h \tag{1}$$

式中: D_w 和 D_h 分别为标准卷积的宽和高。

深度可分离卷积的参数量是: $D_k \times D_k \times M + 1 \times 1 \times M \times N$;深度可分离卷积的计算量是由深度卷积和逐点卷积两部分组成:深度卷积的卷积核尺寸 $D_k \times D_k \times M$,一共要做 $D_w \times D_h$ 次乘加运算;逐点卷积的卷积核尺寸为 $1 \times 1 \times M$,有 N 个,一共要做 $D_w \times D_h$ 次乘加运算,所以深度可分离卷积的计算量 P_2 为

$$P_2 = D_k \times D_k \times M \times D_w \times D_h + M \times N \times D_w \times D_h \tag{2}$$

参数数量和乘加操作的运算量下降倍数 P_3 为

$$P_3 = \frac{P_2}{P_1} = \frac{D_k \times D_k \times M \times D_w \times D_h + M \times N \times D_w \times D_h}{D_k \times D_k \times M \times N \times D_w \times D_h} = \frac{1}{N} + \frac{1}{D_k^2} \tag{3}$$

从式(3)可以看出,深度可分离卷积在参数数量以及计算量上与标准卷积相比,大幅度的降低,提升了模型的运算效率。

深度卷积本身没有改变通道的能力,即输入通道等于输出通道若是输入通道数较少,则深度卷积只能在低维度工作,最终卷积效并不会很好。所以 MobileNetV2^[12] 在 MobileNetV1 的深度卷积层之前,使用 1×1 的卷积进行升维操作,使网络在一个更高维的空间进行特征提取,这就使得提取到的特征更具全局性。同时,为了解决低维度 Relu 运算产生的信息丢失问题,将最后一个 Relu 替换成 Linear 线性激活函数。MobileNetV2 网络结构如表 1 所示,表中: z 为通道扩张倍数; c 为输出通道数; n 为重复次数; v 为步长; f 为输入通道数。

表 1 MobileNetV2 网络结构

Table 1 MobileNetV2 network structure					
输入	卷积操作	z	c	n	v
224 × 224 × 3	Conv2d		32	1	2
112 × 112 × 32	bottleneck	1	16	1	1
112 × 112 × 16	bottleneck	6	24	2	2
56 × 56 × 24	bottleneck	6	32	3	2
28 × 28 × 32	bottleneck	6	64	4	2
28 × 28 × 64	bottleneck	6	96	3	1
14 × 14 × 96	bottleneck	6	160	3	2
7 × 7 × 160	bottleneck	6	320	1	1
7 × 7 × 320	Conv2d 1 × 1		1280	1	1
7 × 7 × 1280	Avgpool 7 × 7			1	
1 × 1 × f	Conv2d 1 × 1		f		

Fig. 1 Schematic diagram of detachable convolution

1.2 MobileFaceNet 人脸特征提取网络

MobileFaceNet 是 MobileNetV2 的改进版本。由于 MobileNetV2 中使用的是平均池化层,但是针对同一张图片,不同像素点的权重是不同的,全局的平均池化将权重平均,网络的表现能力自然会下降。所以在 MobileFaceNet 中,使用一个 $7 \times 7 \times 512$ 的可分离卷积代替原本的全局平均池化层。同时使用 Prelu 代替 Relu 激活函数,并引入归一化层来加快模型收敛,防止模型过拟合。MobileFaceNet 的网络结构如表 2 所示。

表 2 MobileFaceNet 网络结构
Table 2 MobileFaceNet network structure

输入	卷积操作	<i>z</i>	<i>c</i>	<i>n</i>	<i>v</i>
112 × 112 × 3	Conv 3 × 3		64	1	2
56 × 56 × 64	Depthwise Conv 3 × 3		64	1	1
56 × 56 × 64	bottleneck	2	64	5	2
28 × 28 × 64	bottleneck	4	128	1	2
14 × 14 × 128	bottleneck	2	128	6	1
14 × 14 × 128	bottleneck	4	128	1	2
7 × 7 × 128	bottleneck	2	128	2	1
7 × 7 × 128	Conv 1 × 1		512	1	1
7 × 7 × 512	Linear GConv 7 × 7		512	1	1
1 × 1 × 512	Linear Conv 1 × 1		128	1	1

网络结构不仅保留了 MobileNetV2 的升降维层以及线性激活函数,同时还在全连接层之前引入 7×7 的可分离卷积来替代原本的平均池化层,使网络提取到的特征更具泛化性和全局性。

1.3 AdaCos 人脸损失函数

Softmax 目前被广泛应用于图片分类中,主要通过输出样本属于各个类的概率大小来区分不同类别之间的特征。但是 Softmax 并没有对类内和类间距离进行约束,这就大大影响了人脸识别的精度。最新的人脸损失函数 Arcface 以及 AdaCos 就是基于 Softmax 改进的。

1.3.1 Arcface 损失函数

Softmax 函数为

$$L_s = - \frac{1}{m} \sum_{i=1}^m \ln \left(\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^n e^{W_j^T x_i + b_j}} \right) \tag{4}$$

式中: m 为每一个训练批次的大小; i 为图像编号; j 为当前样本的真实标签; $W_{y_i}^T x_i + b_{y_i}$ 表示全连接层输出, $W_{y_i}^T x_i + b_{y_i}$ 和 y_i 分别为第 i 个图像类别的权重、图像的特征、图像类别的偏差和图像的真实类别的标签。通过提高 $W_{y_i}^T x_i + b_{y_i}$ 所占有的比重,提高对该类别的分类效果。但是 Softmax 只考虑到样本的分类正确性,而对于人脸识别这样的多分类问题,缺乏类内和类间距离约束。又因

为 $W^T x = \|W\| \cdot \|x\| \cdot \cos \theta$, θ 为特征向量 x 和权重 W 之间的夹角,即特征向量相乘包含有角度信息。为了让卷积网络提取到的特征可以学习到更可分的角度特性,将特征的权重值固定成一个定值 s ,同时为了简化模型参数,将偏置值设为 0。这就将提取到的特征转化到了角度空间,使得决策边界只与角度有关。最后,通过引入一个新参数 $t(t > 0)$ 来控制角度余弦量值的大小,使得网络可以学习到更有区分度的特征。最终 Arcface 损失函数为

$$L_{\text{Arcface}} = - \frac{1}{m} \sum_{i=1}^m \ln \left(\frac{e^{s(\cos(\theta_{y_i+t}))}}{e^{s(\cos(\theta_{y_i+t}))} + \sum_{j=1, j \neq y_i}^n e^{s \cos \theta_j}} \right) \tag{5}$$

1.3.2 AdaCos 损失函数

AdaCos 损失函数在 Arcface 的基础上引入一个动态自适应缩放系数 s_d ,使得在训练过程中可以动态地调整超参数,同时去掉了控制角度余弦量值的参数 t 。最终 AdaCos 损失函数为

$$L_{\text{AdaCos}} = - \frac{1}{m} \sum_{i=1}^m \ln \left(\frac{e^{\tilde{s}_d^{(t)} \cos \theta_{i,j}}}{C_1 \sum_{k=1} e^{\tilde{s}_d^{(t)} \cos \theta_{i,j}}} \right) \tag{6}$$

式中: C_1 为类别的数量; $\theta_{i,j}$ 为特征向量 x_i 和对应权重之间的夹角; $\tilde{s}_d^{(t)}$ 为一个动态缩放系数,会随着迭代次数 t_d 的不同,取值有 2 种情况,如下:

$$\tilde{s}_d^{(t)} = \begin{cases} \sqrt{2 \ln(C_1 - 1)} & t_d = 0 \\ \frac{\ln B_{\text{avg}}^{(t_d)}}{\cos\left(\min\left(\frac{\pi}{4}, \theta_{\text{med}}^{(t_d)}\right)\right)} & t_d \geq 1 \end{cases} \tag{7}$$

其中: $\theta_{\text{med}}^{(t)}$ 为引入的一个调试指示变量,是所有对应类别的中心角度。当中心角度大的时候,表示网络参数距离最优还很远,需要一个弱监督信号;当中心角度较小时,表示网络参数距离最优已经很近了,需要一个较强的监督信号。 $B_{\text{avg}}^{(t)}$ 为所有非对应 label 的 Logits 之和的平均值,其计算公式为

$$B_{\text{avg}}^{(t)} = \frac{1}{N_1} \sum_{i \in N(t)} \sum_{k_1 \neq y_i} e^{\tilde{s}_d^{(t-1)} \cos \theta_{i,k_1}} \tag{8}$$

式中: $N(t)$ 为一个批次中人脸的类别数量; N_1 为类别总数; k_1 为第 i 个图像的真实类别。

由 AdaCos 的公式可以看出,动态自适应参数在每次迭代的时候对分类概率的影响是不同的,由此就可以根据模型的收敛程度动态的产生合理的缩放系数,加快模型的收敛速度。

2 改进方法

本节在 MobileFaceNet 人脸特征提取网络的

基础上,将风格注意力机制重新校准模块 (Style-based Recalibration Module, SRM)^[13] 加入到 MobileFaceNet 网络中,进一步提高了提取特征的鲁

棒性和全局性。人脸识别方法流程如图 2 所示,多任务卷积神经网络 (MTCNN) 是一种人脸检测方法。

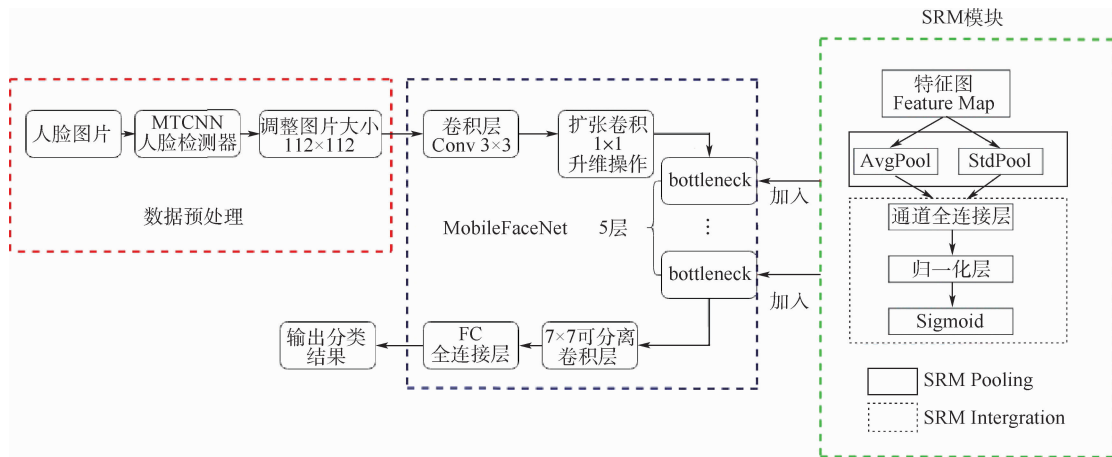


图 2 人脸识别方法流程

Fig. 2 Face recognition method flowchart

方法流程主要分为 3 个部分,即数据预处理、基于 SRM 模块改进的 MobileFaceNet 人脸特征提取网络、输出人脸分类结果。之后使用自适应人脸损失函数 AdaCos 来监督训练过程,减少了设置超参数对训练效果的影响,同时进一步扩大类间间距,缩小类内间距。方法的整体实现是基于 pytorch 框架。

2.1 SRM 模块

通常来说,不同的人脸图像会有不同的属性特征,卷积网络层通过提取图片特征来进行分类和识别。然而,针对不同的人脸图像会有不同的风格,这些风格属性会进一步影响卷积网络对特征的提取,进而影响之后的识别和分类精度。

SRM 通过特征重新校准的形式明确地将风格信息合并到卷积神经网络 (CNN) 表示中。根据不同人脸图像的风格属性,动态地估计各个风格特征的相对重要性,然后根据风格的重要性动态的调整特征权重,这使得网络可以专注于有意义的风格信息,而忽略不必要的风格信息。SRM 模块结构如图 3 所示, C 、 H 、 W 和 d 分别为通道的个数、特征图高度、特征图的宽度和风格特征的数量。

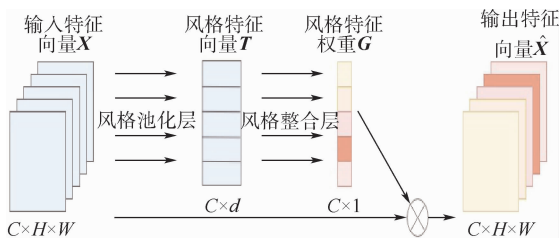


图 3 SRM 模块结构

Fig. 3 SRM module structure

SRM 模块由 2 个主要组件组成:风格池化层 (Style Pooling) 和风格整合层 (Style Integration)。风格池化层通过汇总不通空间维度的特征响应,从每个通道提取风格特征;风格整合层通过逐通道操作,针对图片的不同位置,利用风格特征生成相应的风格权重。同时根据不同的风格权重重新校准特征映射,以强调或隐藏它们的信息。

2.1.1 风格池化层

风格池化层包括 2 个部分,即平均池化层和标准差池化层。对于每一个输入为 $X \in \mathbf{R}^{N_d \times C \times H \times W}$ 的特征图, N_d 为小批次中样本的个数,首先通过平均池化层,具体操作如下:

$$\mu_{nc} = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W x_{nchw} \quad (9)$$

式中: μ_{nc} 为原特征图平均池化后的结果; x_{nchw} 为原特征图, n 、 c 、 h 和 w 分别为 N_d 、 C 、 H 和 W 的分量。其次,将平均池化层的结果与原输入特征图进行标准差的计算,具体过程如下:

$$\sigma_{nc} = \sqrt{\frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W (x_{nchw} - \mu_{nc})^2} \quad (10)$$

式中: σ_{nc} 为经过标准池化层后的结果。最后,经过风格池化层的结果如下:

$$t_{nc} = [\mu_{nc}, \sigma_{nc}] \quad (11)$$

其中: t_{nc} 为经过风格池化层后的特征向量。

2.1.2 风格整合层

风格整合层包括 3 个部分,即通道全连接层 (Channel-wise Fully Connected, CFC)、归一化层 (Batch Normalization, BN)^[14] 以及激活层。将从风格池化层输出的结果作为风格整合层的输入,

具体操作如式(12)~式(16)所示。

$$\mathbf{z}_{nc} = \mathbf{w}_c \cdot \mathbf{t}_{nc} \quad (12)$$

$$\boldsymbol{\mu}_c^{(z)} = \frac{1}{N_d} \sum_{n=1}^{N_d} \mathbf{z}_{nc} \quad (13)$$

$$\boldsymbol{\sigma}_c^{(z)} = \sqrt{\frac{1}{N_d} \sum_{n=1}^{N_d} (\mathbf{z}_{nc} - \boldsymbol{\mu}_c^{(z)})^2} \quad (14)$$

$$\hat{\mathbf{z}}_{nc} = \gamma_c \left(\frac{\mathbf{z}_{nc} - \boldsymbol{\mu}_c^{(z)}}{\boldsymbol{\sigma}_c^{(z)}} \right) + \beta_c \quad (15)$$

$$\mathbf{g}_{nc} = \frac{1}{1 + e^{-\hat{\mathbf{z}}_{nc}}} \quad (16)$$

式中: $\mathbf{z}_{nc}$ 为与可学习的参数 $\mathbf{w}_c$ 点乘后的特征向量; $\boldsymbol{\mu}_c^{(z)}$ 为取平均值后的特征向量; $\boldsymbol{\sigma}_c^{(z)}$ 为取标准差后的特征向量; $\hat{\mathbf{z}}_{nc}$ 为经过归一化层后的特征向量; $\mathbf{g}_{nc}$ 为经过激活层后的特征向量; $\mathbf{w}_c$ 为可学习的参数, $\mathbf{w}_c \in \mathbf{R}^{C \times 2}$; γ_c, β_c 为归一化层中的仿射变换参数。

2.2 基于风格注意力机制改进的 MobileFaceNet 人脸特征提取网络

MobileFaceNet 人脸特征提取网络中有 5 个 bottleneck 层,其结构如图 4 所示。

将 SRM 模块加入到 bottleneck 层中后的结构如图 5 所示。

将 SRM 结构放在了 bottleneck 层中的 Depth-wise 操作之后,使得特征提取网络可以根据特征图的风格信息来动态地增强有用的特征表达,抑制可能的噪声,提高了特征提取的鲁棒性和代表性。这一操作与 MobileNetV3^[15] 的网络架构相似,不同的是 MobileNetV3 的网络架构加入 SE 模块^[16]。

3 仿真实验与结果分析

3.1 实验设置

3.1.1 数据预处理

本文使用 CASIA-Webface 人脸数据集作为训练数据集,包含了 10 575 个人的 494 414 张图像。使用 MTCNN^[17] 人脸检测方法对 CASIA-Webface 中的图片进行再检测,并将检测到的人脸图片裁剪成 112 × 112 个像素大小。

测试数据集选用的是 LFW、CFP-FF、以及 AgeDB 3 个人脸数据集来进行模型的评估。LFW 数据集包含 5 749 人共 13 233 张人脸图像;CFP 数据集包含 500 个身份,每个身份有 10 个正脸,4 个侧脸,本实验使用 CFP 数据集中的 FP (Frontal-Profile) 人脸验证,即 CFP-FP (CFP with Frontal-Profile) 数据集;AgeDB 数据集包含 440 人共 12 240 张人脸图像。

3.1.2 实验环境及参数设置

实验中所有方法都是在 pytorch 框架下通过 python 语言实现的。训练和测试步骤在 NVIDIA GTX2080Ti GPU 上运行。训练总轮数 epoch 设置为 50,训练初始学习率设置为 0.1,迭代到 10、25、40 个 epoch 时,学习率每次除以 10。训练中的 batchsize 设置为 64,总的权重衰减参数设置为 5×10^{-4} ;使用随机梯度下降策略 SGD 优化模型,动量参数设置为 0.9。

3.2 在不同 backbone 上的结果

不同的特征提取网络分别为 Resnet-50、MobileFaceNet、MobileFaceNet-SRM,并使用 AdaCos 人脸识别损失来监督训练过程。ResNet50 是 AdaCos 使用的基础网络;MobileFaceNet 是改进前的特征提取网络;MobileFaceNet-SRM 是加入 SRM 模块后的特征提取网络。评估指标包括在 3 个数据集上的测试准确度以及模型大小,具体表现如表 3 所示。从表 3 可以看出,基于 MobileFaceNet 人脸识别方法在 3 个数据集上的测试准确度与 ResNet50 相比,在 LFW、CFP-FF、AgeDB 3 个数据集上分别提升了 0.49%、0.81%、2.48%,同时模型参数数量大大降低;在将 SRM 模块加入到 MobileFaceNet 后,与改进前的模型相

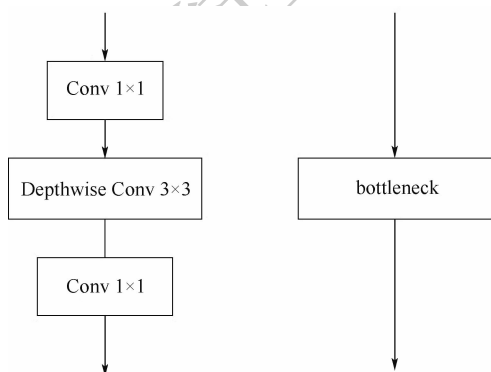


图 4 MobileFaceNet 中 bottleneck 层结构

Fig. 4 Structure of bottleneck layer in MobileFaceNet

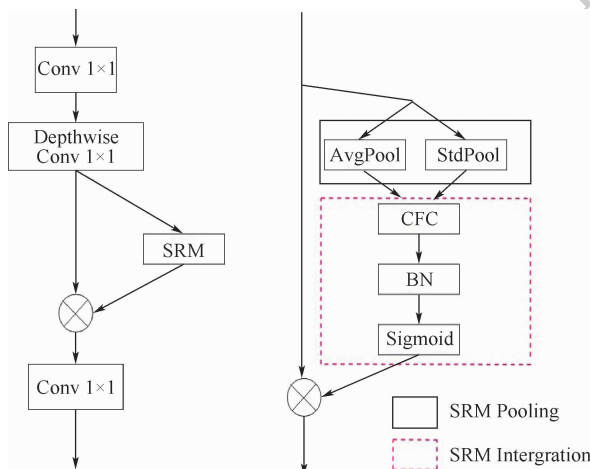


图 5 MobileFaceNet-SRM bottleneck 层结构

Fig. 5 MobileFaceNet-SRM bottleneck layer structure

表 3 基于 AdaCos 的损失函数不同卷积框架
人脸识别模型性能比较

Table 3 Performance comparison of face recognition
models with different convolution frames based on
AdaCos loss function

网络结构	测试准确度/%			模型大小/ MB
	LFW	CFP-FF	AgeDB	
ResNet50	98.41	97.88	88.46	161.5
MobileFaceNet	98.90	98.69	90.94	4.9
MobileFaceNet-SRM	99.15	98.85	91.24	5.2

比,模型参数数量略有提升,但是识别精度在 LFW、CFP-FF、AgeDB 3 个数据集上分别提升了 0.25%、0.16%、0.3%,可以看出改进后的模型在精度上有所提升,证明了改进后模型的有效性。

图 6 是 3 种主干网络在人脸损失函数 AdaCos 监督训练下,在 AgeDB 评测数据集上的受试者工作特征 (Receiver Operating Characteristic, ROC) 曲线。横轴为假正率 (FPR),指的是分类器识别出的假正实例占有负实例的比例;纵轴为真正率 (TPR),指的是识别出的真正实例占有正实例的比例。从图 6 可以看出,MobileFaceNet 相较于 ResNet50 网络识别精度有明显提升,同时模型参数数量大大减少,计算效率有所提高。在向 MobileFaceNet 中引入风格注意力机制后,准确度相较于改进前的模型有所提升,这表明改进后的 MobileFaceNet 人脸识别方法的有效性。

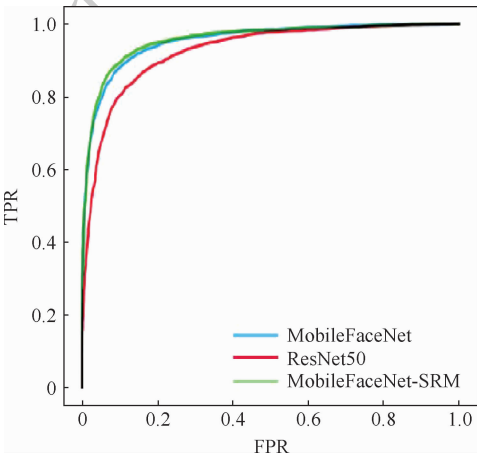


图 6 三种主干网络在 AgeDB 数据集上的 ROC 曲线

Fig. 6 ROC curves of three backbone networks on AgeDB dataset

4 结 论

1) 本文基于 MobileFaceNet 网络提出了一种新的人脸识别方法,通过引入风格注意力机制来增强特征的表达,将改进后的模型作为特征提取网络,大大减少了模型的参数数量,提高了模型的

计算效率,同时提高了提取特征的鲁棒性和代表性。

2) 使用自适应缩放损失函数 AdaCos 作为人脸损失函数来监督训练,通过动态自适应缩放系数,在训练过程中动态调整超参数,加快了模型的收敛速度。在 LFW、CFP-FF 和 AgeDB 3 个人脸数据集上对算法性能进行评测。

3) 实验结果表明,本文方法相较于之前的人脸识别方法,在识别精度有所提升的基础上,大大降低了模型参数的数量,有效地减少了模型的计算量以及模型复杂度。

参考文献 (References)

[1] 徐竞泽,吴作宏,徐岩,等.融合 PCA、LDA 和 SVM 算法的人脸识别[J].计算机工程与应用,2019,55(18):34-37.
XU J Z,WU Z H,XU Y,et al.Face recognition combined with PCA,LDA and SVM algorithms[J].Computer Engineering and Applications,2019,55(18):34-37(in Chinese).

[2] 任飞凯,邱晓晖.基于 LBP 和数据扩充的 CNN 人脸识别研究[J].计算机技术与发展,2020,30(3):62-66.
REN F K, QIU X H. Research on CNN face recognition based on LBP and data augmentation[J]. Computer Technology and Development, 2020, 30(3): 62-66(in Chinese).

[3] SUN Y,WANG X G,TANG X O. Deep learning face representation from predicting 10,000 classes[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2014:1891-1898.

[4] SUN Y,WANG X G,TANG X O,et al. Deep learning face representation by joint identification-verification[C]//Advances in Neural Information Processing Systems,2014:1988-1996.

[5] SUN Y,WANG X G,TANG X O. Deeply learned face representations are sparse,selective,and robust[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2015:2892-2900.

[6] SCHROFF F,KALENICHENKO D,PHILBIN J. FaceNet: A unified embedding for face recognition and clustering[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2015:815-823.

[7] WANG S S,CHEN Y. A joint loss function for deep face recognition[J]. Multidimensional Systems and Signal Processing, 2019,30(3):1517-1530.

[8] DENG J,GUO J K,XUE N N,et al. Arcface: Additive angular margin loss for deep face recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:4690-4699.

[9] ZHANG X,ZHAO R,QIAO Y,et al. AdaCos:Adaptively scaling cosine logits for effectively learning deep face representations[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:10823-10832.

[10] CHEN S,LIU Y,GAO X,et al. MobileFaceNets:Efficient CNNs for accurate real-time face verification on mobile devices[C]//

Chinese Conference on Biometric Recognition. Berlin:Springer, 2018:428-438.

[11] HOWARD A G,ZHU M,CHEN B,et al. MobileNets:Efficient convolutional neural networks for mobile vision applications [EB/OL]. (2017-04-17) [2020-02-20]. https://arxiv.org/abs/1704.04861.

[12] SANDLER M,HOWARD A,ZHU M L,et al. MobileNetV2:Inverted residuals and linear bottlenecks [EB/OL]. (2018-01-13) [2019-03-21]. https://arxiv.org/abs/1801.04381.

[13] LEE H J,KIM H E,NAM H. SRM: A style-based recalibration module for convolutional neural networks[EB/OL]. (2019-05-26) [2020-02-20]. https://arxiv.org/abs/1903.10829?context=cs.

[14] IOFFE S,SZEGEDY C. Batch normalization:Accelerating deep network training by reducing internal covariate shift[EB/OL]. (2015-03-02) [2020-02-20]. http://arxiv.org/abs/1502.03167.

[15] CHU X X,ZHANG B,XU R J. MoGA: Searching beyond MobileNetV3[EB/OL]. (2019-08-04) [2020-02-20]. https://arxiv.org/abs/1908.01314v2.

[16] HU J,SHEN L,ALBANIE S,et al. Squeeze-and-excitation networks[J]. (2017-09-05) [2020-02-20]. https://arxiv.org/abs/1709.01507.

[17] 常思远,李有乘,孙培岩,等. 一种基于 MTCNN 的视频人脸检测及识别方法[J]. 许昌学院学报,2019,38(2):154-157.

CHANG S Y,LI Y S,SUN P Y,et al. A video face detection and recognition method based on MTCNN[J]. Journal of Xuchang University,2019,38(2):154-157 (in Chinese).

作者简介:
张子昊 男,硕士研究生。主要研究方向:行人再识别。
王蓉 女,博士,教授,博士生导师。主要研究方向:模式识别、人工智能。

Improved face recognition method based on MobileFaceNet network

ZHANG Zihao^{1,2,*}, WANG Rong^{1,2,*}

(1. School of Police Information Engineering and Network Security, People's Public Security University of China, Beijing 100038, China;
2. Key Laboratory of Security Technology & Risk Assessment, Beijing 100038, China)

Abstract: In order to solve the problem of more convolutional model parameters and slower convergence speed during training, an improved face recognition method based on MobileFaceNet network is proposed. First, we use the MobileFaceNet network to extract facial features. In the process of extracting features, the number of convolutional layer parameters in the model is reduced by introducing separable convolution. Then, the style attention mechanism is introduced in the MobileFaceNet network to enhance the expression of features. At the same time, the AdaCos face loss function is used to train the model, and the adaptive scaling factor in the AdaCos loss function is used to dynamically adjust the hyperparameters to avoid the effect of artificially setting hyperparameters on the model. Finally, we evaluate the training model on the LFW, AgeDB and CFP-FF test dataset, respectively. The experimental results show that the recognition accuracy of the improved model on the LFW, AgeDB and CFP-FF test dataset has increased by 0.25%, 0.16% and 0.3%, respectively, indicating that the improved model has higher accuracy and robustness than the model before improvement.

Keywords: face recognition; deep learning; MobileFaceNet; AdaCos; convolutional neural network

Received: 2020-02-25; Accepted: 2020-04-18; Published online: 2020-04-21 11:39
URL: kns.cnki.net/kcms/detail/11.2625.V.20200421.0911.001.html
Foundation items: Ministry of Public Security Technology Research Project (2017JSYJB01); the Fundamental Research Funds for the Central Universities (2019JKF111)
* Corresponding author. E-mail: dbdxwangrong@163.com

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0054

一种基于区域划分的改进 ORB 算法



孙浩, 王朋*

(山东交通学院 信息科学与电气工程学院, 济南 250357)

摘 要: 针对传统 ORB 算法所提取的特征点分布不均匀、存在冗杂, 且不具有尺度不变性的问题, 提出了一种基于区域划分的改进 ORB 算法。算法根据需要提取的特征点总数和所划分的区域个数计算每个小区域需要提取的特征点个数, 解决了在特征点提取过程中特征点重叠和特征点冗余的问题; 通过构建图像金字塔, 在每一层图像金字塔上提取特征点, 解决了 ORB 算法提取的特征点不具有尺度不变性的问题。实验结果表明: 在不损失图像匹配精度的同时, 所提算法提取的特征点更加均匀合理, 在提取速度上也较传统 ORB 算法提升了 16% 左右。

关 键 词: ORB 算法; 特征点提取; 区域划分; 图像金字塔; 图像匹配

中图分类号: TP391.4

文献标志码: A

文章编号: 1001-5965(2020)09-1763-07

基于特征的图像匹配算法越来越受到人们的青睐, 特别是在视觉 SLAM (Simultaneous Localization and Mapping)^[1] 和视频监控领域更是成为了一项关键技术。随着多媒体信息技术的发展, 视频监控技术被用于公共区域的视频监控, 以提高公共区域的管理和监控能力。在公共区域视频监控系统的设计中, 需要对公共区域的视频监控图像进行准确的检测和监控, 提取公共区域视频监控的目标特征, 提高公共区域的管理能力, 消除公共区域的危险性。在公共区域除害追踪监测方法的研究上具有十分重要的意义, 而且在公共安全管理 and 刑事侦察领域具有良好的应用价值^[2]。根据图像匹配方法的基本思想可将图像匹配方法分成两大类, 即基于区域的匹配方法和基于特征的匹配方法。两类方法相比而言, 基于特征的匹配方法有计算量小、鲁棒性好、对图像形变不敏感等优点, 所以基于特征的匹配方法成为目前研究的热点。基于特征的图像匹配方法主要包括 3

步: 特征提取、特征描述和特征匹配。当前常见的特征点提取算法有 SIFT (Scale Invariant Feature Transform)^[3]、SURF (Speeded-Up Robust Features)^[4] 以及 ORB (Oriented FAST (Features from Accelerated Segment Test) and Rotated BRIEF (Binary Robust Independent Elementary Features))^[5]。

ORB 算法由 Rublee 等在 2011 年提出^[6], 该算法通过 FAST^[7] 算法提取特征点, 通过 BRIEF^[8] 计算得到描述子, 使得系统对于噪声具有更高的鲁棒性。与之前的算法相比, ORB 算法在计算速度上具有绝对的优势, 在计算速度上是 SURF 的 10 倍, 是 SIFT 的 100 倍^[9], 能满足实时性要求, 但是其鲁棒性不如 SIFT, 且不具备尺度不变性, 在图像匹配中容易造成误匹配^[10]。文献[11]对传统 ORB 算法进行了改进, 提出了一种基于模块化区域分割的 ORB 算法, 算法虽能保证特征点均匀提取, 但无法按照实际要求改变特征点的提取

收稿日期: 2020-02-27; 录用日期: 2020-04-18; 网络出版时间: 2020-05-07 09:47
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200506.1625.008.html
基金项目: 国家自然科学基金 (61502277); 山东省自然科学基金 (ZR2015FL018)
* 通信作者: E-mail: knightwp@126.com

引用格式: 孙浩, 王朋. 一种基于区域划分的改进 ORB 算法[J]. 北京航空航天大学学报, 2020, 46(9): 1763-1769.
SUN H, WANG P. An improved ORB algorithm based on region division[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1763-1769 (in Chinese).

数量,且提取速度也要慢于传统 ORB 算法。文献[12]针对 ORB 算法不具有尺度不变性和匹配精度低的问题,提出了一种结合 SURF 算法和 BBF (Best Bin First)算法的思想对 ORB 算法进行改进,算法虽然提高了匹配精度,但提取速度相较于传统 ORB 算法却慢很多,无法保证系统的实时性。

本文提出了一种全新的基于区域划分的 ORB 算法,在保证特征点提取速度的同时,使得后续的特征点匹配精度也得到了提高。

1 ORB 算法原理

1.1 特征点检测

图像特征点能够简单理解为图像中更重要的点,如轮廓点、较暗区域的高光点、较亮区域的暗点等。ORB 算法运用 FAST 算法来寻找特征点,FAST 的核心思想是找出那些突出的点,即将一个点与周围的点进行比较,如果与周围的大多数点不同,则可以将其挑选出来作为一个特征点。图 1 为 FAST 特征点的提取示意图。

FAST 具体计算过程如下:

步骤 1 从图 1 中挑选一个像素点 P ,判断它是否可以作为特征点。首先假设它的灰度值为 I_p 。

步骤 2 设值一个合适的阈值 t (比如 I_p 的 20%),当 2 个点的灰度值之差的绝对值大于 t 时,可以认为这 2 个点不相同。

步骤 3 以像素点 P 为圆心,挑选出半径为 3 的圆上的 16 个像素点。

步骤 4 假如这 16 个像素点中有连续的 l 个点的灰度都大于 $I_p + t$ 或者都小于 $I_p - t$,那么 P 就可以被认为是一个角点。这里 l 设定为 12,若有至少 12 个点超过阈值,则认为 P 是特征点;否则,认为 P 不是特征点。

步骤 5 为了更高效地检测出特征点,可以增加一项预判操作,这样可以快速地筛选掉大部分不是角点的像素点。通过直接检查邻域圆上的 1、9、5 和 13 这 4 个位置的像素点灰度进行预判。

首先检查 1 和 9,看它们是否和 P 点相同,如果是,再检查 5 和 13。只有当这 4 个像素点中的 3 个同时大于 $I_p + t$ 或者同时小于 $I_p - t$ 时, P 点才被认为是一个角点。如果不能满足上述条件,则不能作为角点,直接排除。通过该操作可以获得 3 种不同的点,如下:

$$S_{p \rightarrow x} = \begin{cases} a & I_{p \rightarrow x} \leq I_p - t \\ b & I_p - t < I_{p \rightarrow x} < I_p + t \\ c & I_p + t \leq I_{p \rightarrow x} \end{cases} \quad (1)$$

式中: $S_{p \rightarrow x}$ 为第 x 个点提取到的点集; $I_{p \rightarrow x}$ 为圆上 16 个点中的第 x 个点的灰度值; a 为圆上比 P 点暗的点; b 为圆上与 P 点相似的点; c 为圆上比 P 点亮的点。

1.2 计算特征点描述子

ORB 使用改进的 BRIEF 算法计算一个特征点的描述子。其核心思想是在特征点 P 的周围以特定模式选取 n 个点对,把这 n 个点对的比对结果组合起来作为描述子^[13]。

计算描述子具体步骤如下:

步骤 1 以特征点 P 为圆心,以 d 为半径做圆 O 。

步骤 2 在圆 O 内选取 n 个点对。为了便于描述,令 $n=4$ 。如图 2 所示,将挑选出的 4 个点对分别标记为: $P_1(A,B)$ 、 $P_2(A,B)$ 、 $P_3(A,B)$ 、 $P_4(A,B)$ 。

步骤 3 定义操作 T 。

$$T(P(A,B)) = \begin{cases} 1 & I_A > I_B \\ 0 & I_A \leq I_B \end{cases} \quad (2)$$

式中: I_A 和 I_B 分别为 A 和 B 的灰度值。

步骤 4 分别对挑选出的点对进行 T 操作,将得到的计算结果进行组合。

假如:

$$\begin{cases} T(P_1(A,B)) = 1 \\ T(P_2(A,B)) = 0 \\ T(P_3(A,B)) = 1 \\ T(P_4(A,B)) = 1 \end{cases}$$

则最终的描述子为:1011。

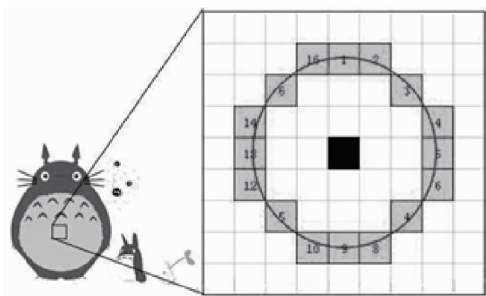


图 1 FAST 特征点提取示意图

Fig. 1 Schematic diagram of FAST feature point extraction

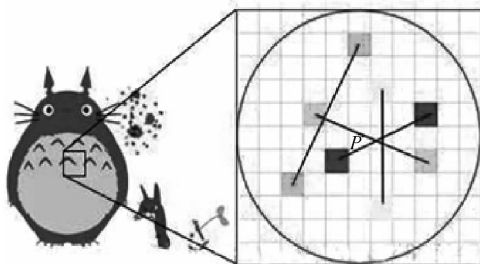


图 2 描述子计算示意图

Fig. 2 Schematic diagram of descriptor calculation

由于 FAST 算法无法检测到特征点方向信息,在 ORB 算法中,利用矩来确定特征点的方向,通过计算特征点附近的图像灰度质心^[14]来表示。在一个小的图像块 B 中,定义图像的矩为

$$m_{pq} = \sum x^p y^q I(x, y) \quad (3)$$

式中: $p, q = 0, 1, 2, \dots$; $I(x, y)$ 为 (x, y) 点处的灰度值。通过式(3)可以得到零阶矩 m_{00} 、一阶矩 m_{10} 和 m_{01} 。

通过矩可以计算其质心为

$$C = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}} \right) \quad (4)$$

接下来连接图像的几何中心 O 与质心 C 求取向向量 $\vec{OC}$ 的方向,假如把 x 和 y 的范围同时保持在 $[-r, r]$ 之间(r 为该特征点邻域的半径),以特征点为坐标原点,则计算得到特征点的方向角为

$$\theta = \arctan(m_{01}/m_{10}) \quad (5)$$

由于 BRIEF 描述子不具备旋转不变性,在对增加方向信息后,在 (x_i, y_i) 的位置,对 n 个点定义矩阵 A 为

$$A = \begin{bmatrix} x_1, x_2, \dots, x_n \\ y_1, y_2, \dots, y_n \end{bmatrix} \quad (6)$$

变换矩阵利用特征点的方向角 θ 与旋转矩阵 R_θ 进行变换:

$$A_\theta = R_\theta A \quad (7)$$

式中: A_θ 为旋转后的特征点对矩阵; $R_\theta = \begin{bmatrix} \cos \theta & \sin \theta \\ -\sin \theta & \cos \theta \end{bmatrix}$ 。

BRIEF 中点对灰度值的比较值又叫做 binary test 值,其定义为

$$\tau(p; x, y) = \begin{cases} 1 & p(x) < p(y) \\ 0 & p(x) \geq p(y) \end{cases} \quad (8)$$

式中: $p(x)$ 和 $p(y)$ 分别为点 x 和点 y 处的像素灰度值。随机选择 n 个点 (x_i, y_i) 就可以生成一个二进制字符串,则生成的特征描述子可以表示为

$$f_n(p) = \sum_{1 \leq i \leq n} 2^{i-1} \tau(p; x_i, y_i) \quad (9)$$

于是,改进后的描述子为

$$g_n(p, \theta) = f_n(p) | (x_i, y_i) \in A_\theta \quad (10)$$

得到改进的描述子后,再进行搜索,在像素块中挑选 n 个点作为最终的描述子^[15]。

2 改进 ORB 算法

ORB 算法使用 FAST 算法检测图像特征点,获得的特征点分布密集,存在冗余。一般来说,特征点越多,则获得的图像匹配越准确,但特征点密集分布对后续的特征描述十分不利,并且会影响

图像匹配精度^[16]。本文提出一种基于区域划分的 ORB 算法,通过对每个区域分别进行特征点提取来解决上述问题。同时,设置自适应阈值,当提取的特征点数量达到条件数量即需要提取的特征点总数时,提前结束来减少特征提取的时间。改进 ORB 算法流程如图 3 所示。

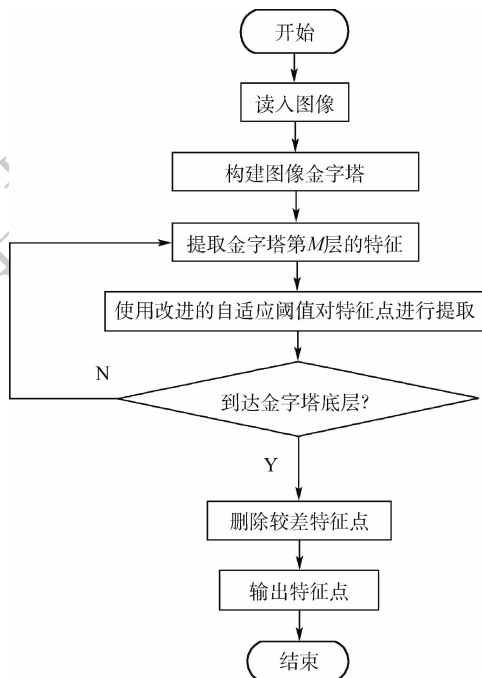


图 3 改进 ORB 算法的流程图

Fig. 3 Flow chart of improved ORB algorithm

算法具体步骤如下:

步骤 1 在构建图像金字塔时,将图像均匀分割成 $M \times N$ 个大小相同的区域, M 为分割的行数, N 为分割的列数,在这些区域中,特征点会随机分布。将区域按照先后列进行排序,表示为 $\{h_1, h_2, h_3, \dots, h_{M \times N}\}$ 。

步骤 2 设置阈值 j 。

$$j = \frac{T_f}{MN} \quad (11)$$

式中:需要提取的特征点总数 T_f 与 ORB 算法提取的特征点总数相同,一般为 500,使得特征点能够更加均匀地分布,而且可以节省筛选最佳特征点的时间。

步骤 3 对每个区域分别进行特征点检测,将每个区域在阈值等于 t 时检测到的 FAST 特征点的个数 n_j 与 j 进行比较,若 n_j 不小于 j ,则选取 j 个最佳点;若 n_j 小于 j ,则降低提取 FAST 特征点时的阈值 t ,重新检测。由于只有当两点之间的灰度值之差的绝对值大于阈值 t 时,才能认为 2 个点是不同的,才可以进行下一步检测。因此,通过降低阈值 t ,能够增加灰度值之差的绝对值大于阈值的点的个数,从而可以检测到更多的角点以供筛选。

步骤 4 经过步骤 3 的检测,有些区域的特征点数会超过需要的特征点数,容易造成检测出来的点彼此相邻,无法成为最佳特征点,需要筛选掉这一部分点。本文采用非最大值抑制的方法。假设 P 和 Q 两点相邻,分别计算两点与周围 16 个点的差分,去掉差分最小的点,直到剩下的点与所需要的点数一致时停止筛选,剩下的即为最佳点。差分和计算公式为

$$V = \max \left(\sum_{x \in S_{\text{bright}}} |I_{p \rightarrow x} - I_p| - t, \sum_{x \in S_{\text{dark}}} |I_p - I_{p \rightarrow x}| - t \right) \tag{12}$$

式中: S_{bright} 为 16 个邻域像素点中灰度值大于 $I_p + t$ 的像素点的集合; S_{dark} 为灰度值小于 $I_p - t$ 的像素点的集合。

步骤 5 遍历所有图像区域,直到提取的特征总数 $\sum n_j$ 达到条件数量时,结束提取。

由于 FAST 特征点不具有尺度不变性,可以采用 SIFT 高斯金字塔来实现,高斯金字塔的构建分为两部分:对图像做不同尺度的高斯模糊和对图像做下采样^[17]。

高斯模糊常被用来模拟物体在人眼中的快速远近效应。其核心是利用高斯函数对输入图像进行模糊模板解卷积运算,去除图像的高频成分,实现图像的模糊化。高斯函数一维和二维表达式分别为

$$G(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \tag{13}$$

$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{14}$$

式中: x,y 分别为原点到 x 轴、 y 轴的距离; σ 为高斯函数 G 的标准差。

本文中通过设置一个比例因子 scaleFactor 和金字塔层数 nlevels ,将原图像 I 按比例因子缩小成 nlevels 幅图像。缩放后的图像 I' 为

$$I' = I / \text{scaleFactor} \cdot k \quad k = 1, 2, \dots, \text{nlevels} \tag{15}$$

分别对 nlevels 幅不同比例的图像通过本文算法提取特征点,提取的特征点总数作为这幅图像的 FAST 特征点以解决 ORB 算法不具有尺度不变性的问题^[18]。

本文中高斯金字塔一共生成 8 层不同尺度的图像,图像金字塔示意图如图 4 所示。图中 $L_0 \sim L_7$ 分别为第 1 幅~第 8 幅下采样图像。

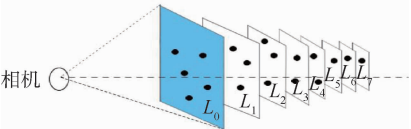
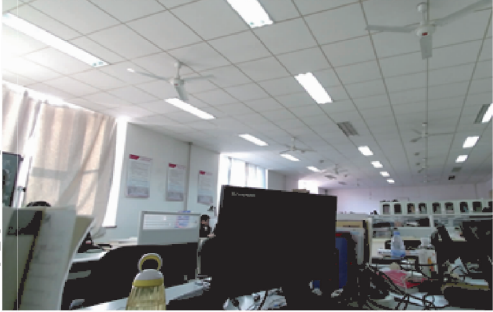


图 4 图像金字塔示意图
Fig. 4 Image pyramid illustration

3 实验结果与分析

本实验在 Linux 系统 + Opencv3.4.1 平台上实现,原始图像如图 5 所示。

传统 ORB 算法和本文算法提取的特征点分别如图 6 所示。图像匹配结果分别如图 7 所示。



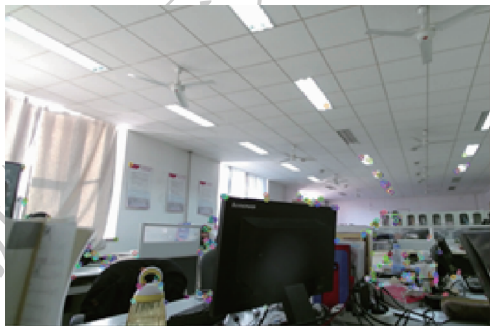
(a) 原始图像1



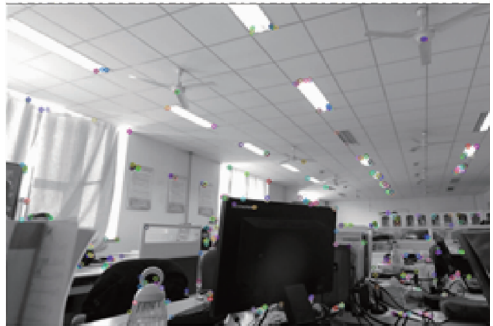
(b) 原始图像2

图 5 原始图像

Fig. 5 Original images



(a) ORB算法



(b) 本文算法

图 6 不同算法提取结果

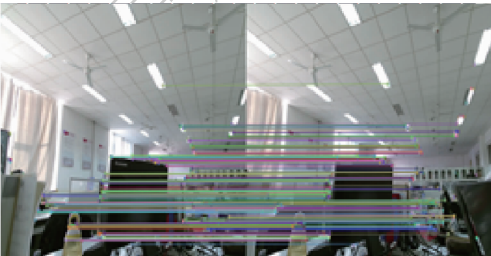
Fig. 6 Extraction results of different algorithms

从图 6 可以看出,相较于图 6(a),图 6(b)的特征点分布更加均匀,大大减少了重叠特征点的数量,说明本文算法提取的特征点更具有代表性,后续的图像匹配也更加准确稳定。匹配准确率对比如表 1 所示。

通过表 1 可以看出,本文算法在匹配精度方面较 ORB 算法提升了 3.4%,较文献[12]改进的 ORB 算法存在一定差距,通过该数据指出了本文算法未来的优化方向。

另一方面,算法运行速度是衡量一个算法优劣的重要标准,在平台相同的情况下,若匹配结果基本相同,则速度越快的算法越好;若运行时间基本相同,则匹配结果越好的算法越优秀。本文将 10 次实验结果取平均值,实验运行时间的部分截图如图 8 所示,算法运行时间和匹配时间比较如表 2、表 3 所示。

从表 2 可以看出,本文算法在提取速度方面比传统的 ORB 算法提升了 16%,较文献[11]改进的 ORB 算法速度提升了 55%,较文献[12]改进的 ORB 算法速度提升了 84%。从表 3 可以看出,在匹配算法相同的前提下,本文算法在匹配速度方面比传统 ORB 算法提升了 9% 左右。本文算法不仅提升了特征点提取速度,也使得所提取的特征点分布更加均匀,匹配速度也得到了提升。



(a) ORB算法



(b) 本文算法

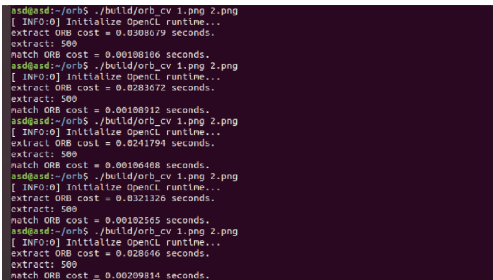
图 7 不同算法匹配结果

Fig. 7 Matching results of different algorithms

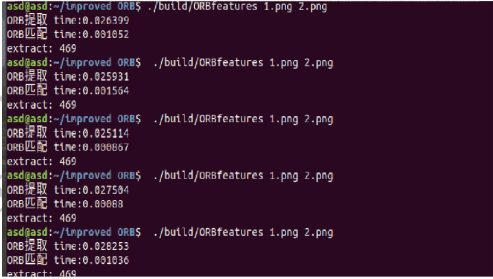
表 1 匹配精度对比

Table 1 Matching accuracy comparison

算法	总提取 点数	正确匹配 点数	匹配准确 率/%
ORB 算法	500	339	67.8
本文算法	469	329	70.1
文献[12]改进的 ORB 算法	253	233	92.1



(a) ORB算法



(b) 本文算法

图 8 不同算法运行截图

Fig. 8 Screenshot of running of different algorithms

表 2 算法运行时间比较

Table 2 Comparison of algorithm running time

算法	总提取 点数	总运行 时间/ms	平均运行 时间/ms
ORB 算法	500	32.7	0.0654
本文算法	469	25.8	0.0550
文献[11]改进的 ORB 算法	262	32.36	0.1235
文献[12]改进的 ORB 算法	253	88.37	0.3493

表 3 算法匹配时间对比

Table 3 Comparison of matching time

算法	总匹配 点数	总匹配 时间/ms	平均匹配 时间/ms
ORB 算法	339	1.50	0.0044
本文算法	329	1.31	0.0040

4 结 论

1) 针对传统 ORB 算法提取的无效特征点较多且不具有尺度不变性的问题,本文提出了一种基于区域划分的改进的特征点提取算法。

2) 算法将图像进行区域划分,提取的特征点都是在区域中表现最佳的点,通过设置自适应阈值,在特征点提取速度方面获得了较大提升,在图像匹配精度较 ORB 算法得到了提升,但对比文献[12]改进的 ORB 算法还存在一定差距。在接下来的工作中,可以通过以下方法对该算法做进一步的优化:在构建图像金字塔时,可以对图像信息进行计算,并自动调节尺度信息,以保证图像的

尺度信息更加合理,使算法在尺度不变性方面得到进一步提升。

参考文献 (References)

[1] 高翔,张涛.视觉 SLAM 十四讲:从理论到实践[M].北京:电子工业出版社,2017:10-14.
GAO X,ZHANG T. 14 lectures on visual SLAM:From theory to practice[M]. Beijing: Publishing House of Electronics Industry,2017:10-14 (in Chinese).

[2] 张全,盛妍,吴佐平,等.公共区域监控视频数据目标特征跟踪定位方法[J].自动化与仪器仪表,2020(1):51-54.
ZHANG Q,SHENG Y,WU Z P,et al. Target tracking and location method for common area surveillance video data[J]. Automation & Instrumentation,2020(1):51-54 (in Chinese).

[3] LOWE D G. Distinctive image features from scale invariant keypoints[J]. International Journal of Computer Vision,2004,60(2):91-110.

[4] BAY H,TUYTELAARS T,VAN GOOL L. Surf:Speeded up robust features[C]//European Conference on Computer Vision. Berlin:Springer,2006:404-417.

[5] ROSTEN E,PORTER R,DRUMMOND T. Faster and better: A machine learning approach to corner detection[J]. Analysis and Machine Intelligence,2008,32(1):105-119.

[6] RUBLEE E,RABAUD V,KONOLIGE K,et al. ORB: An efficient alternative to SIFT or SURF[C]//Proceedings of IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2011:2564-2571.

[7] VISWANATHAN D G. Features from accelerated segment test (FAST)[EB/OL]. (2016-04-15)[2020-01-10]. https://www.cnblogs.com/ronny/p/4078710.html?utm_source=tuicool.

[8] CALONDER M,LEPETIT V,STRECHA C,et al. Brief: Binary robust independent elementary features[C]//European Conference on Computer Vision. Berlin:Springer,2010:778-792.

[9] 刘伟,钱莉.基于 OpenCV 环境的 SIFT、SURF、ORB 算法比较分析[J].化工自动化及仪表,2018,45(9):714-716.
LIU W,QIAN L. Comparative analysis of SIFT and SURF and ORB algorithms based on OpenCV environment[J]. Control and Instruments in Chemical Industry,2018,45(9):714-716 (in Chinese).

[10] ZHAO H Y,ZHAO H C,LV J F,et al. Multimodal image matching based on multimodality robust line segment descriptor[J]. Neurocomputing,2016,177:290-303.

[11] 潘盼.一种改进的 ORB 算法[J].微型机与应用,2017,36(20):23-26.
PAN P. An improved ORB algorithm[J]. Microcomputer & Its Applications,2017,36(20):23-26 (in Chinese).

[12] 马丹,赖惠成.改进 ORB 算法的图像匹配[J].计算机仿真,2018,35(10):274-278.
MA D,LAI H C. Image matching based on improved ORB[J]. Computer Simulation,2018,35(10):274-278 (in Chinese).

[13] 陈思聪,刘晶红,何林阳,等.一种用于图像拼接的改进 BRISK 算法[J].液晶与显示,2016,31(3):324-330.
CHEN S C,LIU J H,HE L Y,et al. Improved BRISK algorithm used in image mosaic[J]. Chinese Journal of Liquid Crystals and Displays,2016,31(3):324-330 (in Chinese).

[14] 范新南,顾亚飞,倪建军.改进 ORB 算法在图像匹配中的应用[J].计算机与现代化,2019(2):1-6.
FAN X N,GU Y F,NI J J. Application of improved ORB algorithm in image matching[J]. Computer and Modernization,2019(2):1-6 (in Chinese).

[15] HUANG C,ZHOU W D. Fast scene matching navigation algorithm based on ORB[J]. Journal of Information and Computational Science,2014,11(11):3857-3863.

[16] 李玉峰,李广泽,谷绍湖,等.基于区域分块与尺度不变特征变换的图像拼接算法[J].光学精密工程,2016,24(5):1197-1205.
LI Y F,LI G Z,GU S H,et al. Image mosaic algorithm based on area blocking and SIFT[J]. Optics and Precision Engineering,2016,24(5):1197-1205 (in Chinese).

[17] 王健,于鸣,任洪娥.一种用于图像拼接的改进 ORB 算法[J].液晶与显示,2018,33(6):520-527.
WANG J,YU M,REN H E. Improved ORB algorithm used in image stitching[J]. Chinese Journal of Liquid Crystals and Displays,2018,33(6):520-527 (in Chinese).

[18] 姚海芳,郭宝龙.一种基于 ORB 的特征匹配算法[J].电子设计工程,2019,27(16):175-179.
YAO H F,GUO B L. An ORB-based feature matching algorithm[J]. Electronic Design Engineering,2019,27(16):175-179 (in Chinese).

作者简介:
孙浩 男,硕士研究生。主要研究方向:图像特征提取与匹配。
王朋 男,博士,副教授,硕士生导师。主要研究方向:数字图像处理、视频编码与传输控制、计算机视觉。

An improved ORB algorithm based on region division

SUN Hao, WANG Peng*

(School of Information Science and Electrical Engineering, Shandong Jiao Tong University, Jinan 250357, China)

Abstract: The feature points extracted by the traditional ORB algorithm are not evenly distributed, are redundant and have no scale invariance. To solve this problem, this paper proposes an improved ORB algorithm based on region division. According to the total number of feature points to be extracted and the number of regions to be divided, the algorithm calculates the number of feature points to be extracted for each region, which solves the problem of feature point overlap and redundancy in the feature point extraction process. By constructing the image pyramid and extracting feature points on each layer, the problem that the feature points extracted by ORB algorithm do not have scale invariance is solved. The experimental results show that the feature points extracted by our algorithm are more uniform and reasonable without losing the accuracy of image matching, and the extraction speed is about 16% faster than that of the traditional ORB algorithm.

Keywords: ORB algorithm; feature point extraction; region division; image pyramid; image matching

Received: 2020-02-27; **Accepted:** 2020-04-18; **Published online:** 2020-05-07 09:47

URL: kns.cnki.net/kcms/detail/11.2625.V.20200506.1625.008.html

Foundation items: National Natural Science Foundation of China (61502277); Natural Science Foundation of Shandong Province (ZR2015FL018)

* **Corresponding author.** E-mail: knightwp@126.com

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0065



基于面部视频分析的生命体征检测

陈辉¹, 郑秀娟^{1,*}, 倪宗军¹, 张昀², 杨晓梅¹

(1. 四川大学 电气工程学院, 成都 610065; 2. 西安交通大学 电子与信息学院, 西安 710049)

摘 要: 在面部视频中提取生命体征相关的生理信号时易受环境光和受试者头部运动的影响, 为了降低外界干扰并提高生命体征检测的准确度, 提出了一种联合集合经验模态分解(EEMD)算法与信号质量检测的面部视频分析方法, 用于精确检测人体的心率与呼吸频率等生命体征。通过公开数据集进行实验验证, 实验结果表明, 所提方法相较于目前已有的常用信号处理方法能够得到更精确的心率与呼吸频率的估计值, 所得估计值与标准值的相关系数分别高于0.9和0.8。同时, 所提方法将为实时活体人脸识别提供一种思路, 也有助于丰富监控视频智能分析的应用研究。

关 键 词: 面部视频; 成像式光电容积描记法; 集合经验模态分解(EEMD); 信号质量检测; 抗干扰

中图分类号: V221⁺.3; TB553

文献标志码: A **文章编号:** 1001-5965(2020)09-1770-08

近年来视频监控技术快速发展, 针对监控视频的智能分析需求深入到公共空间和私人生活的各个方面, 涉及了如交通监控、犯罪预防、人群监控、身份认证以及居家监护等多种不同应用场景^[1]。由于视频中的人脸包含最为丰富的个人特征信息, 因此针对面部视频的智能分析方法也受到广泛的关注。早在20世纪60年代Bledsoe就提出了基于人脸的几何特征的人脸识别方法^[2]; He等也在2015年提出了一种基于局部二值模式(Local Binary Patterns, LBP)关键面部区域中的表情特征融合方法进行面部表情识别^[3]; 2004年, Ahonen等利用LBP人脸单个特征直方图实现人脸识别^[4]; 2016年, Lenc和Kral提出利用一组Gabor滤波器实现对基准点的自动选择, 解决了在使用LBP进行人脸识别时的基准点的位置和数量是固定的问题^[5]。随着计算机视觉理论和技术的发展, 人脸检测与识别技术日臻完

善。在面部视频分析领域中, 之前的许多研究都围绕人脸结构特征展开了深入研究, 但忽略了隐含于视频中的与生命体征相关的生理信号。近年来, 越来越多的研究开始关注对视频及图像序列中的生命体征相关生理信号的提取和分析工作^[6-7], 这将为充分利用监控视频中隐含的人体信息进行活体识别以及智能行为分析提供思路, 有助于丰富监控视频智能分析的应用场景。

2000年, Wu等提出成像式光电容积描记法(imaging Photoplethysmography, iPPG)及其技术框架实现了从视频中获得人体的心率参数估计^[8]。iPPG技术是根据面部视频中的皮下浅层血管血流灌注变化导致的皮肤颜色轻微变化, 从而通过分离血容量脉冲(Blood Volume Pulse, BVP)信号来得到相应的生命体征估计。此后, 多项研究均以iPPG技术为基础实现基于视频的远距离生命体征参数检测工作^[9-10]。然而由于视频中隐含的

收稿日期: 2020-03-02; 录用日期: 2020-04-18; 网络出版时间: 2020-05-08 16:18

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200508.1524.001.html

基金项目: 成都市重点研发支撑计划技术创新研发项目(2020-YF05-00056-SN)

*通信作者: E-mail: xiujuanzheng@scu.edu.cn

引用格式: 陈辉, 郑秀娟, 倪宗军, 等. 基于面部视频分析的生命体征检测[J]. 北京航空航天大学学报, 2020, 46(9): 1770-1777.

CHEN H, ZHENG X J, NI Z J, et al. Vital signs detection via facial video analysis[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1770-1777 (in Chinese).

BVP 信号成分十分微弱,且容易受到环境光照和运动的影响,产生低频基线漂移与高频噪声污染^[11],从而使得 iPPG 方法进行生命体征检测的精准度不稳定。

为抑制噪声对 iPPG 技术的干扰,2010 年 Poh 等使用独立成分分析法 (Independent Component Analysis, ICA) 处理面部视频中 RGB 三通道信号,来获取视频中 0.5 ~ 2 Hz 之间的频率成分,之后重构 BVP 信号,进而获得心率的估计^[12]。随着现代信号处理技术的发展,2009 年,雷恒波等采用经验模态分解 (Empirical Mode Decomposition, EMD) 来自适应处理非平稳信号,将原始信号分解成一系列不同频率成分的本征模态函数 (Intrinsic Mode Functions, IMF),从而重建纯净的 BVP 信号并用于检测相应的生命体征^[13-14]; 2018 年, Zhao 等提出使用改进的 EMD 方法 (Complete Ensemble Empirical Mode Decomposition, CEEMD) 来重建纯净的 BVP 信号^[15]。2019 年, Favilla 等提出使用带通滤波器来获取高帧率视频,并且使用快速 ICA 方法将 BVP 信号与噪声分离,从而得到准确的生命体征的估计值^[16]。

这些信号处理方法都主要解决消除或减弱由运动或光照影响造成的伪影误差,但当视频存在人脸丢失,或者无法准确识别人脸感兴趣区域 (Region of Interest, ROI) 时,仍会使获取的 BVP 信号质量下降,从而影响到生命体征估计的准确率。

由此本文以面部视频中人脸检测为基础,提出了一种新型面部视频分析方法实现自适应非接触生命体征检测。该方法可在面部视频分析时,对存在环境光影响和头部运动的面部视频给出较为精确的心率与呼吸频率的估计。

1 本文方法

本文以人脸检测为基础,使用集合经验模态分解 (Ensemble Empirical Mode Decomposition, EEMD) 算法^[17]与信号质量检测提高基于面部视频的生命体征检测方法的准确性。本文提出的联合集合经验模态分解算法和信号质量检测的面部视频分析方法由 5 个主要步骤构成,流程如图 1 所示。

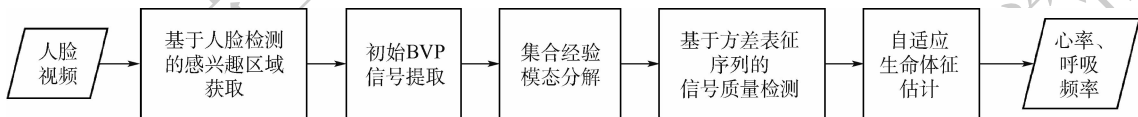


图 1 本文方法流程

Fig. 1 Flowchart of proposed method

1.1 基于人脸检测的感兴趣区域获取

为了更好地检测并跟踪视频中的人脸,本文使用具有 3 个深层卷积网络的级联架构 MTCNN (Multi-Task Convolutional Neural Networks)^[18] 和非极大值抑制策略^[19] 得到人脸识别框与面部标签,同时实现人脸检测与对齐。在确定人脸的范围之后,调用 Dlib 库^[20] 定位人面部特征点。根据面部特征点定位鼻子及额头区域的最大内接矩形可以跟随头部运动实现面部感兴趣区域的自适应获取,如图 2 所示。图中:实心点为 68 个特征点 (粗略 ROI); 矩形框为精细 ROI, 包括鼻子和前额区域。

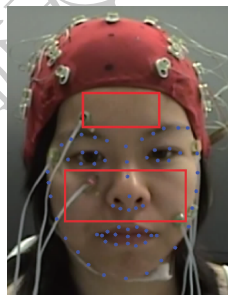


图 2 面部 ROI 选择

Fig. 2 Facial ROI selection

1.2 初始 BVP 信号提取

在确定面部 ROI 后,对 ROI 内像素点进行灰度归一化处理,再进行瑞利分布匹配,去除灰度值落在瑞利分布之外的像素点从而抑制噪声干扰。接着采用 CHROM 模型^[17] 构建 ROI 像素点色彩模型以降低运动带来的噪声影响,即将 RGB 三通道的信号进行线性组合得到初始 BVP 信号。

1.3 集合经验模态分解

为了避免 EMD 算法的模式混合问题,本文采用 EEMD 算法^[21] 将初始 BVP 信号分解为真实的 IMF 函数组合。

因此初始 BVP 信号先由 EMD 算法分解为 N 个 IMF 函数,所有的 IMF 函数和残差 $r(t)$ 可以重构输入信号 $x(t)$ 。

$$x(t) = \sum_{i=1}^N \text{IMF}_i(t) + r(t) \quad (1)$$

根据 EEMD 算法原理,初始 BVP 信号 $x(t)$ 分解步骤如下:

步骤 1 在原始信号 $x(t)$ 中加入均值为 0、标准差为原始信号标准差 0.2 倍的高斯白噪声 $w^{(j)}$, 即 $x^{(j)} = x(t) + aw^{(j)}$ ($j=1,2,\dots,N$), 其中的白噪声的幅度大小为 $a=0.2$ 。

步骤 2 计算 $x^{(j)}$ ($j=1,2,\dots,N$) 的 EMD 算法得到本征模态 $\text{IMF}_k^{(j)}$, 其中 $k=1,2,\dots,K$ 为模态。

步骤 3 将 IMF_k 作为 $x(t)$ 的第 k 阶模态均值, 通过将相应的 $\text{IMF}_k^{(j)}$ 求平均得到。

$$\overline{\text{IMF}}_k = \frac{1}{N} \sum_{j=1}^N \text{IMF}_k^{(j)} \quad (2)$$

当得到所有的 IMF 函数后, 应该舍去与噪声相关的 IMF 函数。由人体血液循环可知, BVP 信号主要受心肺频率的调制, 而呼吸成分频率范围为 0.1~0.5 Hz, 心率成分频率范围为 0.7~4 Hz。为了去除伪影, 计算各 IMF 函数的频谱并确定其主频率, 即获得最大幅度的频率。根据各 IMF 函数的主频率将其归入对应的心肺信号频率范围, 即将 IMF 函数分为心率组和呼吸频率组。为最终从 BVP 信号中得到准确的心率和呼吸频率等生命体征的估计, 将心率组和呼吸频率组的 IMF 函数通过正交变换, 将存在相关性的 IMF 函数转换为一些线性不相关的变量, 即为主成分 (Principal Component signal, PCs)。将主成分进行排序, 第一主成分含有所选 IMF 函数中存在的大部分变化。因此, 在心率组中的 IMF 函数上得到的第一主成分对应着心率, 同样, 呼吸组中的第一主成分对应着呼吸频率, 如图 3 所示。

1.4 基于方差特征序列的信号质量检测

在实际的面部视频分析中, 初始 BVP 信号质量往往受到诸多外界因素的干扰, 即使经过 EEMD 分解后获得的主成分也会存在信号质量不高的问题。因此, 在得到心率主成分 $X_{hr}(t)$ 和呼吸频率主成分 $X_{br}(t)$ 后, 需要检测主成分信号质量来提高生命体征检测的准确性。本文采用方差特征序列 (Variance Characterization series, VCs)^[22] 进行主成分信号的频谱质量检测, 以处理心率主成分 $X_{hr}(t)$ 为例 (呼吸频率主成分 $X_{br}(t)$ 处理方法同理), 具体步骤如下:

步骤 1 根据峰值检测算法挑选出 $X_{hr}(t)$ 所有局部最大值 M_i ($i=1,2,\dots$) 以及局部最小值 m_i ($i=1,2,\dots$)。

步骤 2 将每个局部最大值 M_i 和后续的 7 个局部最大值进行方差计算, 记为 σ_{M_i} 。以相同方法处理局部最小值, 得到方差记为 σ_{m_i} 。当计算最后 7 个局部值的方差时, 以最后 1 个数复制填充至 8 个再进行计算。

步骤 3 计算方差特征序列:

$$\delta_{M_i} = \frac{|\sigma_{M_i}|}{(\log(M_{i+1}) - \log(M_i))^2} \quad (3)$$

$$\delta_{m_i} = \frac{|\sigma_{m_i}|}{(\log(m_{i+1}) - \log(m_i))^2} \quad (4)$$

式中: δ_{M_i} 和 δ_{m_i} 分别为局部最大值下的方差特征序列和局部最小值下的方差特征序列; $\log(*)$, $* \in (M_i, M_{i+1}, m_i, m_{i+1})$ 为在时间序列 $X_{hr}(t)$ 中的局部最大值和局部最小值对应的索引值。

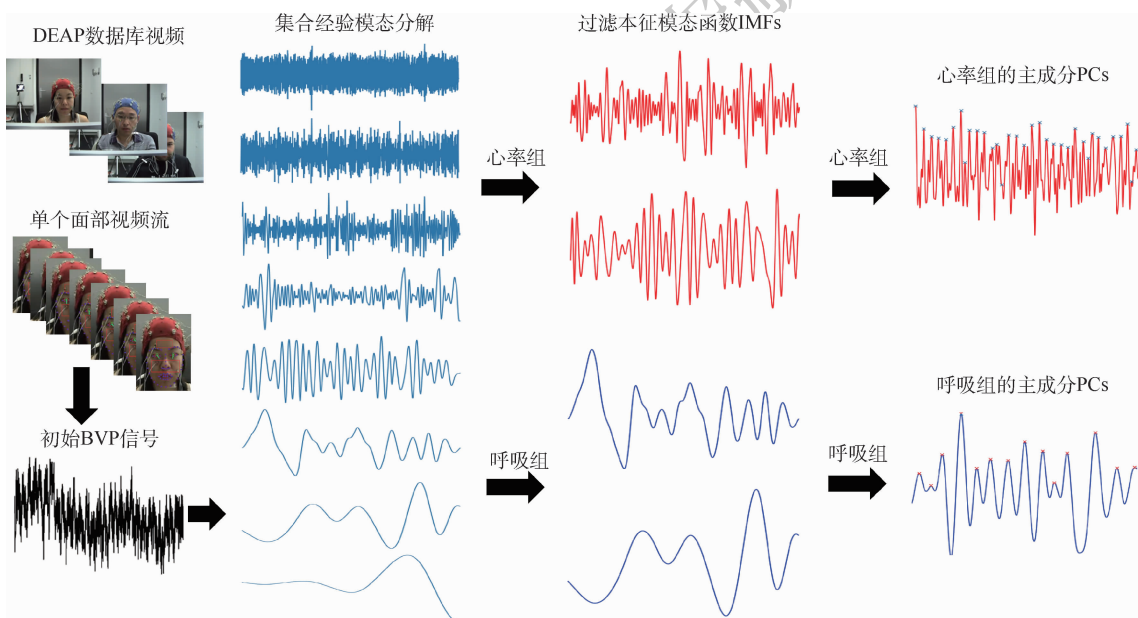


图 3 用 EEMD-PCA 技术从 BVP 信号中提取心率和呼吸频率的主成分的步骤

设定的 8 个局部值计算方差有利于减少基线漂移,由式(3)、式(4)可以得到 $X_{hr}(t)$ 的 2 个方差特征序列。原始主成分信号以及对应的方差特征序列如图 4 所示。图中:三角表示主成分信号的局部最大值(无单位),圆点表示主成分信号的局部最小值(无单位)。可以看到,当不同主成分的波形变化时,方差特征序列的变化情况。

根据获得的方差特征序列,基于 2 个准则来判断 $X_{hr}(t)$ 质量:

1) δ_{M_i} 的值远大于阈值 T 。

2) 第 $i+4$ 个最大值和第 i 个最大值之间的距离远小于 D_1 或者第 $i+2$ 的最大值和第 i 的最大值之间的距离 D_2 。

上述准则中,阈值 T 由先验的无干扰 BVP 信号计算给定,且 D_1 和 D_2 则由前一时间窗估计的心率 (H_{pre}) 和采样频率 ρ 给定,即

$$D_1 = \rho \cdot 180 / H_{pre} \quad (5)$$

$$D_2 = \rho \cdot 60 / H_{pre} \quad (6)$$

式中: H_{pre} 为按照上述同样方法在前一个时间窗的计算值,180 bpm 与 60 bpm 分别为人体在静坐情况下的心率上限值与下限值。

1.5 自适应生命体征估计

若根据 1.4 节计算得出对应的主成分信号质量为不合格,即说明在频谱分析时,会出现大量不符合生理特征区间的频谱高峰值,此时对经过 1.3 节处理之后的主成分信号 $X_{hr}(t)$ 与 $X_{br}(t)$ 使用频谱跟踪算法来获取生命体征的估计值,如图 5 所示,仍以处理 $X_{hr}(t)$ 为例($X_{br}(t)$ 同理),其

具体的步骤如下:

步骤 1 使用三次样条插值平滑处理,以及对主成分信号 $X_{hr}(t)$ 进行三阶的巴特沃斯带通滤波提取相应的频率成分,进一步消除因为环境光照变化和头部运动带来的噪声。

步骤 2 计算将平滑后的主成分 $X_{hr}(t)$ 频谱,选取前 2 个最大的频率值 f_1 、 f_2 ,并且计算对应心率数值为 HR_1 和 HR_2 。

步骤 3 将这 2 个数值 HR_1 和 HR_2 与上一个时间窗估计的频率 f_{pre} 对应的心率 HR_{pre} 作对比。

1) 当 2 个估计值与 HR_{pre} 差值均控制在 10 bpm 之内,且与 HR_{pre} 差值的绝对值均不相等,即选取最为接近 HR_{pre} 的心率估计值作为此次的最终的心率估计值 HR_{final} 。

2) 当 2 个心率估计值与 HR_{pre} 差值均相等且控制在 10 bpm 之内,将时间窗缩短计算各个时间窗之内的数值,依次判断心率变化的趋势。当心率为上升趋势时选取比较大的心率估计值作为 HR_{final} ,反之,则选取较小的心率估计值作为 HR_{final} 。

3) 当 2 个估计值与 HR_{pre} 差值,一个小于 10 bpm,另一个大于 10 bpm,那么,即选取小于 10 bpm 对应的估计值作为 HR_{final} 。

4) 若 2 个心率估计值与 HR_{pre} 的差值均大于 10 bpm,即说明该主成分信号质量不高,不能使用频谱分析,则改为使用时域计数方式获取的心率估计值作为最终的心率估计值 HR_{final} 。

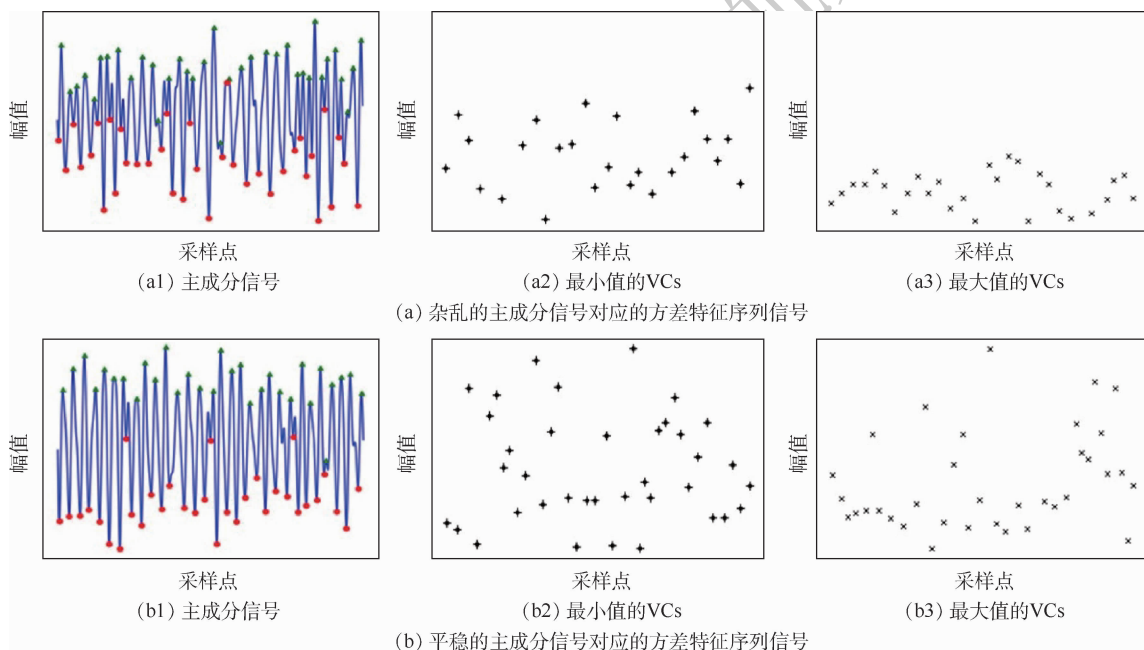


图 4 原始主成分信号及对应的方差特征序列表示

Fig. 4 Original principal component signals and corresponding variance characterization series representation

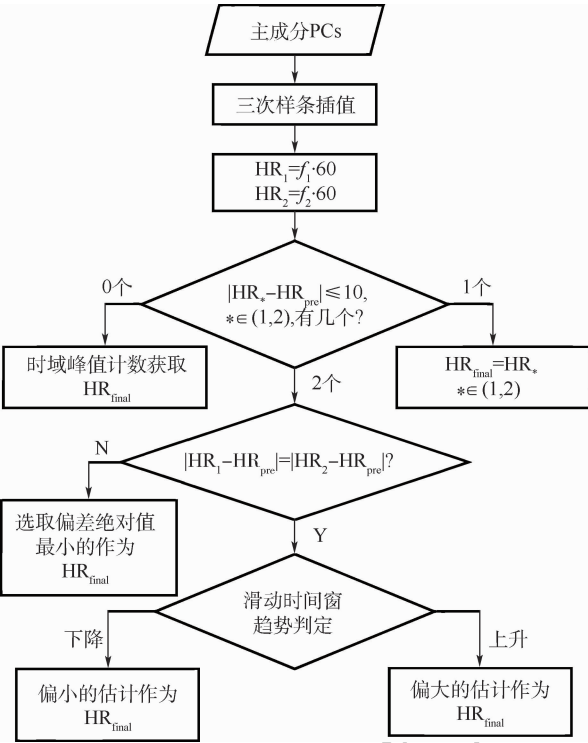


图5 频谱跟踪算法流程图

Fig. 5 Flowchart of spectrum tracking algorithm

如果方差特征序列满足 1.4 节中 2 个基本准则,代表此时获得的主成分信号 $X_{hr}(t)$ 与 $X_{br}(t)$ 质量合格,受到的干扰少,可以直接使用巴特沃斯带通滤波后进行频谱分析,即可得到相应的心肺频率 f_{hr} 、 f_{br} ,通过式(7)和式(8)获得最终心率和呼吸频率的估计值。

$$HR = f_{hr} \cdot 60 \tag{7}$$

$$RR = f_{br} \cdot 60 \tag{8}$$

至此完成了根据主成分信号不同质量情况的自适应生命体征估计。

2 实验结果及分析

2.1 实验数据集

为控制实验变量的单一性,本文使用公开的多模态数据库 DEAP 数据集^[23]进行对比实验。DEAP 数据集中包含有 32 名受试者(年龄分布为 19~37 岁)在观看情绪激发视频时的 1 分钟面部视频以及同步多通道生理信号,其中由呼吸带测得的呼吸频率和指尖 PPG 信号测得的心率将作为本文生命体征估计的标准值。

2.2 实验结果对比

为了验证本文方法的心率和呼吸频率检测的准确率,使用估计误差的平均绝对误差(Mean Absolute Error, MAE)、平均误差百分比(Mean Error Percentage, MEP)以及 P 相关系数(Pearson

correlation coefficient)这些统计指标进行验证分析。选取较早研究中常用的信号处理方法 ICA^[24](2011 年)、AR^[25](2014 年)和近年提出的改进方法 CEEMD^[15](2018 年)、FastICA^[16](2019 年)进行对比实验,验证本文方法性能的有效性。

所有方法都基于同一个实验数据进行,视频预处理方法保持一致,所有的预处理参数保持一致,面部视频时间窗宽 30 s。对比方法保留已报道的参数设定值,ICA^[24]选择的独立成分数为 3,AR^[26]使用的回归参数设定为 3,CEEMD^[15]不用设置其他参量。本文方法中使用的 EEMD 算法也不用特别进行参量设置。

表 1 列出了本文方法与 4 种比较方法在心率估计的准确度比较结果(面部视频时间窗宽为 30 s)。与 ICA、AR、CEEMD、FastICA 4 种方法做比较时,本文方法的心率估计结果 MAE 最小,控制在 3 bpm 之内,同时 MEP 控制在 4% 左右,P 相关系数大于 0.9。表 2 列出了本文方法与 4 种比较方法在呼吸频率估计的准确度比较结果(面部视频时间窗宽为 30 s)。与其他方法相比,本文方法获得了最低的呼吸频率估计 MEP,约为 7.8%,同时 MAE 约为 3 bpm,P 相关系数达到 0.8 以上。由上述结果可见,由面部视频获得的呼吸频率的准确度低于心率的准确度,其原因为呼吸相关信号通常为低频成分,更容易受到外界因素干扰,不容易从视频中将其精确分离。

表 1 不同方法的心率比较

Table 1 Comparison of heart rate among different methods

方法	MAE/bpm	MEP/%	P 相关系数
ICA ^[24]	9.866	11.966	0.733
AR ^[25]	5.334	6.755	0.751
CEEMD ^[15]	3.804	4.881	0.860
FastICA ^[16]	4.541	5.881	0.821
本文方法	2.928	4.161	0.902

表 2 不同方法的呼吸频率比较

Table 2 Comparison of respiratory rate among different methods

方法	MAE/bpm	MEP/%	P 相关系数
ICA ^[24]	7.300	16.292	0.556
AR ^[25]	6.718	14.294	0.612
CEEMD ^[15]	3.804	9.878	0.753
FastICA ^[16]	4.541	15.726	0.623
本文方法	3.085	7.769	0.829

2.3 结果讨论

当从面部视频中获取的生理信号足够稳定,使用以傅里叶变换为基础的频谱分析获得的生命

体征估计具有较高精确度。然而,通常得到的视频信号,尤其是监控视频,常常受到诸多外界因素的干扰,如环境光变化、人体运动等。这些都会给面部视频中隐含的生理信号带来强烈的噪声。因此使用ICA和时域滤波相结合的方法可以在一定程度上抑制外界干扰引入视频的噪声,但使用ICA方法进行信号盲源分离,设定的盲源数量为3。而监控视频中存在的光源和运动噪声源的数量难以确定,因此在实际应用中该方法所获得的生命体征检测结果也不理想。

AR模型常用于时间序列的分析,但要求信号平稳,因此即使引入了背景兴趣区作为环境光噪声参考来提高该方法对光照变化的稳定性,但仍然无法抑制运动噪声干扰。因此,在对比实验中AR^[25]方法相较于ICA^[24]方法所得到的生命体征估计结果的精确度有所提高,但低于CEEMD^[15]和FastICA^[16]方法获得的估计精确度。

使用基于EMD方法以及该类型的拓展方法,如EEMD和CEEMD^[15]方法,其优势在于不需要考虑盲源数量,也无需设定任何基函数,便可以将复杂信号分解为有限个IMF组合,并且各IMF分别包含了原信号的不同时间尺度下的局部特征信号。但同一时间尺度成分由EMD分解成不同的IMF时,会产生模态混叠这一现象。EEMD与CEEMD都是通过改变极值点的分布抑制模态混叠问题,采用的是在原信号中加入白噪声来使真实信号分量在分解的过程逐步显现。使用EEMD时,剩余噪声会随着集成平均的次数逐渐减少,而在使用CEEMD时的剩余噪声一直维持在一个较小的程度,在保证较小的剩余噪声干扰的情况下,能够节省计算时间。在针对面部视频分析时,本文使用EEMD方法无需针对复杂的干扰情况预先设定视频中噪声源数量。同时,本文方法使用信号质量检测和频谱跟踪算法,可以有效处理视频中面部数据缺失或质量降低的问题,来提高生命体征估计的准确度。

最后,本文方法仍然无法高效地抑制剧烈运动干扰的影响。因此,在未来的工作中,需要提高本文方法的泛化能力,进一步针对监控视频特点提高视频中生命体征检测的准确度。

3 结 论

本文基于面部视频分析给出了一种人体生命体征检测方法,并通过公开数据集的对比实验验证得出如下结论。

1) 为了提高视频中生命体征检测的准确度,

抑制外界环境光及头部运动的干扰,提出以人脸检测与跟踪为基础,联合EEMD算法与信号质量检测方法进行面部视频分析,并给出了心率和呼吸频率检测的具体流程与算法。

2) 通过公开数据集验证了所提方法从面部视频中获心率与呼吸频率的准确度,所得估计值与标准值的相关系数分别高于0.9和0.8。

3) 所提方法不仅为实时活体检测提供了思路,而且也有助于丰富监控视频的智能分析的应用场景。

参考文献 (References)

- [1] 杨哲. 基于视频监控的人群异常行为检测[D]. 秦皇岛:燕山大学,2019.
- [2] YANG Z. Detection of crowd abnormal behavior based on video surveillance[D]. Qinhuangdao: Yanshan University, 2019 (in Chinese).
- [3] BLEDSE W. The model method in facial recognition; Technical report: PRI-15[R]. Palo Alto: Panoramic Research, Inc., 1964.
- [4] HE J, CAI J F, FANG L Z, et al. A method of facial expression recognition based on LBP fusion of key expressions areas[C]// The 27th Chinese Control and Decision Conference. Piscataway: IEEE Press, 2015: 4200-4204.
- [5] AHONEN T, HADID A, PIETIKÄINEN M. Face recognition with local binary patterns[C]// European Conference on Computer Vision. Berlin: Springer, 2004: 469-481.
- [6] LENC L, KRÁL P. Local binary pattern based face recognition with automatically detected fiducial points[J]. Integrated Computer-Aided Engineering, 2016, 23(2): 129-139.
- [7] KRANJEC J, BEGUS S, GERŠAK G, et al. Non-contact heart rate and heart rate variability measurements: A review[J]. Bio-medical Signal Processing and Control, 2014, 13: 102-112.
- [8] SUN Y, THAKOR N. Photoplethysmography revisited: From contact to noncontact, from point to imaging[J]. IEEE Transactions on Bio-Medical Engineering, 2016, 63(3): 463-477.
- [9] WU T, BLAZEK V, SCHMITT H J. Photoplethysmography imaging: A new noninvasive and noncontact method for mapping of the dermal perfusion changes[C]// Optical Techniques and Instrumentation for the Measurement of Blood Composition, Structure, and Dynamics. Bellingham: International Society for Optics and Photonics, 2000, 4163: 62-70.
- [10] LI X B, CHEN J, ZHAO G Y, et al. Remote heart rate measurement from face videos under realistic situations[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 4264-4271.
- [11] BEATSON A, GREEN R. Calculating a vagal tone index using webcam photoplethysmography[C]// Proceedings of the 29th International Conference on Image and Vision Computing. New York: ACM, 2014: 242-246.
- [12] CHENG J, CHEN X, XU L X, et al. Illumination variation-resistant video-based heart rate measurement using joint blind source separation and ensemble empirical mode decomposition

- [J]. IEEE Journal of Biomedical and Health Informatics, 2017, 21(3):1422-1433.
- [12] POH M Z, MCDUFF D J, PICARD R W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation[J]. Optics Express, 2010, 18(10):10762-10774.
- [13] 雷恒波, 郭友达, 莫镜清, 等. 基于人脸视频的光电容积脉搏波信号获取与处理[J]. 中国医学物理学杂志, 2016, 33(11):1178-1182.
- LEI H B, GUO Y D, MO J Q, et al. Acquisition and processing of photoelectric volume pulse wave signal based on face video[J]. Chinese Journal of Medical Physics, 2016, 33(11):1178-1182 (in Chinese).
- [14] WU Z, HUANG N E. Ensemble empirical mode decomposition: A noise-assisted data analysis method[J]. Advances in Adaptive Data Analysis, 2009, 1(1):1-41.
- [15] ZHAO J, HUANG J F, JIA G Q, et al. Non-contact method of heart rate using a webcam[C]//IEEE International Conference on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data. Piscataway: IEEE Press, 2018:1902-1906.
- [16] FAVILLA R, ZUCCALÀ V C, COPPINI G. Heart rate and heart rate variability from single-channel video and ICA integration of multiple signals[J]. IEEE Journal of Biomedical and Health Informatics, 2019, 23(6):2398-2408.
- [17] LEI Y G, HE Z J, ZI Y Y. Application of the EEMD method to rotor fault diagnosis of rotating machinery[J]. Mechanical Systems and Signal Processing, 2009, 23(4):1327-1338.
- [18] ZHANG K P, ZHANG Z P, LI Z F, et al. Joint face detection and alignment using multitask cascaded convolutional networks[J]. IEEE Signal Processing Letters, 2016, 23(10):1499-1503.
- [19] NEUBECK A, VAN GOOL L. Efficient non-maximum suppression[C]//18th International Conference on Pattern Recognition. Piscataway: IEEE Press, 2006, 3:850-855.
- [20] KING D E. Dlib-ml: A machine learning toolkit[J]. Journal of Machine Learning Research, 2009, 10:1755-1758.
- [21] DE H G, JEANNE V. Robust pulse rate from chrominance-based rPPG[J]. IEEE Transactions on Biomedical Engineering, 2013, 60(10):2878-2886.
- [22] PANG B, LIU M, ZHANG X, et al. Advanced EMD method using variance characterization for PPG with motion artifact[C]//IEEE Biomedical Circuits and Systems Conference. Piscataway: IEEE Press, 2016:196-199.
- [23] KOELSTRA S, MUHL C, SOLEYMANI M, et al. Deap: A database for emotion analysis using physiological signals[J]. IEEE Transactions on Affective Computing, 2011, 3(1):18-31.
- [24] POH M Z, MCDUFF D J, PICARD R W. Advancements in non-contact, multiparameter physiological measurements using a webcam[J]. IEEE Transactions on Bio-Medical Engineering, 2011, 58(1):7-11.
- [25] TARASSENKO L, VILLARROEL M, GUAZZI A, et al. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models[J]. Physiological Measurement, 2014, 35(5):807-831.

作者简介:

陈辉 男, 硕士研究生。主要研究方向: 信号处理、机器学习。

郑秀娟 女, 博士, 副教授, 硕士生导师。主要研究方向: 信号信息处理、智能检测技术、自然人机交互。

倪宗军 男, 硕士研究生。主要研究方向: 图像处理、机器学习。

张昀 女, 博士, 副研究员。主要研究方向: 多模态数据分析。

杨晓梅 女, 博士, 副教授。主要研究方向: 图像重建、计算机视觉。

Vital signs detection via facial video analysis

CHEN Hui¹, ZHENG Xiujuan^{1,*}, NI Zongjun¹, ZHANG Yun², YANG Xiaomei¹

(1. College of Electrical Engineering, Sichuan University, Chengdu 610065, China;

2. School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: To detect the physiological signals related to vital signs via facial video is easily affected by ambient lights and head motions. In order to reduce the disturbance and increase the accuracy of estimations of vital signs, this paper proposes a facial video analysis method that combines Ensemble Empirical Mode Decomposition (EEMD) algorithm and signals quality detection to accurately detect vital signs such as the heart rate and respiratory rate of human beings. The performance of the proposed method is validated by comparing it with the existing signal processing techniques in a public dataset. The experimental results show that the proposed method can obtain more accurate estimates of heart rate and respiratory rate than the existing methods. The correlation coefficients between the estimates and the golden standards are higher than 0.9 and 0.8, respectively. The vital signs detection method has the potential to benefit real-time living face recognition and intelligent surveillance video analysis.

Keywords: facial video; imaging photoplethysmography; Ensemble Empirical Mode Decomposition (EEMD); signal quality detection; anti-interference

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0066



一种轻量化的多目标实时检测模型

邱博¹, 刘翔^{1,*}, 石蕴玉¹, 尚岩峰²

(1. 上海工程技术大学 电子电气工程学院, 上海 201620; 2. 公安部第三研究所 物联网技术研发中心, 上海 200031)

摘 要: 为实现公安监控系统内容分析的精准智能及提高服务实战能力, 提出一种轻量化的多目标实时检测算法。首先, 基于 CenterNet 检测网络增加了 CBN 的多融合阶梯级联结构, 有效地解决了主干网络在日常监控中特征提取能力不足的问题; 其次, 通过模型剪枝压缩网络减少参数量, 加快了监控视频分析速度。本文利用部分 COCO 数据集和自行采集的现场数据进行训练与测试, 并与其他主流检测算法 (YOLO、Faster-RCNN、SSD 等) 进行消融实验。实验结果表明: 所提模型在公共安全监控中能有效地做到速度与精度的均衡, 并具有较强的普适性。

关 键 词: 目标检测; 深度学习; 模型压缩; 模型蒸馏; 级联融合

中图分类号: TP391.4

文献标志码: A

文章编号: 1001-5965(2020)09-1778-08

视频监控系统作为当前城市公共安全的基础设施, 视频图像信息已成为支撑治安防控、维护社会稳定的重要手段。随着监控系统朝着智能化方向发展, 视频检测手段更加成熟、完善, 检测精度逐渐增高。行人与车辆作为公共安全监控过程中重要的检测目标, 精准性和实时性变得愈发重要。然而现实中复杂场景下, 尤其在车辆堵塞的地区、行人汇集地域, 行人与车辆检测的精度、速度受到了极大的影响。合理均衡模型处理速度、检测精度, 成为视频监控系统中亟待解决的问题。

目前, 针对目标检测算法主要有 2 个研究方向。一是利用传统的方法提取特征并进行检测, 使用的特征包括 HOG、SURF^[1]、SIFT^[2]、Haar^[3] 等。该类算法都是在候选区域内提取特征, 最后使用支持向量机^[4]、随机森林、自适应增强^[5] 等方法进行检测与识别。二是利用深度学习进行目标检测, 可以细分为二阶段与一阶段检测网络。二阶段模型中, Girshick 等基于 RCNN^[6], 利用 se-

lective search 方法输入整张图片生成 2 000 个候选区域, 随后将包含目标区域的图片送入网络中训练, 经过一系列卷积操作提取特征, 最后进行检测, 生成的候选区域中包含大量重叠部分, 导致 RCNN 模型训练速度慢, 处理效率低下。Girshick 提出的 Fast-RCNN^[7] 改进了 RCNN 先提取候选区域后训练的机制, 直接对整张图像进行训练, 模型将候选区域映射到最后一层特征图上, 该做法使得卷积层输出的候选区域经过 ROI Pooling 后得到固定的尺度, 该方法减少了大量的重复计算, 在训练速度、检测速度与训练空间上都进一步地减少。尽管如此, Fast-RCNN 也未能达到实时检测的性能。Ren 等提出的 Faster-RCNN^[8] 中, 取代了 selective search 提取候选区域的手段, 创造性地采用卷积网络 RPN 自行产生建议框, 并且和目标检测网络共享卷积网络, 极大地提升了候选框的质量, 进一步提升了检测的精准度。

虽然二阶段网络的检测精度不断地提高, 但

收稿日期: 2020-03-02; 录用日期: 2020-04-18; 网络出版时间: 2020-05-07 15:56

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200507.1114.002.html

基金项目: 国家重点研发计划 (2016YFC0801304); 上海市“科技创新行动计划”高新技术领域项目 (17511106803)

* 通信作者。E-mail: xliu@sues.edu.cn

引用格式: 邱博, 刘翔, 石蕴玉, 等. 一种轻量化的多目标实时检测模型[J]. 北京航空航天大学学报, 2020, 46(9): 1778-1785.
QIU B, LIU X, SHI Y Y, et al. A lightweight multi-target real-time detection model[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1778-1785 (in Chinese).

训练繁琐、检测速度不佳等情况时常发生。YOLO^[9]、SSD^[10]等算法的出现大大缩短了检测时间。YOLO 预定义预测区域,在每个指定区域内生成若干个不同大小、比例的潜在边界框,将候选区边界框尺度调整到更接近真实的目标尺度。SSD 则设计特征金字塔结构提取不同尺度特征,不同卷积块负责预测不同大小的物体。模型训练过程中采用大尺度特征图检测小物体,小尺度特征图检测大物体的策略。一阶段模型使得目标检测更加具有实时性,但精度却受到了不小的影响。

Anchor-free 是一种新型的无边框检测方法。网络不需要设定预选框来限定实际感受野,通过学习一些关键点的特征或者密集区域特征来实现检测目的。例如 CornerNet^[11]学习了真值框的左上与右下 2 个角点,利用内嵌向量分组进行角点匹配。ExtremeNet 在热力图上对每个目标预测 TOP 4 个点,并且预测中心点,通过几何方法进行分组。FoveaBox^[12]使用特征金字塔网络(FPN)作为主干网络,通过比例压缩真值的中心区域,形成中心区域的正样本集,对区域内每个像素预测分类与偏移量。CenterNet^[13]对真值热力图进行高斯核处理,网络学习高斯核后的真值的中心点坐标,预测偏移量与类别信息,并且在训练过程中摒弃了非极大值抑制手段,实现了真正意义上的端到端。

本文提出了一种基于 CenterNet 模型的轻量化行人车辆等多目标实时检测算法,考虑了复杂场景中目标的大小与目标距离的远近对模型精度造成的影响。同时对模型进行压缩、剪枝,在精度与速度方面进行权衡,实现了占用内存量低、精准性高、检测速度快的目的,适用于日常监控场景。

1 CenterNet 网络

CenterNet 是一种端到端的无边框目标检测算法。Zhou 等^[13]分别使用 ResNet-18^[14]、hourglass^[15]、dla34 作为网络主干。主干网络输出的特征图传入检测模块,该模块拥有 3 个分支,每个

分支都由 3×3 的卷积层与 1×1 的卷积层相连。模型简图如图 1 所示。

网络对输入图片进行高斯核处理,提升真值中心位置的响应热度。随后将热力图进行卷积操作,模型学习不同类别的峰值点坐标(目标的中心点)。每个类别仅有一个局部峰值点,此时局部峰值点作为正样本,其余为负样本。主干网络由若干个卷积层组成,主要功能是提取图像特征。与其他检测模型进行 32 倍下采样不同,CenterNet 只进行 4 倍下采样。使用更大分辨率的特征图作为主干网络的输出,能有效降低特征点的损失,避免特征过小导致后续网络无法学习有效特征。

1.1 损失函数

大部分检测网络学习真值框的位置,CenterNet 则学习真值中心点坐标,进一步降低了学习成本。Heatmap 分支输出形状为 (C, H, W) ,其中 C 为类别数, H 为输出图片的高度, W 为输出图片的宽度。类别中最高的响应分数即为该物体的类别。由于负样本数与正样本比例过大,因此采用 Focal-loss 函数平衡正负样本,计算公式为

$$L_k = \frac{1}{N} \sum_{xyc} \begin{cases} (1 - \hat{Y}_{xyc})^\alpha \lg(\hat{Y}_{xyc}) & Y_{xyc} = 1 \\ (1 - \hat{Y}_{xyc})^\beta (\hat{Y}_{xyc})^\alpha \lg(1 - \hat{Y}_{xyc}) & \text{其他} \end{cases} \quad (1)$$

式中: L_k 为中心点与分类损失函数,下标 k 为第 k 张输入图像; N 为图片中关键点的数量;下标 xyc 为图像中的正负样本; Y_{xyc} 为真值的标签; $\hat{Y}_{xyc}$ 为预测标签; α, β 为超参数。当 $Y_{xyc} = 1$ 时, $(1 - \hat{Y}_{xyc})^\alpha$ 充当了矫正作用。当 $\hat{Y}_{xyc} = 1$ 时,表示样本容易被分辨,此时应该减少 loss 比重,相应地对难分类的样本增加权重。当 Y_{xyc} 接近 1 时, $\hat{Y}_{xyc}$ 应该预测为 0。此时使用 $(\hat{Y}_{xyc})^\alpha$ 增加 loss 权重,加大训练。

但由于 Y_{xyc} 接近 1,于是利用 $(1 - \hat{Y}_{xyc})^\beta$ 进行补偿。当 Y_{xyc} 接近 0 时,则补偿与惩罚互相替换,起到双方制约的作用,弱化了实际中心点周围的

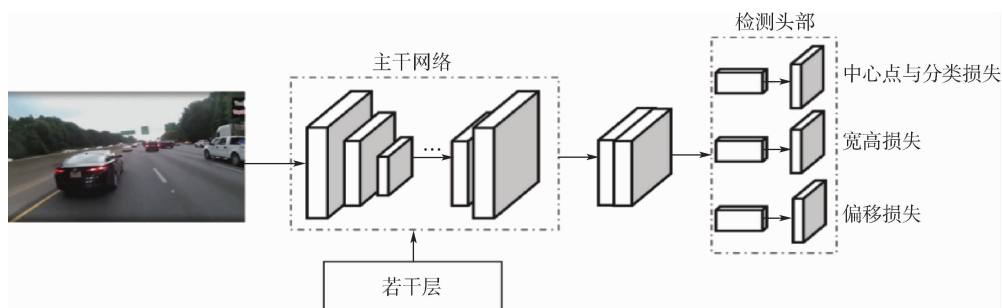


图 1 CenterNet 模型结构

Fig. 1 CenterNet model structure

负样本的比重,进一步处理正负样本的不平衡。

位置损失函数采用了 L1 loss;输入图片在主干网络中进行 4 倍下采样后,模型将特征图重新映射到原图时会发生像素偏移从而产生精度误差。目标中心位置损失函数为

$$L_{\text{off}} = \frac{1}{N} \sum_p \left| \hat{O}_p - \left(\frac{p}{R} - \hat{P} \right) \right| \quad (2)$$

式中: $\frac{p}{R} - \hat{P}$ 为实际的中心偏移量, p 为原图的目标真值的坐标, R 为下采样倍数, $\hat{P}$ 为模型预测的中心点坐标, $\hat{O}_p$ 为预测的偏移量。本文针对每一个中心点,都采用一个偏移量进行偏差补偿。

回归损失 L1:网络预测出中心点坐标,对每个目标尺度进行回归得出宽高,与位置损失一样使用 L1 loss,如下:

$$L_{\text{size}} = \frac{1}{N} \sum_{k=1}^N |\hat{S}_{pk} - S_k| \quad (3)$$

式中: L_{size} 为回归损失函数; S_k 为原图尺度; $\hat{S}_{pk}$ 为回归后的尺度。

3 类损失函数相加,乘以参数,整个训练过程的目标损失函数为

$$L_{\text{det}} = L_k + \lambda_{\text{size}} L_{\text{size}} + \lambda_{\text{off}} L_{\text{off}} \quad (4)$$

式中: λ_{size} 、 λ_{off} 为参数。

综上所述,网络模型对每个真值进行预测,输出 $(C+4)$ 维的矢量,分别对应物体的类别、宽高、中心偏差。

1.2 模型检测原理

模型检测时采用最大池化思想,对 8 联通区域的响应点数值进行比较。如果中心点响应大于其他位置的响应,则判断该点为峰值点,反之则舍弃。最后设置阈值,选取 TOP 100 个结果,作为目标的中心点坐标,如图 2 所示。另外两个分支根据目标中心坐标回归到边界框,调整回归损失。

2 轻量化多目标实时检测算法

2.1 级联融合层的设计

公安监控系统对固定区域进行实时监控和信息记录,主要在交通要道、公共聚集场所以及治安复杂场所安装视频监控设备。环境变化成为影响监控系统精度的主要原因。处于十字路口、商业区等复杂环境中,人流车流明显增多,此时摄像头不能根据行人远近、车辆大小调整拍摄角度、高度。行人的增加、重叠现象会加大主干网络特征提取的难度,造成漏检,同时车辆经常会对后方行人形成遮挡,此时模型会过分关注于大型车辆而忽略小目标。ResNet-18 对输入图片进行 8 倍下采样,两个重叠的行人经过下采样后处于同一个像素点中,难以提取到行人信息。随着网络深度加深,模型的表征能力增强,但相应也造成模型计算冗余。ResNet-101 层级太多,在公安监控系统中不能起到实时检测的效果。本文提出了一种基于 CBNet^[16] 级联差级交互的 hourglass 模型,主要针对主干网络(ResNet-18)做出如下修改:首先增加残差块与 hourglass 模块的错级级联结构。第一个残差块输出的特征与第一个 hourglass 的输入特征进行堆叠,第二个残差块的输出的特征与第一个 hourglass 的输出特征进行堆叠,以此类推。网络提取细条状的目标特征时,通过减少下采样的倍数来减少特征丢失,同时加深网络层数增强网络学习困难标签的能力。如果目标区域面积占比较大时,网络提取目标特征时,相应地增加下采样倍数,降低像素点过多对视觉造成的影响,将模型注意力转移到其他目标。本文提出的主干网络模型如图 3 所示。

输入图片尺度为 320×192 ,经过第 1 个卷积层与最大池化层后,图片分辨率已经缩小 4 倍。传统的模型通常进行 16 倍下采样。当行人重叠或目标过小时,经过下采样后,重叠行人往往聚集在同一个像素中。即使之后进行反卷积,特征也已被破坏,因此一般模型难以分辨目标重叠情况。为防止此现象的发生,本文新增了一个并行分支网络 ResNet-18。分支网络包含 4 个基本模块,每个基本模块都由 2 个残差块组成,残差块中不使用下采样,保证了尺度不变性。图片经过分支网络后,尺度变成 80×48 ,通道数增加至 256。

主干网络采用 hourglass 的漏斗状结构,对下采样的特征图双线性插值返回到原有尺度。碰到大目标时,该结构能够“转移”网络的注意力,防止对大型目标过多“关注”,从而产生过拟合现象。

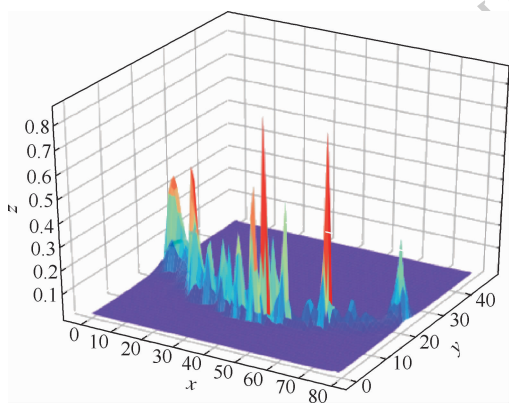


图 2 峰值响应可视化

Fig. 2 Peak response visualization

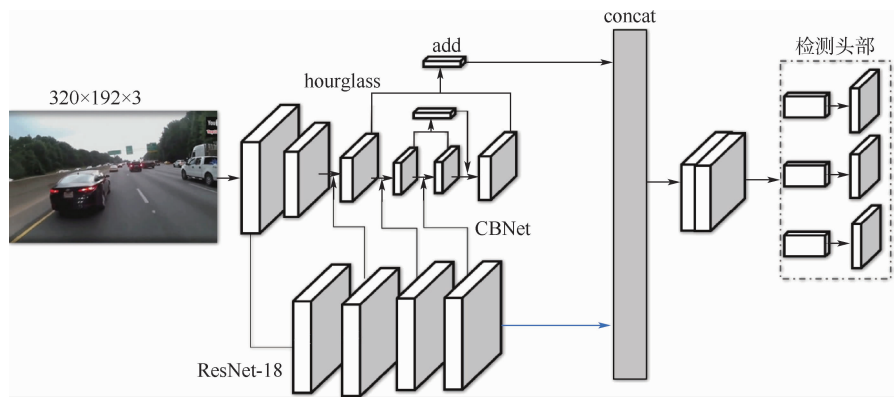


图 3 主干网络简图
Fig. 3 Backbone network illustration

本文提出的主干网络结构结合了 SSD^[10] 的思想:大尺度的特征图检测小的目标,小尺度的特征图检测大的目标。但是 SSD 并没有使用特征融合,仅仅在每层预设边界框的个数。将主干网络与分支网络最后输出的特征进行通道叠加,促使网络的通道响应更加灵敏。本文利用 CBNet 结构错级连接 ResNet-18 与 hourglass。分支之间的前一个残差块经过下采样与主干网络的后一个残差块进行差级相连,相连之前保证了两者输出尺度一致性,进而进行通道叠加。对不同尺度,不同层级的特征进行复用回传,以便获得更强的特征学习能力。传统的级联融合模型类似于 FPN 网络进行下采样后使用反卷积与相同尺度的特征进行融合。这种做法使得参数量快速增加,推理时间随之大幅度增加。使用 FPN 的形式去训练一个端到端的模型,增加了收敛难度。FPN 本身有利于多尺度的特征提取,但是分支网络需要保证尺度不变,因此主干网络与分支网络存在意义相悖。实验表明,CBNet 特征融合结构与原有模型所占内存的大小相近,并没有增加模型的参数量。在不破坏分支网络尺度不变性的前提下,融合主干网络特征与分支网络提取的特征后进行下采样。该级联方式让并行网络进行互补,对各自分支并无损失,提高了在训练速度与检测精度。

2.2 增加通道响应与模型加速

图像送入模型进行训练,4 倍下采样后密集目标尺度进一步缩小。不同类别的质心相互重合,热力图峰值差异缩小。网络在所有通道上进行特征提取,按照通道响应的大小来判断所提取的特征对应的物体种类。后续卷积层又将这些特征映射到后续的通道中。因此,主干网络需要增加通道数来减少通道响应重合的可能性。通道数的增加使得分类的准确率、召回率明显上升,但是导致模型参数量越来越大。针对该问题本文采取 2 种减少参数量方法:深度可分离卷积^[17]和模型压缩。

2.2.1 深度可分离卷积

深度可分离卷积^[17]是一种能有效减少参数的方法,其能将传统的卷积分解为一个深度卷积 (Depthwise Convolution) 和一个逐点卷积 (Pointwise Convolution),如图 4 所示, F 和 G 分别为输入特征图和输出特征图。

输入特征图的尺寸为 $D_f \times D_f \times M$, D_f 和 M 分别为特征图的边长和通道数量。传统卷积核尺寸为 $D_k \times D_k$, D_k 为卷积核的边长,经过 N 个卷积核,输出特征图尺寸为 $D_i \times D_i \times N$, N 为通道数量。同样是输出相同尺度的特征图,2 种卷积的参数量相差很大。传统卷积的计算量为 $D_k \times D_k \times M \times N \times D_i \times D_i$,深度可分离卷积的计算量为

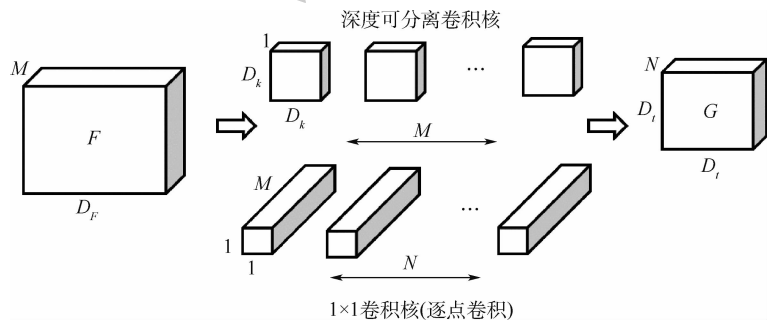


图 4 深度可分离卷积原理
Fig. 4 Depthwise separable convolution principle

$$D_k \times D_k \times M \times D_t \times D_t + M \times N \times D_t \times D_t。$$

从参数量看,2 个不同卷积的参数量之间相差一个数量级。从原理看,传统卷积直接对通道与区域同时进行卷积,深度可分离卷积首先对每个通道做区域卷积,再用的卷积核合并通道卷积。不仅提高了网络的检测速度,还实现了通道和区域的分离。

本文对分支网络的 4 个模块都进行了深度可分离卷积操作,实验表明模型的推理速度提升了 13%,较好地解决了模型因为参数量大而导致推理缓慢的问题。

2.2.2 模型压缩

目前深度学习的模型压缩方法主要分为以下 4 个方向:设计更加精致的主干网络、模型剪枝、核的稀疏化及模型量化。

本文考虑现有的 4 种方法优缺点,核的系数化与模型量化都需要平台上稀疏矩阵操作库的支持,对模型的框架、移植的系统都有严格的要求。优化主干网络^[18]提升模型效果最大,剪枝可以更好地降低模型的复杂度。因此本文选取了较为成熟的 2 种方法,优化主干网络和模型剪枝。Zhou 等^[13]使用 ResNet-18 作为主干网络,网络的深度和宽度较大,在训练过程中易造成参数冗余,且精度、召回率均不高。由于通道数量较大,上采样耗时严重,本文裁剪掉最后一层上采样。为了保证最后的输出尺度是 4 倍下采样,对应删除前面一层下采样的基础块。所有的 ResNet-18 支路保留原有尺度参数,而在 hourglass 支路中,保持下采样,并且在下采样过程中成倍增加通道数量,提高精度与召回率。

模型压缩方面主要进行了通道压缩^[19]与蒸馏操作。通道剪枝的核心思想是移除一些冗余的通道,同时裁剪与这些通道对应的卷积核,减少模型前向计算所需的 FLOPs。本文受到 L1 稀疏化的启发,对 hourglass 4 个卷积模块的通道增加缩放因子 γ ,训练后得到 4 个缩放系数 γ 。此时模型的检测精度、召回率损失最低。ResNet-18 分支则增加了空间注意力机制。利用全局均匀池化将上一层输出的值归一化作为权重系数,下一层特征图中的每个位置像素与该位置权重做乘积操作。其次对预训练的模型进行剪枝,分析每层的稀疏性对精度的影响,第二层卷积层稀疏性对精度影响如图 5 所示。当卷积层的稀疏性不断上升,分类准确性波动。稀疏性为 20% 时,准确性达到最低;当稀疏性接近 25%,精度最高。因此这个卷积层选择 25% 的稀疏性。

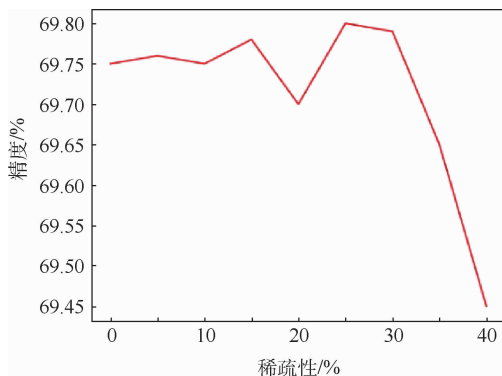


图 5 稀疏性与精度关系

Fig. 5 Relationship between sparsity and precision

2.3 IOU 协同过滤

CenterNet 中 Heatmap 分支利用最大池化机制预测出中心点坐标,再由损失函数分支预测的宽度、高度计算出边界框。

实际应用场景中,同一类别目标在空间尺度上的变化很大,针对如何设定阈值适应场景中目标尺度变化的问题,本文进行如下处理:首先,构建掩模矩阵,预测的边界框超出原图坐标则设置为 0,这样大大减少了边界框的个数。其次,对每个类别设置 IOU 协同过滤。根据 W/H 计算出边界框的面积。本文根据人眼的可见范围的对每个类别手动设置最小的面积阈值,即感兴趣区域的最小的边缘范围面积。当某个目标的面积小于该类别设定的阈值时,则自动忽略目标。划分检测区域过滤大目标,当目标超过图像原尺度或者预测框的坐标超过预设的检测区域将进行过滤。实际系统中经常会使用多摄像头协同工作,此时可以利用本文策略根据实际需求对各个摄像头划分检测区域。当检测目标的左上角、右下角坐标任意一点越过检测区域,及时过滤,确保单个目标能被最清晰地捕捉。最后,在热力图分支中对响应分数做非极大值抑制,加上可视化阈值,减少过小物体显示。

3 实验结果与分析

3.1 实验环境与数据集

本实验以 Pytorch 为框架,操作系统为 Windows7,实验环境为 4 块 Nvidia-1080Ti 显卡,batch-size 设置为 16,初始学习率为 5×10^{-4} ,训练 500 个 epochs。本文采用交叉训练方式,首先训练残差支路,增加网络结构并加载预训练权重训练 hourglass 结构。最后结合 CBN 模块进行训练,学习率降为 9×10^{-4} 。

本文从 COCO 数据集中挑选出 400 张车辆、

行人图片,结合来自广角监控摄像头的 1 600 张数据构成新的数据集。为了保证两者数据同分布,都选择了白天的街景。使用上下翻转,均值漂移等数据增强手段,设置 80 个种子随机进行数据增强,最后获得 12 000 张图片。部分图片如图 6 所示,(a)、(b)来自 COCO 数据集街景图,(c)、(d)来自监控摄像头街景图。



(a) COCO街景图一



(b) COCO街景图二



(c) 监控摄像头街景图一



(d) 监控摄像头街景图二

图 6 数据集部分图片

Fig.6 Partial pictures of dataset

3.2 消融实验

3.2.1 主干网络对比

本文使用 CenterNet 模型进行精度测试。尝试多个版本的主干网络,并与主流的框架 SSD、slim YOLOV3 进行实验对比。结果如表 1 所示。

与 SSD、slim YOLOV3 相比,CenterNet 本身并不具备优势。主干网络经过改进后,模型没有增

表 1 主流模型精度对比

Table 1 Mainstream models' precision comparison

模型	召回率	准确率	内存/KB
ResNet-18	0.83	0.80	25 014
ResNet-18 + hourglass + CBNet	0.92	0.90	24 254
ResNet-18 × 2 + CBNet	0.86	0.83	22 254
Hourglass × 2 + CBNet	0.85	0.83	23 280
SSD	0.87	0.84	23 020
slim YOLOV3	0.90	0.88	33 894

加参数量,但是精度、召回率均提高到了 90%。实验表明,CBNet 对 hourglass 与 ResNet 交互级联很好地提升了模型的学习能力。

3.2.2 模型压缩对比

本文对卷积层通道剪枝,按比例减少通道数。并且对每个卷积层进行蒸馏,测试稀疏性与准确性的关系。该方法在模型精度没有大幅下降的前提下有效地压缩了 80% 的内存,压缩前后的模型精度变化如表 2 所示。

表 2 模型压缩前后精度变化

Table 2 Model's accuracy before and after compression

模型	召回率	准确率	内存/KB
ResNet-18 + hourglass + CBNet	0.92	0.90	24 254
ResNet-18 + hourglass + CBNet(剪枝后)	0.90	0.88	3 676

3.2.3 模型推理速度对比

本文对压缩前后的模型进行性能分析,实验环境为单块 Nvida-1080Ti 显卡。实验结果如表 3 所示。模型重新设计以后,推理速度提升了 50%,证明了主干网络有助于模型加速。其次将部分模型的卷积模块替换成深度可分离卷积,模型速度提升了 13%。经过蒸馏与剪枝后模型推理速度大约为原模型速度的 3 倍,验证了本文压缩、加速方法的有效性。

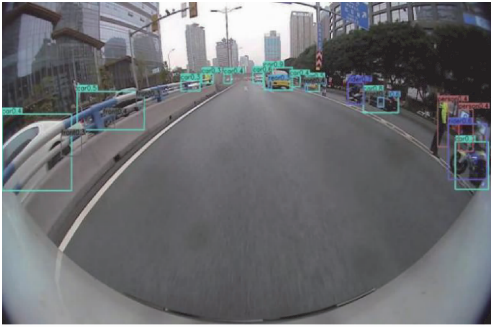
表 3 模型推理速度对比

Table 3 Model's inference speed comparison

模型	帧数/s
CenterNet(ResNet-18)	6.6
ResNet-18 + hourglass + CBNet	9.9
ResNet-18(深度可分离卷积) + hourglass + CBNet	11.2
ResNet-18 + hourglass + CBNet(剪枝后)	19.2

3.2.4 协同过滤可视化结果

由图 7 协同过滤对比可以看出,本文算法很好地排除了小目标的干扰。马路对面的车辆无需进行检测便可以排除,并且右侧靠近摄像头的人与车辆并没有漏检与误检。



(a) 推理过程中未进行过滤



(b) 推理过程中进行了面积与IOU的协同过滤

图7 协同过滤对比

Fig.7 Collaborative filtering comparison

4 结 论

本文提出了一个基于 CenterNet 模型的轻量化的多目标实时检测算法,使用了学习能力更强的 CBN 差级级联融合结构进行特征提取,并对原有模型进行压缩及蒸馏。

1) 模型准确率与召回率均可达到 90%,解决了复杂的交通场景下,对于行人车辆密集的区域检测精度过低的问题。

2) 模型压缩、蒸馏后,召回率下降 2%,准确率下降 2%,但内存占用量仅为原来的 15%,增强了模型的普适性。

3) 模型进行重新设计与加速,弥补了原始模型推理过程中速度过慢的缺陷,使得检测具有实时性,极大地提高了监控服务实时作战能力。

参考文献 (References)

[1] KALIA R,LEE K D,SAMIR B V R,et al. An analysis of the effect of different image preprocessing techniques on the performance of SURF;Speeded up robust features[C]// Workshop on Frontiers of Computer Vision. Piscataway:IEEE Press,2011: 1-6.

[2] LOWE D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision,2004,60 (2):91-110.

[3] MUNRO S,THOMAS K L,ABU-SHAAR M. Molecular characterization of a peripheral receptor for cannabinoids[J]. Nature,

1993,365(6441):61-65.

[4] PLATT J C. A fast algorithm for training support vector machines[J]. Journal of Information Technology,1998,2(5): 1-28.

[5] FREUND Y,SCHAPIRE R E. A decision-theoretic generalization of on-line learning and an application to boosting[C]// Proceedings of the 2nd European Conference on Computational Learning Theory. Berlin:Springer,1995:22-37.

[6] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2014: 580-587.

[7] GIRSHICK R. Fast-RCNN[C]// Proceedings of 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press,2015:10-15.

[8] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:Towards real-time object detection with region proposal networks[C]// Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: MIT Press,2015: 1-15.

[9] REDMON J,FARHADI A. YOLO9000: Better, faster, stronger [EB/OL]. (2016-12-25) [2020-02-27]. <https://arxiv.org/abs/1612.08242>.

[10] LIU W,ANGUELOV D,ERHAN D,et al. SSD: Single shot multibox detector[C]// Proceedings of 2016 European Conference on Computer Vision and Pattern Recognition. Berlin: Springer,2016:13-17.

[11] LAW H,DENG J. CornerNet: Detecting objects as paired keypoints[J]. International Journal of Computer Vision,2018, 128:642-656.

[12] KONG T,SUN F,LIU H,et al. FoveaBox: Beyond anchor-based object detector[EB/OL]. (2019-04-08) [2020-02-27]. <https://arxiv.org/abs/1904.03797>.

[13] ZHOU X,WANG D,KRÄHENBÜHL P. Objects as points[EB/OL]. (2019-04-16) [2020-02-27]. <https://arxiv.org/abs/1904.07850>.

[14] HE K M,ZHANG X Y,REN S Q. Deep residual learning for image recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2016:770-778.

[15] NEWELL A,YANG K,JIA D. Stacked hourglass networks for human pose estimation[EB/OL]. (2016-03-22) [2020-02-27]. <https://arxiv.org/abs/1603.06937>.

[16] LIU Y,WANG Y,WANG S,et al. CBNNet: A novel composite backbone network architecture for object detection[EB/OL]. (2019-09-09) [2020-02-27]. <https://arxiv.org/abs/1909.03625>.

[17] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017:1800-1807.

[18] TAN M,PANG R,LE Q V. EfficientDet: Scalable and efficient object detection[EB/OL]. (2019-11-20) [2020-02-27]. <https://arxiv.org/abs/1911.09070>.

[19] HE Y,ZHANG X Y,SUN J. Channel pruning for accelerating very deep neural networks[EB/OL]. (2017-08-21)[2020-02-27]. <https://arxiv.org/abs/1707.06168>.

作者简介:

邱博 男,硕士研究生。主要研究方向:计算机视觉、深度学习、目标检测。

刘翔 男,博士,副教授,硕士生导师。主要研究方向:计算机视觉及人工生命。

石蕴玉 女,博士,讲师,硕士生导师。主要研究方向:视频大数据智能分析。

尚岩峰 男,博士,副研究员。主要研究方向:模式识别和视频大数据。

A lightweight multi-target real-time detection model

QIU Bo¹, LIU Xiang^{1,*}, SHI Yunyu¹, SHANG Yanfeng²

(1. School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;

2. Internet of Things Technology R & D Center, The Third Research Institute of the Ministry of Public Security, Shanghai 200031, China)

Abstract: For the public security monitoring system, a lightweight multi-target real-time detection algorithm is proposed in order to realize the accurate intelligence of the content analysis and improve the actual service ability. First, the multi-fusion gradient cascade structure of CBNet is added based on CenterNet detection network, which effectively solves the problem of insufficient feature extraction capability of the backbone network in daily monitoring videos. Second, the number of parameters is reduced through the model pruning and compression, which can speed up the analysis speed of monitoring videos. During the experiments, the dataset for training and testing consists of a part of COCO datasets and a number of field data collected by ourselves. The ablation experiments are conducted with other mainstream detection algorithms (YOLO, Faster-RCNN, SSD, etc.). The experimental results show that the presented model can effectively balance the speed and precision in the analysis of monitoring videos for public security and has stronger universality.

Keywords: target detection; deep learning; model compression; model distillation; cascade fusion

Received: 2020-03-02; **Accepted:** 2020-04-18; **Published online:** 2020-05-07 15:56

URL: kns.cnki.net/kcms/detail/11.2625.V.20200507.1114.002.html

Foundation items: National Key R & D Program of China (2016YFC0801304); Shanghai Science and Technology Innovation Action Plan in Hi-tech Field (17511106803)

* **Corresponding author.** E-mail: xliu@sues.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0069

基于跨尺度特征聚合网络的多尺度行人检测

曹帅, 张晓伟*, 马健伟

(青岛大学 计算机科学技术学院, 青岛 266071)



摘 要: 行人的空间尺度差异是影响行人检测性能的主要瓶颈之一。针对这一问题,提出了跨尺度特征聚合网络(TS-FAN)有效检测多尺度行人。首先,鉴于不同尺度空间呈现出的特征差异性,引入一种基于多路径区域建议网络(RPN)的尺度补偿策略,其在多尺度卷积特征层上自适应地生成一系列与其感受野大小相对应的候选目标尺度集。其次,考虑到不同层次卷积特征在视觉语义上的互补性,提出了跨尺度特征聚合网络模块,其通过横向连接、自上而下路径和由底向上路径,有效地聚合具有语义鲁棒性的高层特征和具有精确定位信息的低层特征,实现对卷积层特征的增强表示。最后,联合多路径 RPN 尺度补偿策略和跨尺度特征聚合网络模块,构建了一种尺度自适应感知的多尺度行人检测网络。实验结果表明,所提方法与当前一流的行人检测方法 TLL-TFA 相比,在整个 Caltech 公开测试数据集上(All:行人高度大于 20 像素)的行人漏检率降低到 26.21% (提高了 11.94%),尤其对于 Caltech 小尺寸行人子数据集上(Far:行人高度在 20~30 像素之间)的行人漏检率降低到 47.30% (提高了 12.79%),同时在尺度变化剧烈的 ETH 数据集上的效果也取得显著提升。

关键词: 行人检测; 尺度感知; 特征金字塔; 特征聚合; 非极大值抑制

中图分类号: V221+.3; TB553

文献标志码: A

文章编号: 1001-5965(2020)09-1786-11

随着人工智能与计算机视觉技术的发展以及人们对公共安全领域的日益重视,智能视频监控技术已经发展为当前的研究热点。行人检测是智能化视频监控系统中的核心关键技术之一,并为后续的更高层次的视频分析与理解提供可靠的数据支持。因此,行人检测技术直接影响着智能视频监控的智能化水平,具有重要的研究意义和应用价值。

近年来,基于深度卷积神经网络的通用目标检测^[1-3]取得了巨大的成功,尤其是 Girshick 等提出的基于 R-CNN^[4]目标检测模型,代替了手工设计特征^[5-6]极大地提高了行人检测的性能。然而,在视频监控系统中行人存在不同空间尺度的差异,尤其是大尺度、低分辨率的行人目标对行人

检测技术带来了极大的挑战。为解决行人在空间尺度变化下的检测问题,目前主要分为 2 种策略:图像金字塔^[7-8]和特征金字塔^[2,9-10]。基于图像金字塔的多尺度行人检测方法通过采样输入图像得到不同尺度的图像金字塔集合,以预测最终的检测结果。其中,具有代表性的图像金字塔的尺度归一化网络(SNIP)^[7]及其加强版具有高效重采样的图像金字塔的尺度归一化网络(SNIPER)^[8],其选取若干个正样本区域和负样本区域作为图像金字塔。然而,这种基于图像金字塔的行人检测方法在时间和内存方面的消耗是巨大的,限制了在实时智能监控视频任务中的应用。

为兼顾精度与速度,基于特征金字塔的多尺度方法被广泛应用到目标检测。与图像金字塔相

收稿日期: 2020-03-02; 录用日期: 2020-04-09; 网络出版时间: 2020-04-17 13:01

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200417.0944.001.html

基金项目: 国家自然科学基金(61902204); 山东省自然科学基金(ZR2019BF028)

* 通信作者. E-mail: xiaowei19870119@sina.com

引用格式: 曹帅, 张晓伟, 马健伟. 基于跨尺度特征聚合网络的多尺度行人检测[J]. 北京航空航天大学学报, 2020, 46(9): 1786-1796. CAO S, ZHANG X W, MA J W. Trans-scale feature aggregation network for multiscale pedestrian detection[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1786-1796 (in Chinese).

比,特征金字塔需要的内存空间和计算量要少的多,并且结构简单,能够有效地嵌入到各类目标检测器中。其中,单镜头多盒检测器(SSD)^[2]就是基于视觉几何组网络(VGG-16)提取不同分辨率的多层特征图集合用于多尺度目标检测。感受野块网络(RFB-Net)^[9]在SSD网络的基础上采用2个模拟人类视觉感受野的特征提取模块替换原有卷积层,并使用6个不同层级的特征图用于检测。类似的,渐进定位网络(ALFNet)^[10]在ResNet-50^[11]后3个阶段的最后一个卷积层以及新添加卷积层上堆叠多个目标分类和空间位置坐标回归模块,形成渐进定位网络,从而实现对多尺度目标精确定位。

大量研究表明^[12-13],通过添加横向连接和自上而下信息传播路径进行特征融合,生成的特征表达能力更强。特征金字塔网络(FPN)^[14]通过自上而下的跨层路径有效地融合了高层特征所具有的鲁棒语义信息生成特征金字塔网络。路径聚合网络(PANet)^[15]基于FPN网络结构添加自底向上的扩展路径,以精确的定位信息增强整个特征金字塔表达能力,有效提高了目标分割的准确度。多级特征金字塔目标检测器(M2Det)^[16]提出了多层次的特征金字塔网络,使用多层次重复的网络结构生成具有更强表达能力的特征金字塔。受上述研究工作的启发,为充分利用不同尺度特征层在视觉语义信息上的互补性,本文提出了跨尺度特征聚合网络(TS-FAN)模块,在几乎没有增加任何时间耗费的前提下实现了不同层次特征信息的聚合,以增强特征金字塔的语义鲁棒性和定位精确性。

另一方面,为检测视频图像中的不同尺度目标,Faster R-CNN^[1]基于多尺度区域建议网络(RPN)生成多尺度初始候选目标区域检测多尺度目标。进而FPN利用多个RPN子网络来尽可能的覆盖图像中目标的所有尺度,但这种方式忽略了不同路径RPN网络生成的候选目标集存在行人分类与定位回归之间的不一致性,从而影响多尺度行人的检测性能。为解决这一问题,本文引入一种基于多路径RPN的尺度补偿策略来有效处理多尺度目标,以提高不同尺度行人的召回率。受尺度自适应的三叉戟网络(TridentNet)^[17]多分支检测的启发,为避免极端尺度行人对不同路径模型训练影响,本文利用尺度感知的策略使不同路径RPN检测与特征层感受野相匹配的多尺度行人候选框,并将不同路径RPN生成的候选目标区域单独处理,避免了不同路径下不同候选

目标相互之间产生的影响。同时,根据Li等^[18]的研究,不同尺度行人实例在不同特征层上具有不同的特征表达,因此本文对不同尺度行人使用不同的特征映射函数,为多路径RPN生成的不同尺度候选目标区域集匹配相适应的聚合特征层,形成多尺度行人检测网络。

综上所述,本文主要贡献如下:

1) 引入一种基于多路径RPN的尺度补偿策略,依据不同分辨率行人实例构建多路径RPN网络,使各路径RPN网络分支基于有效感受野大小自适应地生成候选目标尺度集,以提高多尺度目标的召回率,并通过非极大值抑制的方法得到多尺度的目标候选区域集合。

2) 根据不同分辨率特征层在视觉语义信息和精确定位信息上的差异性,本文提出了TS-FAN模块,通过跨层连接聚合多尺度特征信息,极大地缩短了底层特征信息传播到顶层特征层的路径,增强了特征金字塔的语义鲁棒性和定位精确性。

3) 基于尺度感知的端到端训练方案,将多路径RPN中得到的不同尺度候选目标集映射到与之匹配的聚合特征层中进行特征提取,形成多尺度行人检测网络。实验结果表明,在Caltech^[19]和ETH^[20]数据集上明显优于目前一流行人检测方法TLL-TFA^[21],尤其对大尺寸、低分辨率行人的检测性能提升较为显著。

1 TS-FAN网络结构

TS-FAN总体结构如图1所示,主要包含3个部分:基于尺度补偿策略的多路径RPN、TS-FAN模块和多尺度行人检测网络。TS-FAN网络模型基于端到端的训练方式联合不同路径RPN子网络和TS-FAN模块,通过多路径RPN产生得到的行人候选区域自适应感知其在相应特征聚合网络模块上的有效特征,并采用尺度感知的策略形成多尺度行人检测网络。图中: $C_1 \sim C_5$ 分别为ResNet-50的5个不同阶段; $H_3 \sim H_5$ 分别为不同分辨率的聚合特征。

1.1 基于尺度补偿策略的多路径RPN

RPN在Faster R-CNN^[1]中被提出,因其引入了多尺度滑动窗口遍历特征图的每个空间位置,极大地提高了目标检测的召回率。然而,RPN只在某一深度卷积特征层上提取候选目标,其固定尺寸的卷积核限制了单一特征层的视觉感受野大小。对此,FPN^[14]在多个特征层上生成多尺度候选目标,进一步提升了目标检测的召回率。在此

基础上,本文开展了对多路径 RPN 行人召回率的实验分析,发现不同深度卷积特征层对不同尺度行人候选目标召回率具有较大的性能差异。大尺寸行人在高层特征图具有较高的召回率,而小尺寸行人在分辨率高的低层特征具有较高的召回

率。为此,根据各深度卷积特征层的有效感受野大小^[22],本文采用尺度补偿策略,将行人候选目标划分为 3 个路径的 RPN 来适应行人的多尺度变化,如图 2 所示。其中左、右分支作为辅助检测网络,中间分支则为主检测网络。

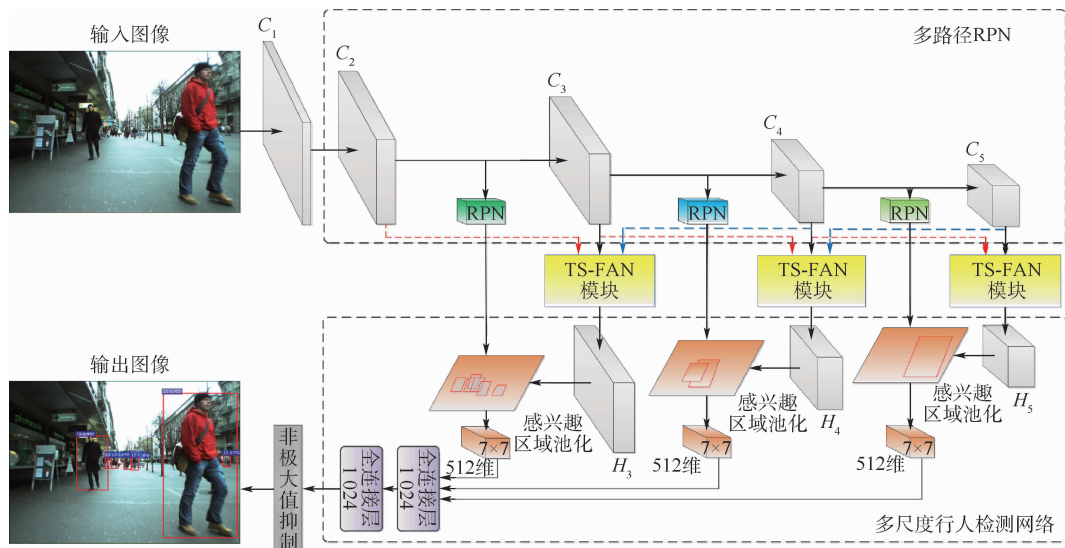


图 1 TS-FAN 总体网络架构

Fig. 1 TS-FAN overall network architecture

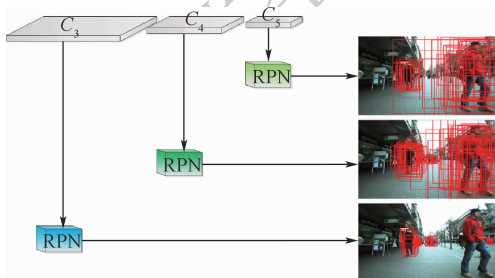


图 2 多路径 RPN

Fig. 2 Multipath region proposal network

本文使用 ResNet-50 作为特征提取基础网络,定义 C_3 、 C_4 、 C_5 代表基础网络结构中每个阶段的最后一个残差块 res3d、res4f、res5c。不同分支 RPN 中设置有效真实标注框的高度(行人实例高度像素值)分别在小于 50 像素、所有像素、大于 100 像素范围内,跨越该范围的真实标注视为无效标注,不参与该 RPN 分支训练。由于每个 RPN 路径针对不同尺度的行人目标进行训练,所以不同 RPN 路径使用独立损失函数,其中 RPN 多任务损失函数定义为

$$L = l_{cls} + \phi[y = 1]l_{loc} \quad (1)$$

式中: l_{cls} 为分类损失采用交叉熵损失函数^[1]; l_{loc} 为位置回归损失采用 Smooth-L1 损失函数^[1]; ϕ 为一个超参数; $y = 1$ 表示只有正样本进行位置回归。基于单个 RPN 损失函数,给出总体损失函数,其定义为

$$L_{rpn} = L_1 + L_2 + L_3 \quad (2)$$

式中: L_1 、 L_2 、 L_3 分别为左、中、右分支的多任务损失函数。

基于上述多路径 RPN 得到具有不同尺度范围的候选区域集 $P = \{P_s, P_a, P_l\}$, 其中 P_s 和 P_l 分别为小尺度集和大尺度集,它们是对所有尺度集 P_a 的尺度补偿。对于上述候选区域集使用阈值为 0.7 的非极大值抑制减少重叠候选目标框,为目标识别阶段提供高质量的候选区域。

1.2 跨尺度特征聚合网络

特征金字塔被广泛应用到多尺度检测的模型中,如图 3 所示,SSD^[2] 和 STDN^[23] 网络都是基于自底向上的信息传播方式生成不同空间分辨率的特征金字塔。然而这种方法没有考虑到不同层次信息的互补性^[24-26], TLL-TFA^[21]、CSP^[27] 等利用反卷积等上采样操作将不同层次的特征层归一化到同一分辨率,并通过特征通道叠加的方式进行特征融合。FPN^[14] 和 PANet^[15] 模型构建了自上而下和自底向上的信息传播路径,融合不同层次特征层作为目标检测特征层。而 M2Det^[16] 提出了多层次的 FPN,使用更加复杂的网络结构生成具有更强表达能力的特征金字塔。然而这些方法都是通过复杂的网络构建更多的特征金字塔,从而得到更加鲁棒的特征表示,其忽略了低层次特征的重要性,丢失大量的

细节特征信息。

本文提出的 TS-FAN 模块是在 FPN 网络模型的基础上通过添加自底向上快速路径,缩短低层次高分辨率特征图到高层次特征图的传播路径,以有效聚合低层特征图中的局部细节特征信息。本文在自底向上的特征编码路径引入平均池化层,来丰富用于检测特征层的特征信息,实现不同尺度卷积特征的增强表示。

本文提出的 TS-FAN 模块有效地融合了自顶向下、由底向上和同层映射三种路径特征,如图 4 所示。在特征融合之前,首先使用 1×1 卷积核对当前特征层 C_i 和上、下相邻特征层 C_{i+1} 、 C_{i-1} ($i \in \{3, 4, 5\}$) 实现特征维度的统一,得到空间分辨率不同但是特征通道数相同的特征层 C'_{i+1} 、 C'_i 、 C'_{i-1} 。在自上而下的特征传播路径中,使用双线性插值的上采样方法将 C'_{i+1} 特征层的空间分辨率扩大到原来的 2 倍,并且保持特征维度不变,保留其高层特征图中较为鲁棒的语义特征信息。另一

方面,为保留低层有利于目标定位较为敏感的局部位置信息,在自底向上的特征增强路径中,采用平均池化方法对 C'_{i-1} 层特征层下采样缩放至原来一半的空间分辨率大小,并且不改变其特征维度,保留其低层特征图中较为精确的定位信息。特征聚合通过对特征图逐像素相加的方式实现,其能够增加特征的信息量,但特征维度本身没有增加,这对于最终的图像分类是有益的。最后为减少上采样过程中的混叠效应,添加了一个 3×3 卷积处理融合后的特征图生成最终的具有强表达能力的特征图。TS-FAN 模块的计算公式为

$$H_i = \sum_{i=1}^c (\text{Avgpooling}(C'_{i-1}) + \text{Upsampling}(C'_{i+1}) + C'_i) * K_i \quad (3)$$

式中: c 为特征通道维度; K_i 为 3×3 卷积核;“ $*$ ”为卷积操作;Avgpooling 为平均池化操作;Upsampling 为上采样操作; H_i 为 TS-FAN 模块得到的增强特征表示。

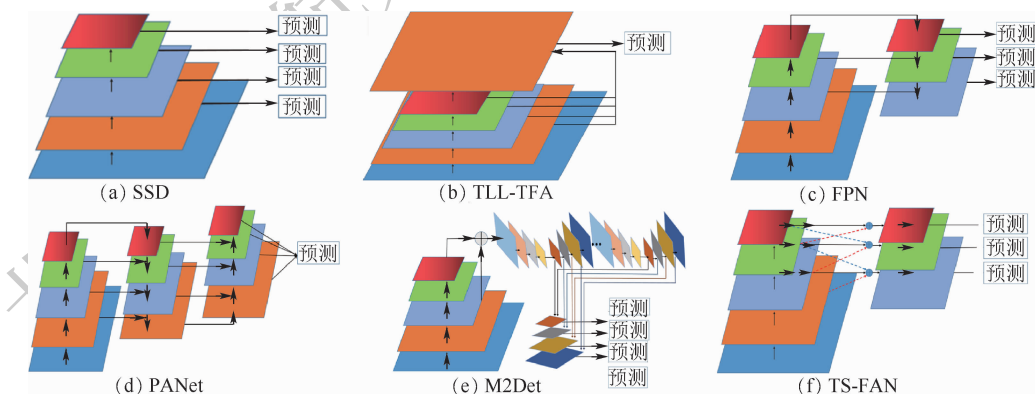


图 3 多种特征金字塔模型示意图

Fig. 3 Schematic diagram of multiple feature pyramid models

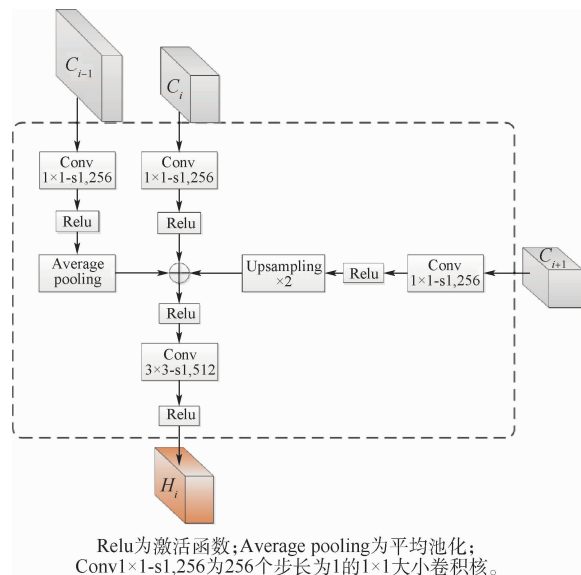


图 4 特征聚合模块

Fig. 4 Feature aggregation module

1.3 多尺度行人检测网络

本文根据不同分辨率的特征层对于不同尺度行人的有效性,通过多尺度检测方法联合多路径 RPN 生成的多尺度行人候选集 $P_i = \{P_s, P_a, P_l\}$ 和 TS-FAN 模块得到的聚合特征 $H_i = \{H_3, H_4, H_5\}$ 提取候选区域特征编码。如多路径 RPN 中的主检测分支生成 P_a 集合中的行人候选区域匹配到相应的 TS-FAN 生成的聚合特征 H_4 ,从而得到该特征层的感兴趣区域,利用 RoI-pooling 归一化提取的特征编码得到 $7 \times 7 \times 512$ 特征,将提取的特征编码由全连接层变换到 1 024 维高维特征向量,精确计算候选区域的置信度分数和 4 个坐标偏移量,得到最终的检测结果,其他 2 个辅助检测分支类似。对于不同尺度集的候选区域使用对应的检测分支,每个检测分支训练都有真实类别标注 p^* 和真实标注框 $b^* = (b_x^*, b_y^*, b_w^*, b_h^*)$, b_x^* 、

b_y^* 、 b_w^* 和 b_h^* 分别为真实标注框的左下角坐标和宽、高。本文单分支行人检测训练的损失函数定义如下:

$$L_{\text{det}}(p, b) = L_{\text{cls}}(p, p^*) + \lambda p^* L_{\text{reg}}(b, b^*) \quad (4)$$

式中: L_{cls} 为分类交叉损失函数; L_{reg} 为候选目标的回归损失函数, $L_{\text{reg}}(b, b^*) = R(b - b^*)$, R 为 Smooth-L1 损失函数; p 和 $b = (b_x, b_y, b_w, b_h)$ 为网络候选目标框置信度分数和空间位置; λ 为平衡分类与回归任务的损失函数, 本文中 $\lambda = 10$ 。预测得到的候选目标框与任何一个真实标注框的重叠度大于 0.5 时 $p^* = 1$, 否则 $p^* = 0$ 。多尺度行人检测的具体实现过程如算法 1 所示。

算法 1 基于 TS-FAN 的多尺度行人检测。

输入: 训练样本 $U = \{(p_i^*, b_i^*)\}_{i=1}^N$, 迭代次数 K 。

输出: 网络模型权重 ω 和 ω_b 。

初始化: 加载 ImageNet 数据集上的预训练权重, 为新添加的卷积层使用高斯函数初始化权重。设定学习率 ζ_ω 和 ζ_{ω_b} 。

迭代循环:

for $t = 1$ to T **do**

随机样本 mini-batches: B

for mini-batch **do**

基于尺度补偿的多路径 RPN:

$$L_{\text{far}} = l_{\text{cls}} + \phi[y = 1]l_{\text{loc}}, L_{\text{medium}} = l_{\text{cls}} +$$

$$\phi[y = 1]l_{\text{loc}} \quad b_h^* \leq 50$$

$$L_{\text{near}} = l_{\text{cls}} + \phi[y = 1]l_{\text{loc}} \quad b_h^* \geq 100$$

多路径 RPN 输出结果: $P_i = \{P_s, P_a, P_l\}$

TS-FAN 模块:

$$H_i = \sum_{i=1}^c (\text{Avapooling}(C'_{i-1}) + \text{Upsampling}(C'_{i+1}) + C'_i) * K_i$$

TS-FAN 模块输出:

$$H_i = \{H_3, H_4, H_5\}$$

多尺度行人检测网络:

$$\begin{cases} L_{\text{far}}(p, b) = L_{\text{cls}}(p, p^*) + \lambda p^* L_{\text{reg}}(b, b^*) \\ L_{\text{medium}}(p, b) = L_{\text{cls}}(p, p^*) + \lambda p^* L_{\text{reg}}(b, b^*) \\ L_{\text{near}}(p, b) = L_{\text{cls}}(p, p^*) + \lambda p^* L_{\text{reg}}(b, b^*) \end{cases}$$

使用随机梯度下降(SGD)优化方法更新

ω 和 ω_b :

$$\omega \leftarrow \text{SGD}(\nabla \omega(L_{\text{cls}} + \lambda p^* L_{\text{reg}}), \omega, \zeta_\omega)$$

$$\omega_b \leftarrow \text{SGD}(\nabla \omega_b(L_{\text{cls}} + \lambda p^* L_{\text{reg}}), \omega_b, \zeta_{\omega_b})$$

end for

end for

输出: ω 和 ω_b 。

其中: t 和 T 分别为当前迭代次数和总迭代次数; L_{far} 、 L_{medium} 、 L_{near} 分别为大尺寸、中尺寸、小尺寸分支路径的损失函数; ζ_ω 和 ζ_{ω_b} 分别为权重的学习率和偏置项的学习率。

2 实验与分析

2.1 数据集与实验设置

本节在 2 个公开基准数据集 Caltech 和 ETH 上测试本文的 TS-FAN 方法对多尺度行人检测的有效性。本文基于 Caltech 评估标准^[19]: 平均每幅图像假阳性 (FPPI) 在 $[10^{-2}, 10^0]$ 之间的行人漏检率, 用 MR^{-2} 表示。依据 Caltech 测试集数据划分标准^[19], 划分为: Reasonable 子集(行人高度大于 50 像素和可见度在 65% 以上); All 子集(高度最小为 20 像素和可见度在 20% 以上); Large、Near、Medium 和 Far 子集分别表示行人高度范围为大于 100 像素、大于 80 像素、30 ~ 80 像素之间和 20 ~ 30 像素之间的测试子集。设置学习率为 0.001、权重衰减为 0.0005、梯度更新权重为 0.9, 在单 GPU 上每个 mini-batch 使用 2 张图片, 选择使用 SGD 优化器。实验所使用环境为 Ubuntu14.0、caffe2、CUDA8.0.61、python2.7.12, 硬件配置为 NVIDIA GeForce GTX 1080Ti(一块)、Intel(R) Xeon(R) CPU E5-2609v4@1.70 GHz × 16。

2.2 消融实验

2.2.1 RPN 尺度补偿策略的重要性

为验证基于尺度补偿策略的多路径 RPN 对多尺度行人候选目标生成的有效性, 本文在 Caltech 数据集上通过 RPN 获取 300 个目标候选框, 以评估行人检测的召回率(表示为 R^{300})。本实验设置预测目标框与真实标注框重叠度阈值大于 0.5 即为判断为正样本, 否则为负样本。

首先, 在 ResNet-50 的不同层次特征层 (C_3 、 C_4 、 C_5) 中引入 RPN, P_{34} 表示为 FPN 网络中联合使用 P_3 、 P_4 特征层, C_{34} 表示为联合使用 C_3 、 C_4 特征层, 其他依次类推。从表 1 数据可以看出, 小尺寸的行人在高分辨率的特征层上具有较高的召回率, 如 C_3 要比 C_5 表现得更好。而在 C_4 层, 该层能够更好地兼顾不同尺度行人实例, 对于多尺度检测表现出良好的效果, 但对于小尺寸行人实例召回率仅为 75.2%, 这意味着单卷积层的 RPN 并不能有效覆盖图像中行人实例的所有尺度。本文引入多路径 RPN 尺度补偿策略提取多尺度目标候选框, 在整个 Caltech 多尺度行人集合上取得了 97.2% 的行人召回率。而且从表 1 中还可以看到, 在卷积特征层上联合多路径 RPN 生成行人候

表 1 在 Caltech 数据集上对于 RPN 的消融实验

Table 1 Ablation experiment of RPN on Caltech dataset

特征层	$R^{300}/\%$			
	All 子集	Far 子集	Medium 子集	Near 子集
C_3	87.7	71.5	90.6	91.9
C_4	92.8	75.2	95.9	97.7
C_5	82.4	59.7	85.4	95.2
P_{34}	95.5	89.1	96.8	97.3
C_{34}	95.3	93.7	95.7	97.9
P_{45}	92.9	76.2	96.3	97.3
C_{45}	93.3	77.3	97.7	97.9
P_{345}	93.7	91.1	94.5	93.4
C_{345}	97.2	93.7	97.7	97.9

选目标比在 FPN 中更为有效,其原因可归结为经过卷积后的特征层比融合后的 FPN 特征含有更多的局部细节信息。

2.2.2 跨尺度聚合特征对于行人检测的有效性

为验证本文 TS-FAN 模块对行人检测的有效性,本节将其与 FPN 的行人检测结果进行了实验对比。表 2 中的 Proposal 为单路径 RPN 的输入,如 FPN-P3 和 TS-FAN-H3 分别表示 FPN 网络和 TS-FAN 网络在 ResNet-50 第 3 阶段检测,其余类似,TS-FAN-H3H4H5 表示多分支检测。从表 2 中可以观察到,TS-FAN 模块 TS-FAN-H3 和 TS-FAN-H4 相较于 FPN 的 FPN-P3 和 FPN-P4 在 Caltech 数据集上均有明显的检测性能提升。尤其是 TS-FAN-H3 比 FPN-P3 在 Caltech 的 Reasonable、Near、Medium 子集上的行人漏检率 MR^{-2} 降低了 17.45%、28%、11.25%,TS-FAN-H4 相对于 FPN-P4 在 Far 子集上的行人漏检率 MR^{-2} 提升了 9.91%。这可归因于聚合低层次特征的细节信息有利于提升行人的检测效果。

表 2 Caltech 数据集上验证跨尺度聚合特征的有效性

Table 2 Verification of validity of trans-scale aggregation features on Caltech dataset

方法	Proposal	$MR^{-2}/\%$			
		Reasonable 子集	Near 子集	Medium 子集	Far 子集
FPN-P3	C_3	31.29	43.31	31.75	54.06
TS-FAN-H3	C_3	13.84	15.31	20.50	52.80
FPN-P4	C_4	5.33	0.72	24.65	75.41
TS-FAN-H4	C_4	5.12	0.47	20.08	65.50
FPN-P5	C_5	28.45	2.05	75.82	100.00
TS-FAN-H5	C_5	37.96	1.97	82.73	100.00
TS-FAN-H3H4H5	C_4	6.16	1.57	17.24	50.38
TS-FAN-H3H4H5	C_{345}	5.53	0.47	13.76	47.30

值得关注的是在 FPN-P5 加入低层次特征后,TS-FAN-H5 只在 Caltech 的 Near 测试子集上表现出性能提升,其原因可归结为该特征层分辨率较低,更加倾向于大尺寸行人实例的检测。而且表 2 中联合各跨尺度聚合特征在 Caltech 的 Reasonable 和 Near 测试子集上表现效果略低于单跨尺度聚合特征,而在 Medium 和 Far 测试子集上行人漏检率 MR^{-2} 为 17.24% 和 50.38%,明显优于单跨尺度聚合特征。这是由于低层次特征的加入,使得网络更加关注中、小尺度行人。最后,本文相对于单路径 RPN,在多路径 RPN 下联合各跨尺度聚合特征 TS-FAN-H3H4H5,能够更有效地检测多尺度行人实例,在 Reasonable、Near、Medium、Far 测试子集上行人漏检率 MR^{-2} 分别达到 5.53%、0.47%、13.76%、47.30%。其不同尺度行人的检测性能上均有明显的提升,其原因可归结为多路径 RPN 为第二阶段行人识别和预测行人目标包围框提供了高召回率、高质量行人候选区域集。

2.3 与目前一流行人检测方法的比较

本节为横向对比本文方法对多尺度行人检测的有效性,首先在 Caltech 测试数据集上与目前表现较好的行人检测方法 FasterRCNN + ATT^[28]、RPN + BF^[29]、AdaptFasterRCNN^[30]、F-DNN + SS^[31]、PCN^[32]、GDFL^[33]、F-DNN2 + SS^[34]、TLL-TFA 和 AR-Ped^[35] 进行了实验对比。从表 3 中可以看出,本文 TS-FAN 方法在 Caltech 数据集上取得了最好的检测性能,其在 Reasonable、All、Near、Medium 和 Far 的子集上,行人漏检率 MR^{-2} 分别为 5.53%、26.21%、0.47%、13.76% 和 47.30%。在 Caltech 测试数据集 Reasonable 子集上,本文方法相较于当前一流的 AR-Ped 方法,行人漏检率 MR^{-2} 降低了 0.92%。与当前领先的 TLL-TFA 方法相比,在 All、Near、Medium 和 Far 的子集上行人漏检率 MR^{-2} 分别降低了 11.94%、0.25%、9.16% 和 12.79%。量化的实验对比结果如图 5 中所示,可以明显地观察到本文提出的 TS-FAN 网络对于不同尺度的行人实例均表现出较好的行人检测效果。

低层次特征带来局部细节特征信息和较为精确的定位信息,使得网络对于位置感知能力更强,从而在行人间的遮挡问题上同样表现出具有竞争力的行人检测效果。TS-FAN 方法在 Caltech 数据集的 Partial 和 Heavy 子集上与当前领先的 TLL-TFA 方法相比,在 Partial 和 Heavy 子集上行人漏检率 MR^{-2} 分别降低了 7.81% 和 10.84%。在 Partial 子集上与 AR-Ped 方法相比,行人漏检率

MR⁻²降低了 1.25%。

在 ETH 测试数据集上,TS-FAN 方法与目前检测性能较好的行人检测方法 ChnFtrs^[36]、Joint-Deep^[37]、MultiSDF^[38]、DBN-Mut^[39]、TA-CNN^[40]、RPN + BF 和 F-DNN2 + SS 进行对比实验,实验结果如图 6 所示。见图 6(a)和(b),TS-FAN 方法在 All 和 Reasonable 测试子集上比目前检测效果较好的 F-DNN2 + SS 方法行人漏检率 MR⁻²降低了 7.28% 和 2.73%。尤其在 Medium 和 Far 子集上,TS-FAN 方法行人漏检率 MR⁻²提升了 19.97% 和

29.21%,如图 6(e)和(f)所示。而在 Near 子集上,TS-FAN 方法较 RPN + BF 方法和 F-DNN2 + SS 方法行人漏检率 MR⁻²分别降低了 1.51% 和 3.19%。其原因可归结为本文 TS-FAN 方法没有使用更深层次的卷积层(如特征图分辨率较原图下采样 64 倍)。

为形象地观测本文 TS-FAN 模型在 Caltech 数据集和 ETH 数据集上的检测效果,图 7 和图 8 显示了本文方法与当前一流行人检测方法的输出结果。

表 3 在 Caltech 数据集不同重叠评估设置上,本文方法与目前一流方法的比较

Table 3 Comparison of proposed method with some state-of-the-art methods on the Caltech dataset under different overlapping evaluation protocols

方法	MR ⁻² /%						
	Reasonable 子集	All 子集	Near 子集	Medium 子集	Far 子集	Partial 子集	Heavy 子集
FasterRCNN + ATT ^[28]	10.33	54.51	1.43	40.75	90.94	22.29	45.18
RPN + BF ^[29]	9.58	64.66	2.26	53.93	100	24.23	74.36
AR-Ped ^[35]	6.45	58.83	1.37	49.31	100	11.93	48.80
TLL-TFA ^[21]	7.40	38.15	0.72	22.92	60.09	18.49	28.66
TS-FAN(本文)	5.53	26.21	0.47	13.76	47.30	10.68	17.82

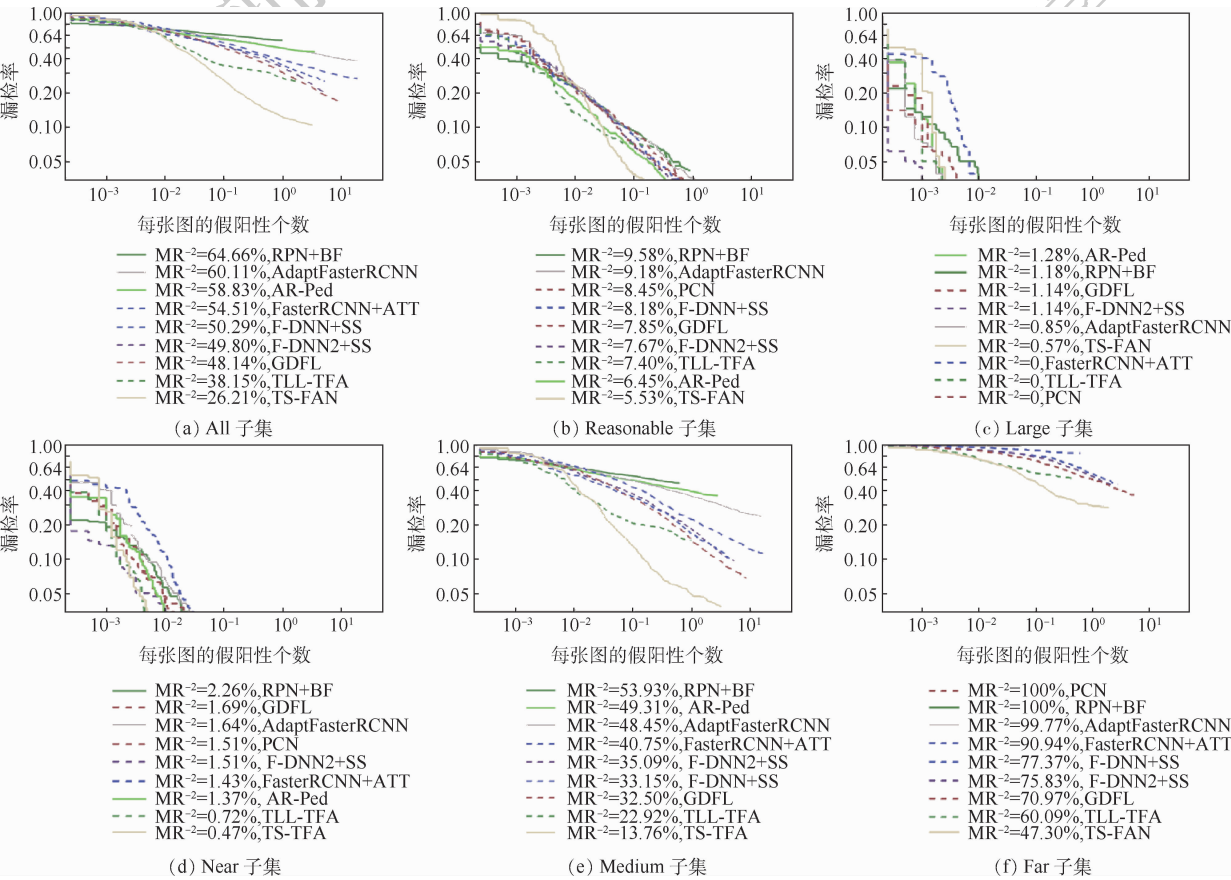


图 5 在 Caltech 数据集上,本文方法与目前一流方法的对比

Fig.5 Comparison of proposed method with some state-of-the-art methods on Caltech dataset

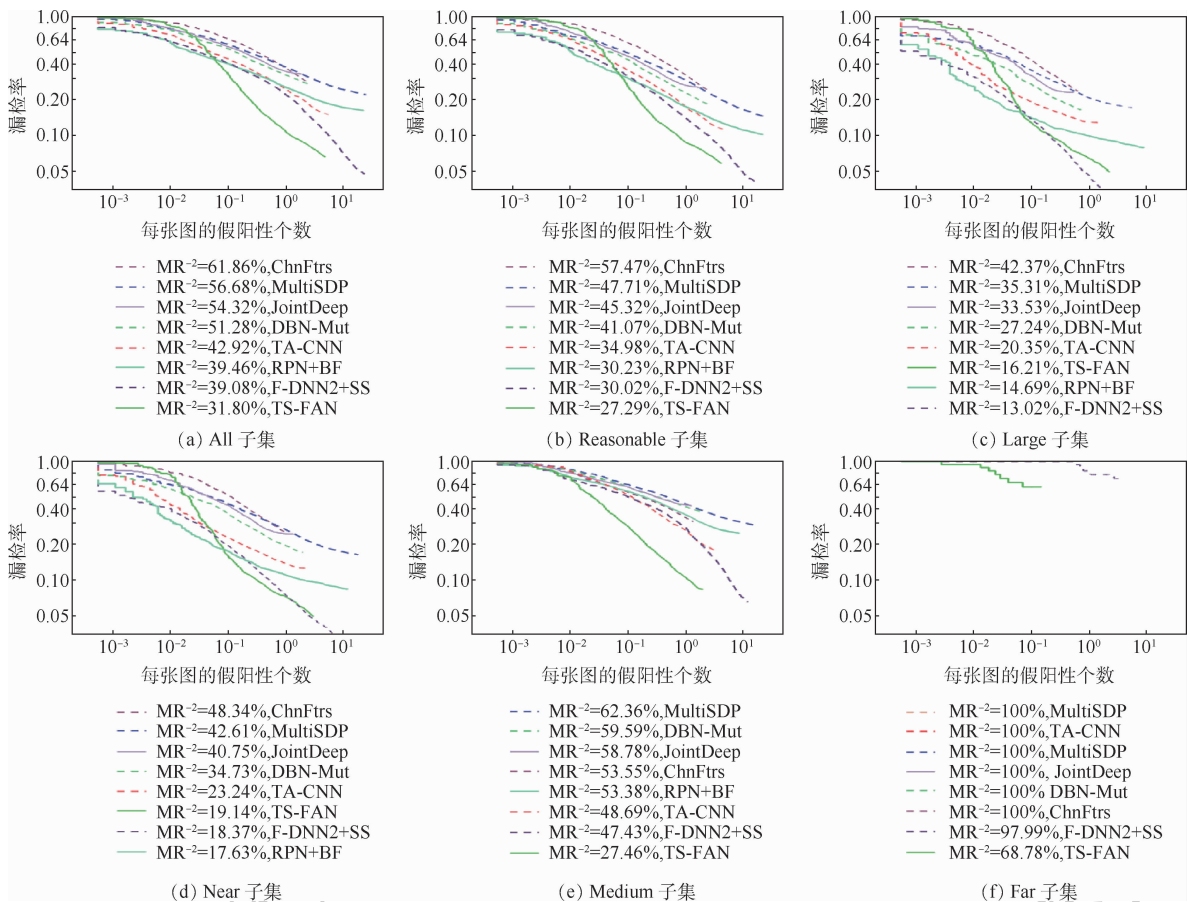


图 6 在 ETH 数据集上,本文方法与目前一流方法的对比

Fig. 6 Comparison of proposed method with some state-of-the-art methods on ETH dataset

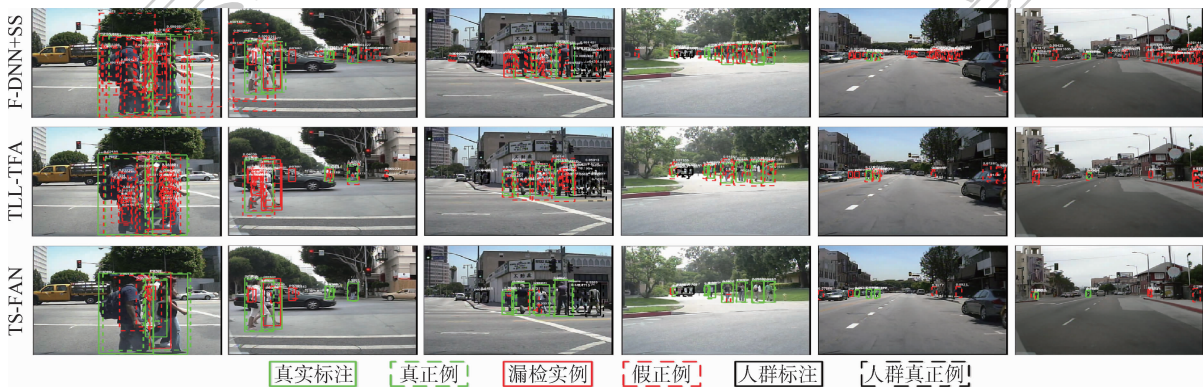


图 7 在 Caltech 数据集上,本文方法与目前一流方法可视化效果对比

Fig. 7 Comparison of visualized effects of proposed method with some state-of-the-art methods on Caltech dataset

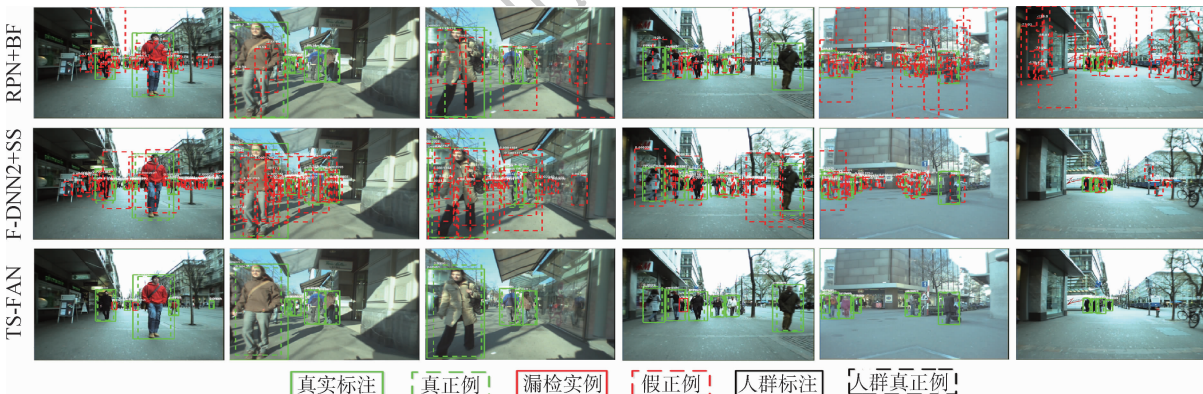


图 8 在 ETH 数据集上,本文方法与目前一流方法可视化效果对比

Fig. 8 Comparison of visualized effects of proposed method with some state-of-the-art methods on ETH dataset

3 结 论

1) 本文针对多尺度行人检测任务,设计了一种跨尺度特征聚合的多尺度行人检测网络,其通过多路径 RPN 尺度补偿策略为行人识别阶段提供了高质量的目标候选区域。

2) 提出的 TS-FAN 网络模块为多尺度行人检测网络提供了高鲁棒性的特征层用于特征提取,并通过实验验证 TS-FAN 网络模块能显著提高行人检测性能。

3) 通过多尺度行人检测网络,联合多路径 RPN 得到的候选目标区域和跨尺度聚合特征进行行人识别和精细化空间位置。实验结果表明,本文方法 TS-FAN 在 Caltech 数据集和 ETH 数据集上取得了一流的多尺度行人检测性能。

参考文献 (References)

- [1] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C] // International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2015: 91-99.
- [2] LIU W, ANGLELOW D, ERHAN D, et al. SSD: Single shot multibox detector[C] // European Conference on Computer Vision. Berlin: Springer, 2016: 21-37.
- [3] ZHANG X W, CHENG L C, LI B, et al. Too far to see? Not really! — Pedestrian detection with scale-aware localization policy[J]. IEEE Transactions on Image Processing, 2018, 27(8): 3703-3715.
- [4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 580-587.
- [5] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 886-893.
- [6] 种衍文, 匡湖林, 李清泉. 一种基于多特征和机器学习的分级行人检测方法[J]. 自动化学报, 2012, 38(3): 375-381.
- [7] ZHONG Y W, KUANG H L, LING Q Q. Two-stage pedestrian detection based on multiple features and machine learning[J]. Acta Automatica Sinica, 2012, 38(3): 375-381 (in Chinese).
- [8] SINGH B, DAVIS L. An analysis of scale invariance in object detection snippet[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 3578-3587.
- [9] SINGH B, NAJIBI M, DAVIS L. SNIPER: Efficient multi-scale training[C] // International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2018: 9310-9320.
- [10] LIU S T, HUANG D, WANG Y H. Receptive field block net for accurate and fast object detection[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 385-400.
- [11] LIU W, LIAO S C, HU W D, et al. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 643-659.
- [12] HE K Y, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
- [13] HONARI S, YOSINSKI J, VINCENT P, et al. Recombinator networks: Learning coarse-to-fine feature aggregation[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 5743-5752.
- [14] 谭红臣, 李淑华, 刘彬, 等. 特征增强的 SSD 算法及其在目标检测中的应用[J]. 计算机辅助设计与图形学学报, 2019, 31(4): 63-69.
- [15] TAN H C, LI S H, LIU B, et al. Feature enhancement SSD for object detection[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(4): 63-69 (in Chinese).
- [16] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2117-2125.
- [17] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8759-8768.
- [18] ZHAO Q J, SHENG T, WANG Y T, et al. M2Det: A single-shot object detector based on multi-level feature pyramid network[C] // AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2019: 9259-9266.
- [19] LI Y, CHEN Y, WANG N, et al. Scale-aware trident networks for object detection[C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 6054-6063.
- [20] LI J, LIANG X, SHEN S M, et al. Scale-aware fast R-CNN for pedestrian detection[J]. IEEE Transactions on Multimedia, 2017, 20(4): 985-996.
- [21] DOLLA P, WOJEK C, SCHIELE B, et al. Pedestrian detection: An evaluation of the state of the art[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34(4): 743-761.
- [22] ESS A, LEIBE B, GOOL L V. Depth and appearance for mobile scene analysis[C] // IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2007: 1-8.
- [23] SONG T, SUN L Y, XIE D, et al. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 536-551.
- [24] 王刚, 陈永光, 杨锁昌, 等. 鲁棒的红外小目标视觉显著性检测方法[J]. 北京航空航天大学学报, 2015, 41(12): 2309-2318.
- [25] WAGN G, CHEN Y G, YANG S C, et al. Robust visual saliency detection method for infrared small target[J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41(12): 2309-2318 (in Chinese).
- [26] ZHOU P, NI B B, GENG C, et al. Scale-transferrable object detection[C] // IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 528-537.
- [27] 李晓光, 付陈平, 李晓莉, 等. 面向多尺度目标检测的改进 Faster R-CNN 算法[J]. 计算机辅助设计与图形学学报, 2019, 31(7): 1095-1101.
- [28] LI X G, FU C P, LI X L, et al. Improved Faster R-CNN for multi-scale object detection[J]. Journal of Computer-Aided Design & Computer Graphics, 2019, 31(7): 1095-1101 (in Chinese).

- [25] 裴伟,许晏铭,朱永英,等.改进的SSD航拍目标检测方法[J].软件学报,2019,30(3):248-268.
- PEI W,XU Y M,ZHU Y Y,et al. The target detection method of aerial photography images with improved SSD[J]. Journal of Software,2019,30(3):248-268(in Chinese).
- [26] 许冰,牛燕雄,吕建明.复杂动态场景下目标检测与分割算法[J].北京航空航天大学学报,2016,42(2):310-317.
- XU B,NIU Y X,LYU J M. Object detection and segmentation algorithm in complex dynamic scene[J]. Journal of Beijing University of Aeronautics and Astronautics,2016,42(2):310-317(in Chinese).
- [27] LIU W, LIAO S C, REN W Q, et al. High-level semantic feature detection: A new perspective for pedestrian detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5187-5196.
- [28] ZHANG S J, YANG J, SCHIELE B. Occluded pedestrian detection through guided attention in CNNs[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6995-7003.
- [29] ZHANG L L, LIN L, LIANG X D, et al. Is Faster R-CNN doing well for pedestrian detection? [C]//European Conference on Computer Vision. Berlin: Springer, 2018: 618-634.
- [30] ZHANG S S, BENENSON R, SCHIELE B. Citypersons: A diverse dataset for pedestrian detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 3213-3221.
- [31] DU X Z, EL-KHAMY M, LEE J, et al. Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection[C]//IEEE Winter Conference on Applications of Computer Vision. Piscataway: IEEE Press, 2017: 953-961.
- [32] WANG S G, CHENG J, LIU H J, et al. PCN: Part and context information for pedestrian detection with CNNs [EB/OL]. (2018-04-12) [2020-01-27]. <https://arxiv.org/abs/1804.04483>.
- [33] LIN C Z, LU J W, WANG G, et al. Graininess-aware deep feature learning for pedestrian detection[C]//European Conference on Computer Vision. Berlin: Springer, 2018: 732-747.
- [34] DU X, EL-KHAMY M, MORARIU V I, et al. Fused deep neural networks for efficient pedestrian detection[EB/OL]. (2018-05-02) [2020-01-27]. <https://arxiv.org/abs/1805.08688>.
- [35] BRAZIL G, LIU X M. Pedestrian detection with autoregressive network phases[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 7231-7240.
- [36] DOLLAR P, TU Z, PERONA P, et al. Integral channel features[C]//British Machine Vision Conference, 2009: 91.1-91.11.
- [37] OUYANG W L, WANG X G. Joint deep learning for pedestrian detection[C]//IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2013: 2056-2063.
- [38] ZENG X Y, OUYANG W L, WANG X G. Multi-stage contextual deep learning for pedestrian detection[C]//IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2013: 121-128.
- [39] OUYANG W L, ZENG X Y, WANG X G. Modeling mutual visibility relationship in pedestrian detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2013: 3222-3229.
- [40] TIAN Y L, LUO P, WANG X G, et al. Pedestrian detection aided by deep learning semantic tasks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 5079-5087.

作者简介:

曹帅 男,硕士研究生。主要研究方向:行人检测和计算机视觉。

张晓伟 男,博士,讲师。主要研究方向:图像/视频分析和理解、计算机视觉和机器学习。

Trans-scale feature aggregation network for multiscale pedestrian detection

CAO Shuai, ZHANG Xiaowei*, MA Jianwei

(College of Computer Science and Technology, Qingdao University, Qingdao 266071, China)

Abstract: Space scale variation of pedestrian instance is one of the main bottlenecks affecting pedestrian detection performance. For this issue, a Trans-Scale Feature Aggregation Network (TS-FAN) is proposed to effectively deal with multi-scale pedestrian detection. First, in view of the feature differences among different scale spaces, we introduce a scale compensation strategy based on multi-path Region Proposal Network (RPN). According to the effectiveness of the convolutional feature layers of different scales, a series of candidate regional scale sets are generated adaptively from the feature maps corresponding to the size of the receptive field. Second, considering the semantic complementarity of convolutional features at different levels, a trans-scale feature aggregation module is proposed to effectively aggregate with semantic robustness highlevel features and with accurate location information of low-level features and achieve enhanced representation ability of convolutional features, by aggregating horizontal connection, top-down path and bottom-up path. Finally, combining the multi-path RPN scale compensation strategy and trans-scale feature aggregation module, we construct a multi-scale pedestrian detection network by adaptive scale perception. The experimental results show that, compared with the state-of-the-art method TLL-TFA, the log-average miss rate of pedestrian detection on widely-used Caltech dataset is reduced to 26.21% (increased by 11.94%) for whole-scale pedestrians (above 20 pixel in height), and 47.30% (increased by 12.79%) for small-scale pedestrian (between 20 – 30 pixels in height). And the similar improvement is also achieved on ETH dataset with drastic scale variation.

Keywords: pedestrian detection; scale perception; feature pyramid; feature aggregation; non-maximum suppression

Received: 2020-03-02; Accepted: 2020-04-09; Published online: 2020-04-17 13:01
URL: kns.cnki.net/kcms/detail/11.2625.V.20200417.0944.001.html
Foundation items: National Natural Science Foundation of China (61902204); Natural Science Foundation of Shandong Province of China (ZR2019BF028)
* Corresponding author. E-mail: xiaowei19870119@sina.com

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0073

基于联合正则化策略的人脸表情识别方法



兰凌强, 李欣, 刘淇缘, 卢树华*

(中国人民公安大学 警务信息工程与网络安全学院, 北京 102600)

摘 要: 针对目前人脸表情识别大多采用基于深度学习的端到端特征提取及分类方法的现象,提出了一种新的深度模型优化方法。基于 ResNet18 残差网络架构和正则化思想,提出了联合正则化策略,即将过滤器响应正则化和批量正则化、实例正则化和组正则化、组正则化和批量正则化分别嵌入网络之中,平衡和改善特征数据分布,弥补单一正则化的缺点,提升模型性能。在 2 个公开数据集 FER2013 和 CK+ 进行了验证和测试,最高准确率分别达到了 73.558% 和 94.9%,实验结果表明,联合正则化策略提高了基础网络的性能,其表现优于诸多当前较新的人脸表情识别方法。

关 键 词: 表情识别; 联合正则化策略; 过滤器响应正则化; 批量正则化; 组正则化

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2020)09-1797-10

人类表达情感的方式多种多样,但是最为直观的即是人脸的面部表情。在人与人的日常交流中,人们可以通过他人的表情,判断其心理状态,加深相互间的理解,故人脸表情识别一直受到学者的较多关注。此外,自动人脸表情识别在智能控制、安全防范、辅助医疗、自动驾驶和商业推广等人机交互领域得到了广泛的应用,成为计算机视觉领域较为活跃的研究课题。近年来,随着计算能力的不断提升以及神经网络的兴起,深度学习技术广泛应用于人脸表情识别研究,取得了显著的进展,特别是各种卷积神经网络的应用,大大提高了识别的准确率^[1-3]。文献[1-3]都引入了卷积神经网络用于提取更加有效的特征,相对于传统方法^[4]如 LBP (Local Binary Pattern)^[5]提取底层特征送入 SVM (Support Vector Machine) 进行分类来说,准确率有大幅度的提升。然而深度学习模型的训练需要大量的数据样本和较高的算力,使得自动人脸表情识别的准确率和实时性仍

需进一步提高,因此研究加速模型训练和提高模型性能的方法成为学者关注的热点之一。

目前,大多神经网络架构在训练和测试时使用了批量正则化 (Batch Normalization, BN)^[6]、实例正则化 (Instance Normalization, IN)^[7]和组正则化 (Group Normalization, GN)^[8]或过滤器响应正则化 (Filter Response Normalization, FRN)^[9]等方法,实践表明这些方法可以加速模型训练,防止过拟合和提高模型性能。但是不同的方法仍面临一定的局限性,例如,BN 方法对于 batch size 数值较为敏感,在其数值较大时,训练和测试结果较为理想,但是在复杂背景下或输入图片像素较大时,由于 GPU (Graphics Processing Unit) 显存有限,只能降低 batch size,而其数值较小时,模型性能并不理想,为克服此类问题,IN 和 GN 等方法相继提出,最近, Singh 和 Krishnan^[9]提出 FRN 方法,克服 BN 方法的 batch size 依赖,但是该方法的方差对输出通道大小较为敏感。

收稿日期: 2020-03-02; 录用日期: 2020-04-09; 网络出版时间: 2020-04-15 11:00

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200415.1029.002.html

基金项目: 国家重点研发计划 (2016YFC0801005); 中央高校基本科研业务费专项资金 (2019JKF225)

* 通信作者: E-mail: lushuhua@ppsuc.edu.cn

引用格式: 兰凌强, 李欣, 刘淇缘, 等. 基于联合正则化策略的人脸表情识别方法[J]. 北京航空航天大学学报, 2020, 46(9): 1797-1806. LAN L Q, LI X, LIU Q Y, et al. Facial expression recognition method based on a joint normalization strategy[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1797-1806 (in Chinese).

因此,本文提出了 FRN + BN、IN + GN、BN + GN 等联合正则化策略,弥补单一优化带来的不足,尽可能使网络保留有效的特征信息,改善数据分布,提高模型性能。以 ResNet18 作为基本框架,在标准公开数据库(FER2013^[10]数据集和 CK + ^[11]数据集)上运行实验,取得了较高的准确率,实验结果表明所提方法能够提高经典网络结构性能,具有较好的鲁棒性。

1 相关工作

传统的基于图片的表情识别通常是使用手工设计并提取底层特征,例如 LBP、HOG (Histogram of Oriented Gradient)^[12]、SIFT (Scale-Invariant Feature Transform)^[13]等,在早期,这些方法在一些数据集上都取得了较好的结果^[11,14-15]。随着 ILSVRC2013 (Large Scale Visual Recognition Challenge 2013)的落幕,深度神经网络得以迅速的发展,并逐步应用到人脸表情识别中。深度神经网络特别是深度卷积神经网络(Convolutional Neural Networks, CNN)是提升机器视觉任务方面(例如人脸识别^[16]、属性识别^[17,18]和表情识别^[19-20]等)重要方法。卷积神经网络可进行人脸表情深层特征提取和分类,对于各种干扰更加具有鲁棒性,因此更适合分类,使识别准确率得到显著提高。以下从人脸表情识别的神经网络架构和数据正则化两方面简要介绍。

卷积神经网络的结构受到学者较多关注。Yu 和 Salzmann^[21]提出了新的网络,在经过卷积后的特征图中提取协方差矩阵,计算矩阵参数的二阶变换,对矩阵进行参数向量化,由此二阶统计特征也开始运用到识别上。Gao 等^[22]运用全局二阶统计特征提出了 Gsop (Global second-order pooling)网络,可以捕获沿通道尺寸或位置维度的全局二阶统计量,充分利用图像中的上下文信息并且内存和计算复杂度都减小。Huang 和 Luc^[23]提出了新颖的黎曼网络架构,为黎曼流形上的对称正定矩阵(Symmetric Positive Definite, SPD)深度非线性学习开辟了新的方向,且把 Riemannian 结构并入深度网络体系结构以压缩数据空间和权重空间。随着 SPD 网络的提出,Acharya 等^[24]把 SPD 网络与 VGG 网络相结合,把得到的卷积特征图先展开计算其协方差矩阵,再把协方差矩阵送入 SPD 网络中进行运算。Hamester 等^[25]提出了一种多通道卷积神经网络,通过结合卷积神经网络和自动编码器的信息来识别面部表情。Molla-

hosseini 等^[26]等提出具有 Inception 层的网络架构,把 Inception 运用在人脸表情识别上面,结合面部动作进行表情分析。Chen 和 Hu^[27]提出类间关系学习网络,同时输入两张不同表情的图片经过几个 Inception 层得到特征图进行融合学习不同表情之间的关系,再用注意力机制送入分类器。Liu 等^[28]提出了一种具有可变动作部分约束的 3D 卷积神经网络,该网络可以检测到特定的面部部分动作并且获取到易区分的表征特征。Nguyen 等^[29]提出了一个基于残差网络为基础的多特征水平的融合方法,利用底层特征和高层特征的融合来辅助提高其准确率。

深度学习模型的加速和优化发展迅速,2015 年,Ioffe 和 Szegedy^[6]提出批量正则化,对经过卷积层之后的数据进行标准化,加速网络的处理速度,但是批量正则化对于批处理的大小敏感,导致批处理越小其效果越差。2017 年,Ulyanov 等^[7]提出实例正则化,对单个图片进行处理保留图像中颜色、风格,虚拟的或现实的不敏感的特征,其效果在图像风格迁移,高分辨重组等方面具有良好表现。Wu 和 He^[8]在 2018 年为克服批量正则化的局限提出了组正则化,提高了在批处理较小时的效果,但是在批处理大时并没有批量正则化的效果好。2019 年,谷歌团队^[9]提出了过滤器响应正则化,对于组正则化来说又是进一步的提升,其保留了组正则化的优点,同时保留了批量正则化在批处理大时的效果。

2 基本原理

正则化的基本原理是改善神经网络学习模型的数据分布情况,若没有做正则化处理,那么每一批次训练数据的分布都不一样,由于每层输入数据分布都在不断的变化,这会导致网络难以收敛。而正则化主要是对输入的数据进行标准化以提高代码的处理速度^[30],防止过拟合,进而增强其泛化能力。在深度神经网络中常用的几种正则化方法有:批量正则化、实例正则化、组正则化和过滤器响应正则化。然而每一种正则化都应用在不同的领域上,批量正则化主要用于图像分类和识别问题上;实例正则化主要用于风格迁移、图像超分辨率上;而组正则化是弥补了批量正则化在批处理过小时表现效果较差而提出,在一些极端条件下可以等价于实例正则化;过滤器响应正则化不仅仅弥补批量正则化对于批处理过小时表现差同时还保持了批量正则化对于批处理大时的效果。

2.1 批量正则化

批量正则化是最早提出,也是效果最好的一种正则化的方式,其主要是通过式(1)计算其均值,再利用式(2)对每个通道进行正则化^[6]。

$$\mu_{BN} = \frac{1}{NHW} \sum_{n=1}^N \sum_{h=1}^H \sum_{w=1}^W x_{nchw} \quad (1)$$

$$\hat{x}_{BN} = \frac{x_{nchw} - \mu_{BN}}{\sqrt{\frac{1}{NHW} \sum_{n=1}^N \sum_{h=1}^H \sum_{w=1}^W (x_{nchw} - \mu_{BN})^2 + \omega}} \quad (2)$$

式中: x_{nchw} 为特征张量元素值, $\mathbf{x}=[N,C,H,W]$ 为特征张量, N 为批处理的大小, C 为通道数, H 和 W 分别为特征图的高和宽; μ_{BN} 为计算后特征图的均值; ω 为常数; $\hat{x}_{BN}$ 为特征经过正则化操作后的值。

批量正则化主要是针对每一个神经元,使数据在进入激活函数之前,沿着通道计算批处理的均值、方差,强制使得数据保持在均值为0,方差为1的正态分布,减少训练卷积神经网络时内部协变量的偏移,避免梯度消失,提高学习速度并且加快收敛速度,这使得批量正则化成为了一些主流的卷积神经网络架构的一部分,例如ResNeXt^[31]、Inception^[32]、Dense Net^[33]等,批正则化也有一定的缺点,因为均值和方差是根据训练集预先计算得出,运行通常使用平均值,因此在测试阶段没有进行正则化,当目标数据分布发生变化,训练计算出的结果也可能产生变化,这些问题会导致在训练和测试时产生矛盾。此外,批处理的大小对于合理估计训练数据的均值和方差有较大的影响。

2.2 实例正则化

与批量正则化不同的是,批量正则化注重每一次批处理的大小进行正则化,保证数据分布一致,因为判别模型中结果决于数据整体的分布情况。然而实例正则化是对每个批次中的单个图片进行正则化,所以实例正则化的信息都来自于自身的图片,相当于对全局信息做了一次整合和调整,但是对于各个通道之间没有信息的交流,其主要是通过式(3)计算其均值,接着利用式(4)对单个特征图进行正则化^[7]。

$$\mu_{IN} = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W x_{nchw} \quad (3)$$

$$\hat{x}_{IN} = \frac{x_{nchw} - \mu_{IN}}{\sqrt{\frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W (x_{nchw} - \mu_{IN})^2 + \omega}} \quad (4)$$

式中: μ_{IN} 为计算后特征图的均值; $\hat{x}_{IN}$ 为特征经过正则化操作后的值。

实例正则化主要用于风格迁移和类似的任务上,因为它可以从内容中过滤出特定实例的对比

信息。但是把实例正则化主要用于图像分类的问题上,相比批量正则化的结果来说稍差^[34]。

2.3 组正则化

组正则化是为了解决批量正则化中批处理的大小过于小而提出的,组正则化是将通道分成组,并在每组内计算正则化的均值和方差。组正则化的计算与批量大小无关,所以其准确率在各种批量大小下都很稳定,组正则化首先是通过式(5)计算分组的均值,接着通过式(6)对分组进行正则化操作, G 为对通道数 C 的分组大小^[8]。

$$\mu_{GN} = \frac{1}{(C/G)HW} \sum_{c=gC/G}^{(g+1)C/G} \sum_{h=1}^H \sum_{w=1}^W x_{nchw} \quad (5)$$

$$\hat{x}_{GN} = \frac{x_{nchw} - \mu_{GN}}{\sqrt{\frac{1}{(C/G)HW} \sum_{c=gC/G}^{(g+1)C/G} \sum_{h=1}^H \sum_{w=1}^W (x_{nchw} - \mu_{GN})^2 + \omega}} \quad (6)$$

式中: G 为通道分组数量; μ_{GN} 为计算后特征图的均值; $\hat{x}_{GN}$ 为特征经过正则化操作后的值; g 为通道分组后的组标。

2.4 过滤器响应正则化

过滤器响应正则化和实例正则化是在同一个维度进行操作,对每一个样本的每一个通道:

$$v^2 = \frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W x_{nchw}^2 \quad (7)$$

$$\hat{x}_{FRN} = \frac{x_{nchw}}{\sqrt{\frac{1}{HW} \sum_{h=1}^H \sum_{w=1}^W x_{nchw}^2 + \omega}} \quad (8)$$

$$y_i = \gamma \frac{x_{nchw}}{\sqrt{v^2 + \omega}} + \beta \quad (9)$$

式中: γ 和 β 为可学习的参数,通过网络的训练学习得到; v^2 为特征图的二次范数的平均值; $\hat{x}_{FRN}$ 为正则化后特征取值; y_i 为特征经过正则化操作后网络输出的值。

2.5 联合正则化

目前,绝大多数的深度网络使用的都是单一的正则化策略,该策略只能够保留其自身的优势,同时也会带来一定不足。联合正则化策略,使用了两种正则化相结合的优化方法,保存了两种正则化策略的优点,还互补了两者的不足。就4种正则化策略来说,批量正则化和组正则化保留了各个样本之间的区别,但是也使得卷积神经网络容易受到外观变换的影响。实例正则化消除了个体的对比,但是同时消减了有用的信息,过滤器响应正则化弥补了批量正则化的缺点,同时还保持了批量正则化的优点,但是不适用于非卷积运算,并且过滤器响应正则化的方差对输出通道大小敏感。上

述方法都有着各自的局限性, 综合考虑, 提出联合正则化策略引入到网络结构中以保留各自的优势, 并形成互补性。借鉴文献[35]思路, 对 ResNet18 残差网络的基本模块进行改善, 在通过第一个卷积产生 64 个通道后, 将 64 个通道分成两个 32 通道, 把这两个 32 通道分别送到不同正则化函数如 FRN + BN、IN + GN 以及 BN + GN 再分别通过式(10)~式(12)进行正则化处理, 接着把正则化过后的两个 32 通道拼接在一起, 送到下一个卷积。

$$\hat{x}_{FRN+BN} = \text{cat}\left(\frac{C}{2}\hat{x}_{FRN}, \frac{C}{2}\hat{x}_{BN}\right)$$
 (10)

$$\hat{x}_{IN+GN} = \text{cat}\left(\frac{C}{2}\hat{x}_{IN}, \frac{C}{2}\hat{x}_{GN}\right)$$
 (11)

$$\hat{x}_{BN+GN} = \text{cat}\left(\frac{C}{2}\hat{x}_{BN}, \frac{C}{2}\hat{x}_{GN}\right)$$
 (12)

式中: $\hat{x}_{FRN+BN}$ 、 $\hat{x}_{IN+GN}$ 、 $\hat{x}_{BN+GN}$ 表示经过联合正则化处理后的特征; cat 函数用于连接不同正则化函数处理后的特征。采用这种方法引入两种正则化函数, 尽可能的保留有效的特征, 辅助提高其准确率。由于没有增加通道数量, 故联合正则化策略并没有增加计算量。

3 实验

3.1 数据集

为评估所提出的联合正则化方法的效果和泛化性, 本文在 2 个典型的国际公开数据集 (FER2013 和 CK+) 上进行了实验。以下将简要介绍一下这 2 个数据集。

3.1.1 FER2013 数据集

该数据集是由 ICML (International Conference on Machine Learning) 2013^[10] 挑战产生, 是表情识别数据集中比较大的数据集, 这些图片都是从谷歌图片中获取并且每张图片是在非限制条件下拍摄, 所以这些图片可能带有噪声或者是质量相对较差。所有的图片都经过预处理过并且裁剪为 48 × 48。包含了 28 709 张训练图片, 3 589 张验证图片和 3 589 张测试图片都带有 7 种表情标签, 即

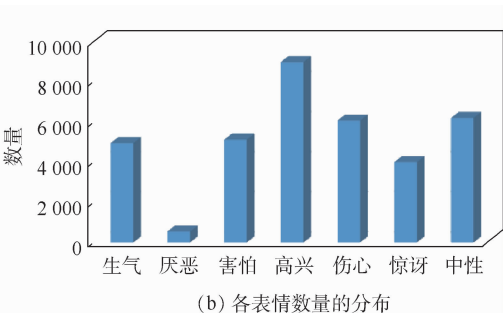
生气、厌恶、害怕、高兴、伤心、惊讶和中性。这 7 种表情分类是基于心理学家 Ekman 与 Friesen^[36] 对表情做出了定义, 将表情分为 6 种 (生气、厌恶、害怕、高兴、伤心和惊讶) 适用于不同种族和文化差异的基础上, 后期又添加了中性表情与其中一共构成 7 种表情。图 1 (a) 展示了 FER2013 各个表情分类的示例, 图 1 (b) 给出了 FER2013 各个表情数量的分布情况。

3.1.2 CK + 数据集

扩展 CK + 数据集用于评估人脸识别系统最广泛的实验室控制数据集。CK + 包含了从 123 个对象中提取的 593 视频序列。这些序列持续时间从 10 帧到 60 帧不等, 展示了从中性脸表情到高峰表情的转变。在这些视频中来自 118 个对象的 327 个序列带有 7 个基础表情标签 (生气、蔑视、厌恶、害怕、高兴、伤心和惊讶) 都是基于面部运动编码系统。与 FER2013 不同, CK + 中用了蔑视代替中性表情并且 CK + 没有提供专门的训练集、验证集、测试集, 所以算法的评估并没有统一。基于静态的图片方法, 大多数人使用的是提取最后 1~3 帧具有高峰表情的帧和每个序列的第 1 帧 (中性面), 然后把这些数据分为 n 组, 进行 n 倍的交叉验证实验, 一般 n 取 5、8、10。图 2 (a) 展示了 CK + 表情分类的示例, 图 2 (b) 给出了 CK + 各个表情数量的分布情况。

3.2 网络结构及其训练

基于 ResNet18 网络架构基础, 提出 3 种联合正则化模型对网络进行优化, 并用于表情识别中, 把过滤器响应正则化与批量正则化、实例正则化与组正则化、组正则化和批量正则化分别组合。ResNet18 包含了 4 个残差块, 如图 3 中 (a) 表示使用了单一正则化策略 (BN、FRN、GN 和 IN) 作为基础模块。笔者把残差网络第 1 层卷积 7 × 7 用 3 × 3 卷积替代, 因为 FER2013 和 CK + 都是灰度图像, 于是把这 2 个数据集的图片复制成 3 份使得残差网络的输入为 3 通道, 在最后一个卷积



(a) 表情分类实例

(b) 各表情数量的分布

图 1 FER2013 数据集表情分类示例及表情数量分布

Fig. 1 Samples of FER2013 dataset for facial expression and distribution of number of each facial expression

层和全连接层之间加入了丢弃层为了防止过拟合,并且在网络的最后只加了一个带有分类器的全连接层,减少了计算量。仅将所提的 3 种组合优化方式添加到残差网络中的前 3 个残差块中,最后一个残差块保留不动。对于每一个残差块,在第一个卷积层后面的批量正则化改成 3 种组合模式进行优化,如图 3 中(b)、(c)、(d)所示。

之所以不在残差网络的恒等路径上加入优化函数,是基于文献[37]中提出的对于残差网络来说不需要对恒等路径做过多的处理,否则会导致残差网络的效率降低。也不在第 2 个卷积层后加入组合的优化策略是为了避免匹配的问题,这样的设计是对模型容量大小来考虑,另一方面可以使用不同优化函数的优点相互结合,具有互补性。

如图 3 中(b)所示,把过滤器响应正则化与批量正则化相结合,弥补了批量正则化的缺点,同时还继承了过滤器响应正则化优点。如图 3 中(c)所示,把实例正则化和组正则化相结合,这样可以利用实例正则化学习到对形状变换,比如颜色,风格,虚拟的或现实的不敏感的特征,利用组正则化对于批处理大小的不敏感性来弥补批量正则化的不足,保留其纹理信息。如图 3 中(d)所示,把组正则化与批量正则化相结合,用组正则化来弥补批量正则化对批处理大小敏感问题,利用批量正则化来进行各通道信息交流。

实验是在 Ubuntu16. 04 系统下利用 Pytorch 0. 4. 1 运行并且使用了 Nvidia RTX 2070 GPU 进行训练,权重更新使用了随机梯度下降将其设置

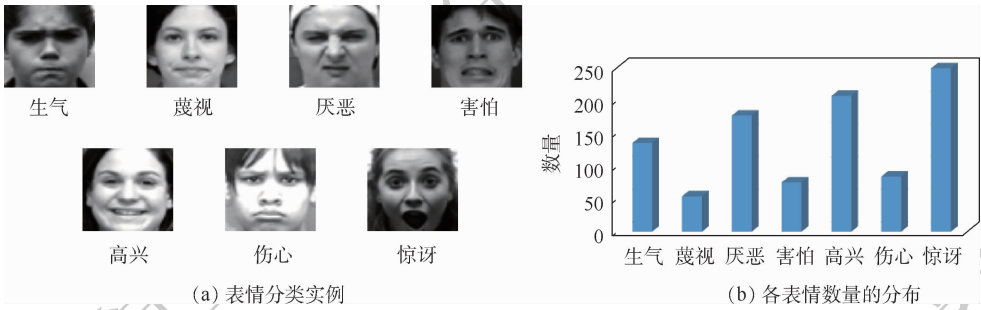
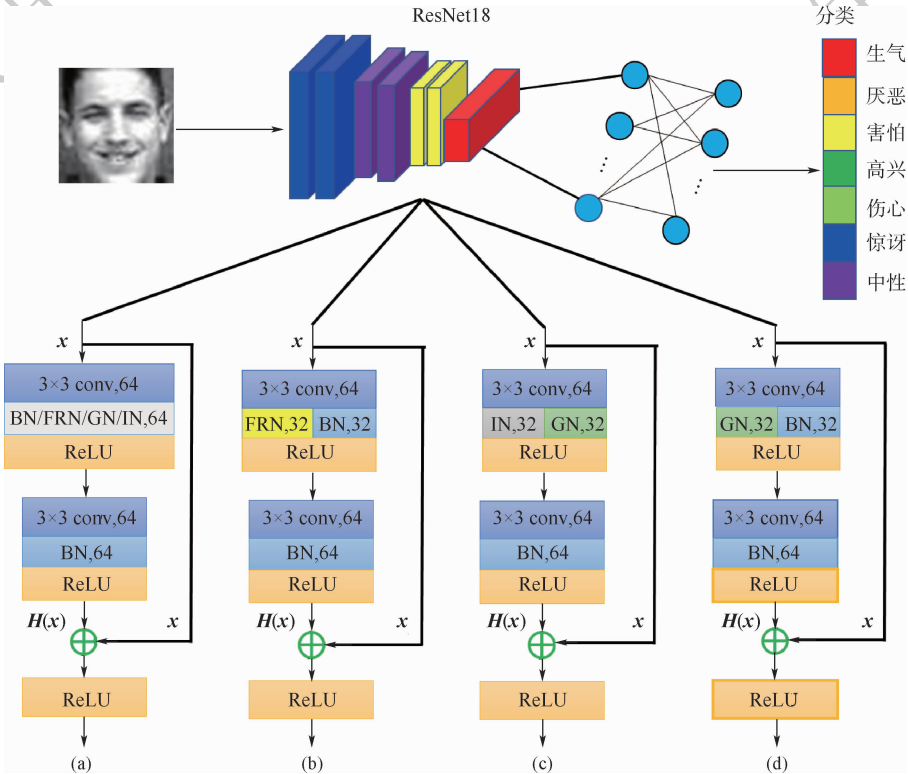


图 2 CK+ 数据集表情分类示例及表情数量分布

Fig. 2 Samples of CK+ dataset for facial expression and distribution of number of each facial expression



(a) 为采用 BN/FRN/GN/IN 单一正则化的残差模块,箭头指向(a)表示整个网络以(a)的模块为基础;(b)、(c)、(d)为所提 3 种优化策略在残差块中的应用,箭头指向(b)、(c)、(d)分别表示使用(b)、(c)、(d)作为基础模块的残差网络。

图 3 网络架构

Fig. 3 Network architecture

动量为 0.9,权重衰减为 0.000 5。初始的学习率设置为 0.01,迭代次数设置为 350 并且在 80 次迭代后开始衰减。在训练过程中,首先对面部表情分类器进行了微调,为了防止过拟合提高其准确率,其次对数据集都采取了数据增强的策略,对每一张数据集的图片随机创建了 10 张大小为 44×44 的裁剪图片。此外,还收集了每个面部表情的 10 张处理过的图像,通过裁剪图片的右上角、左上角以及中心等区域来测试,最后通过对这 10 张处理过后的图片取平均误差减少分类错误,做出判断。

在实验中,运用了所提出的 3 种联合正则化方案的基本框架是采用了文献[37]中的 ResNet18 网络,其中的 ResNet18 使用单一的批量正则化进行优化。对所提出的每一种正则化方案进行了命名,将 FRN + BN 联合正则化称为 Model1;IN + GN 联合正则化称为 Model2;BN + GN 联合正则化称为 Model3。

4 结果与讨论

4.1 FER2013 数据集实验结果

表 1 展示了所提出的联合正则化方法及基础框架在数据集 FER2013 上私有测试集的性能比较,从表 1 可以看出,本文所提出的 3 种联合正则化策略相较于文献[38]中的残差网络框架的准确率都有一定的提高,其中 Model1 取得了最优的效果。

表 1 基础框架以及添加联合正则化策略后的实验结果
Table 1 Experimental results of basic framework and adding joint normalization strategies

模型	准确率/%
文献[38]	71.190
Model1(本文)	73.558
Model2(本文)	73.534
Model3(本文)	73.031

为了比较不同数量的联合正则化对于 ResNet18 的效果,在数据集 FER2013 上进行了测试,结果如表 2 所示。其中,残差网络具有 4 个残差块,0 表示不加入联合正则化,1-2、1-3 以及 1-4 分别表示在前 2 个残差块、前 3 个残差块以及全部残差块都加入了联合正则化。从实验结果得出,Model1、Model2 和 Model3 加入联合正则化在浅层时能够提高网络性能,其中在前 3 个残差块中使用 Model1 取得了最优的效果,可能源于在底层中使用联合正则化能够最大可能保留了有效特征,辅助提高了准确率。但是加入在最后一个残差块

表 2 残差网络添加联合正则化数量的效果比较
Table 2 Comparison of impact of adding number of joint normalization based on residual network

数量	准确率/%		
	Model1	Model2	Model3
0	71.190	71.190	71.190
1	72.555	72.722	72.499
1-2	72.053	72.417	71.691
1-3	73.558	73.530	73.031
1-4	72.778	72.416	72.723

时 3 种联合正则化模型的性能都开始下降,表明在高层中使用 BN 保留其抽象信息可能更为重要。

此外,笔者对单一优化以及联合正则化进行了对比,通过对不同数量的联合正则化的效果实验,采用前 3 个残差块加入联合正则化(ResNet18(1-3)),为了进行对比实验,将联合正则化部分替换成单一优化,具体实验结果如表 3 所示。从表 3 可以看出,仅单一使用 FRN 优化的准确率为 72.276%,但是 Model1(FRN + BN)联合正则化达到了 73.558%,与单一使用 FRN 相比效果有较为明显提升。仅使用 IN 优化与 Model2 相比,Model2(IN + GN)的效果提升了 0.362%,Model2 与 Model3 中都使用了 GN 与其他正则化函数的结合,其结果与仅使用 GN 的效果相比也有一定的提高。可见,联合正则化策略能够融合 2 种正则化方法的优点,提高网络性能。

表 3 单一正则化与联合正则化(在前 3 个残差块中使用的)的比较
Table 3 Comparison between individual normalization and joint normalization(used in the first three residual blocks)

优化策略	准确率/%
BN	71.190
IN	73.168
GN	73.029
FRN	72.276
Model1	73.558
Model2	73.530
Model3	73.031

为进一步研究所提正则化策略的性能,将本文 Model1、Model2 以及 Model3 与目前已有比较新的几种方法(SHCNN^[39]、文献[40]、IcRL^[27]和文献[41])在 FER2013 数据集上进行了比较,结果如表 4 所示。SHCNN^[39]提出了一个 3 个卷积层($5 \times 5, 3 \times 3, 5 \times 5$)的浅层网络,为了防止过拟合以及梯度爆炸的问题。在文献[40]中,对于基础的 Softmax 函数进行了研究并进行了改善,通

过控制余弦值和输入特征图的大小来分析和提高 Softmax 函数,从而获得更加具有区分力的特征,以助于提高表情识别的准确率。文献[27]提出了一个新的网络架构(IcRL),同时从两个不同表情的图像中提取特征,然后将两个提取的特征以随机比率进行整合,从而获得混合特征。还提出了一个关注模块,为混合特征的每个像素分配权重。最后,将加权后的混合特征输入到后续的分类模块中。整个网络训练的目的是输出每个表达式的正确比例,从而学习不同类别的表情之间的相互关系,并扩大类间距离与类内距离之比。文献[41]是在类感知余量和对人脸表情识别三元组损失具有异常值抑制的基础上提出了一个新的 Triplet 损失函数,对于每一对表情例如高兴和害怕都分配了自适应的余量参数,并且根据特征距离分布抛弃异常的 Triplet。这些方法都在

FER2013 数据集上都取得了良好的效果。相比较而言,本文所提 Model1 取得了最高的准确率,至少提高了约 1%。图 4 展示了 Model1 在 FER2013 上私有测试集和公有测试集的混淆矩阵。

表 4 本文方法与目前较新的方法在 FER2013 数据集上准确率比较

模型	准确率/%
SHCNN ^[39]	69.100
文献[40]	70.910
IcRL ^[27]	72.360
文献[41]	72.640
Model1(本文)	73.558
Model2(本文)	73.534
Model3(本文)	73.031

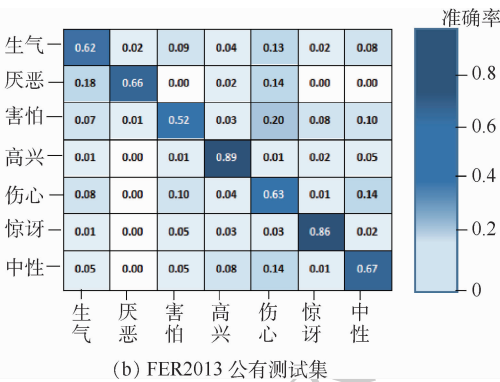
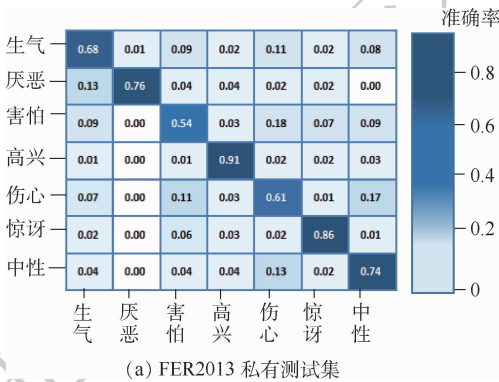


图 4 FER2013 私有和公有测试集混淆矩阵

Fig.4 Confusion matrix for FER2013 private and public test sets

4.2 CK + 数据集实验结果

为了研究所提正则化策略在其他数据集上的性能,选取 CK + 数据集进行了实验,并与一些方法的结果进行了比较,具体内容如表 5 所示。对于 CK + 数据集使用了十倍交叉验证法进行验证。从表 5 中可以得出,对于 CK + 数据集来说,所提出的联合正则化模型相对于文献[38]的基本框架

来说准确率有较为明显的提高,最多提高了 5.6%。此外本文展示了 Model1 在 CK + 数据集上的混淆矩阵,如图 5 所示。但是其中 7 种表情中,蔑视使用的训练样本只有 54 张图片,数量较少导致其识别准确率较低,只有 78%,从而影响了平均准确率。从结果可以看出所提出的正则化策略性能较好,优于部分所比较的方法,表明了联合正则化具有一定的优越性。

表 5 本文方法与目前较新的方法在 CK + 数据集上准确率比较

模型	准确率/%
3DCNN-DAP ^[28]	92.4
Inception ^[26]	93.2
文献[1]	93.2
DAM-CNN ^[42]	95.9
文献[38]	89.3
Model1(本文)	94.9
Model2(本文)	93.6
Model3(本文)	94.1



图 5 CK + 数据集混淆矩阵

Fig.5 Confusion matrix for CK + dataset

4.3 结果讨论

从 4.1 节、4.2 节结果可以看出,相对于无正则化和单一正则化方法,所提出的联合正则化策略在非限制条件下和限制条件下两种数据集上表现较好,表明将联合正则化引入卷积神经网络中,可使网络既能利用单一正则化的优点,又能保留与表情相关的有效特征,从而提高网络的性能。由于没有增加通道数量,故联合正则化策略并没有增加计算量。

Model1 采用 FRN + BN 策略,FRN 弥补了 BN 批处理大小的问题,同时消除尺度带来的问题,而 BN 保留纹理相关的特征,两者结合尽可能地保留了与表情相关的有效特征,提高了识别准确率。Model2 采用 IN + GN 策略,在底层使用 IN 保留例如颜色、风格等不敏感的特征,同时结合 GN 保留纹理信息消除对批处理大小的依赖,改善单一正则化的弱点。Model3 采用 BN + GN 策略,表明对于单一使用 BN 不如使用 BN 和 GN 联合正则化的效果,在同时使用 GN 和 BN 的过程中,部分的特征图使用了 GN 消减了仅使用 BN 时对于批处理大小的过度依赖,提高了其准确率。

此外,使用联合正则化在底层保留部分除纹理以外的特征能够辅助提高分类准确率,但在高层,相关内容信息可能更为重要,因此,在底层使用不同的正则化策略利用各自的优点存储不同的特征能够更好地提高模型性能。

5 结 论

1) 本文提出了 3 种联合正则化策略,将过滤器响应正则化与批量正则化、实例正则化和组正则化、组正则化和批量正则化两两结合,弥补单一正则化所带来的不足,尽可能保留有效的特征信息,提高网络分类的准确性。

2) Model1 在国际公开数据集 FER2013 取得 73.558% 的准确率,CK + 上取得了 94.9% 的准确率,表现效果比 Model2 和 Model3 更好并且优于诸多当前人脸表情识别方法。

3) 联合正则化策略,可以尝试嵌入到绝大多数的网络框架中,加速模型训练,辅助提高其效果以及泛化能力。

参考文献 (References)

- [1] JAIN D K, SHAMSOLMOALI P, SEHDEV P. Extended deep neural network for facial emotion recognition [J]. Pattern Recognition Letters, 2019, 120: 69-74.
- [2] HU S H, HU Y M, LI J Q, et al. Natural scene facial expression recognition based on differential features [C] // 2019 Chinese Automation Congress (CAC). Piscataway: IEEE Press, 2019: 2840-2844.
- [3] LI Y, CAO G T, CAO W M. Stacking-based deep neural network for facial expression recognition [C] // 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway: IEEE Press, 2019: 1338-1342.
- [4] HE J, CAI J F, FANG L Z, et al. A method of facial expression recognition based on LBP fusion of key expressions areas [C] // The 27th Chinese Control and Decision Conference (2015 CCDC). Piscataway: IEEE Press, 2015: 4200-4204.
- [5] OJALA T, PIETIKÄINEN M, MÄENPÄÄ T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24 (7): 971-987.
- [6] IOFFE S, SZEGEDY C. Batch normalization: Accelerating deep network training by reducing internal covariate shift [C] // Proceedings of the 32nd International Conference on Machine Learning (ICML). New York: ACM, 2015: 448-456.
- [7] ULYANOV D, VEDALDI A, LEMPITSKY V. Instance normalization: The missing ingredient for fast stylization [EB/OL]. (2017-11-06) [2020-02-20]. <https://arxiv.org/abs/1607.08022>.
- [8] WU Y X, HE K M. Group normalization [C] // The European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 3-19.
- [9] SINGH S, KRISHNAN S. Filter response normalization layer: Eliminating batch dependence in the training of deep neural networks [EB/OL]. (2019-11-21) [2020-02-20]. <https://arxiv.org/abs/1911.09737>.
- [10] GOODFELLOW I J, ERHAN D, CARRIER P L, et al. Challenges in representation learning: A report on three machine learning contests [C] // International Conference on Neural Information Processing. Berlin: Springer, 2013: 117-124.
- [11] LUCEY P, COHN J F, KANADE T, et al. The extended Cohn-Kanade dataset (CK+) : A complete dataset for action unit and emotion-specified expression [C] // 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2010: 94-10.
- [12] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2005, 1: 886-893.
- [13] LOWE D G. Object recognition from local scale-invariant features [C] // Proceedings of the 7th IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 1999, 2: 1150-1157.
- [14] LYONS M J, AKAMATSU S, KAMACHI M, et al. The Japanese female facial expression (JAFPE) database [C] // Proceedings of 3rd International Conference on Automatic Face and Gesture Recognition. Piscataway: IEEE Press, 1998: 14-16.
- [15] ZHAO G, HUANG X, TAINI M, et al. Facial expression recognition from near-infrared videos [J]. Image and Vision Computing, 2011, 29 (9): 607-619.
- [16] ELSAYED A, MAHMOOD A, SOBH T. Effect of super resolution based on differential features [C] // 2019 Chinese Automation Congress (CAC). Piscataway: IEEE Press, 2019: 2840-2844.

- tion on high dimensional features for unsupervised face recognition in the wild [C] // 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR). Piscataway: IEEE Press, 2017: 1-5.
- [17] WANG P Y, SU F, ZHAO Z C. Joint multi-feature fusion and attribute relationships for facial attribute prediction [C] // 2017 IEEE Visual Communications and Image Processing (VCIP). Piscataway: IEEE Press, 2017: 1-4.
- [18] TAHERKHANI F, NASRABADI N M, DAWSON J. A deep face identification network enhanced by facial attributes prediction [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2018: 553-560.
- [19] KUO C M, LAI S H, MICHEL S. A compact deep learning model for robust facial expression recognition [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2018: 2121-2129.
- [20] YANG J, ZHANG F, CHEN B, et al. Facial expression recognition based on facial action unit [C] // 2019 10th International Green and Sustainable Computing Conference (IGSC). Piscataway: IEEE Press, 2019: 1-6.
- [21] YU K, SALZMANN M. Second-order convolution neural networks [J]. Clinical Immunology & Immunopathology, 2017, 66 (3): 230-238.
- [22] GAO Z, XIE J, WANG Q, et al. Global second-order pooling convolutional networks [C] // 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 3024-3033.
- [23] HUANG Z, LUC V G. A Riemannian network for SPD matrix learning [C] // 31st AAAI Conference on Artificial Intelligence. San Francisco: AAAI Press, 2017: 2036-2042.
- [24] ACHARYA D, HUANG Z, PAUDEL D P, et al. Covariance pooling for facial expression recognition [C] // 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2018: 367-374.
- [25] HAMESTER D, BARROS P, WERMTER S. Face expression recognition with a 2-channel convolutional neural network [C] // 2015 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2015: 1-8.
- [26] MOLLAHOSSEINI A, CHAN D, MAHOOR M H. Going deeper in facial expression recognition using deep neural networks [C] // 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2016: 1-10.
- [27] CHEN Y, HU H. Facial expression recognition by inter-class relational learning [J]. IEEE Access, 2019, 7: 94106-94117.
- [28] LIU M, LI S, SHAN S, et al. Deeply learning deformable facial action parts model for dynamic expression analysis [C] // Asian Conference on Computer Vision (ACCV). Berlin: Springer, 2014: 143-157.
- [29] NGUYEN D H, KIM S H, LEE G S, et al. Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks [EB/OL]. (2019-10-10) [2020-02-20]. <https://ieeexplore.ieee.org/document/8863974>.
- [30] MONTAVON G, ORR G B, MULLER K R. Neural networks: Tricks of the trade [M]. Berlin: Springer, 1998: 9-50.
- [31] XIE S Y, GIRSHICK R, DOLLAR P, et al. Aggregated residual transformations for deep neural networks [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 1492-1500.
- [32] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C] // 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 1-9.
- [33] HUANG G, LIU Z, WEINBERGER K Q, et al. Densely connected convolutional networks [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 4700-4708.
- [34] ULYANOV D, VEDALDI A, LEMPITSKY V. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis [C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 6924-6932.
- [35] PAN X, LUO P, SHI J, et al. Two at once: Enhancing learning and generalization capacities via IBN-net [C] // The European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 464-479.
- [36] EKMAN P, FRIESEN W V. Facial action coding system: A technique for the measurement of facial movement [M]. Palo Alto: Consulting Psychologists Press, 1978: 32-96.
- [37] HE K, ZHANG X, REN S, et al. Identity mappings in deep residual networks [C] // The European Conference on Computer Vision. Berlin: Springer, 2016: 630-645.
- [38] QIN Z Y, WU J. Visual saliency maps can apply to facial expression recognition [EB/OL]. (2018-11-12) [2020-02-20]. <https://arxiv.org/abs/1811.04544>.
- [39] MIAO S, XU H Y, HAN Z Q, et al. Recognizing facial expressions using a shallow convolutional neural network [J]. IEEE Access, 2019, 7: 78000-78011.
- [40] ZHOU J C, JIA X, SHEN L L, et al. Improved softmax loss for deep learning-based face and expression recognition [J]. Cognitive Computation and Systems, 2019, 1(4): 97-102.
- [41] TIAN Y, WEN Z W, XIE W C, et al. Outlier-suppressed triplet loss with adaptive class-aware margins for facial expression recognition [C] // 2019 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2019: 46-50.
- [42] XIE S Y, HU H F, WU Y B. Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition [J]. Pattern Recognition, 2019, 92: 177-191.

作者简介:

兰凌强 男, 硕士研究生。主要研究方向: 计算机视觉。

李欣 男, 博士, 副教授, 硕士生导师。主要研究方向: 信息技术。

刘淇缘 女, 硕士研究生。主要研究方向: 计算机视觉。

卢树华 男, 博士, 副教授, 硕士生导师。主要研究方向: 安全防范技术。

Facial expression recognition method based on a joint normalization strategy

LAN Lingqiang, LI Xin, LIU Qiyuan, LU Shuhua*

(College of Police Information Technology and Cyber Security, People’s Public Security University of China, Beijing 102600, China)

Abstract: As for that end-to-end feature extraction and classification based on deep learning often used in facial expression recognition, a new method of depth model optimization has been proposed. This paper proposes the joint optimization strategies learned from ResNet18 residual network and normalization ideas, that is, filter response normalization and batch normalization, instance normalization and group normalization, as well as group normalization and batch normalization were embedded in the network, respectively, to balance and improve the distribution of feature data, make up for the shortcomings of single regularization, and improve model performance. The validation and test were carried out on the two public datasets FER2013 and CK+, and the highest accuracy rates are 73.558% and 94.9%, respectively. The experimental results indicate that the joint optimization strategy enhances the performance of the basic network, which is better than most of the latest facial expression recognition methods.

Keywords: expression recognition; joint strategy; filter response normalization; batch normalization; group normalization

Received: 2020-03-02; **Accepted:** 2020-04-09; **Published online:** 2020-04-15 11:00
URL: kns.cnki.net/kcms/detail/11.2625.V.20200415.1029.002.html
Foundation items: National Key R & D Program of China (2016YFC0801005); the Fundamental Funds for the Central Universities (2019JKF225)
* **Corresponding author.** E-mail: lushuhua@ppsuc.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn

DOI: 10.13700/j.bh.1001-5965.2020.0309

基于 U-Net 的掌纹图像增强与 ROI 提取



陆展鸿¹, 单鲁斌², 苏立循², 焦雨欣¹, 王家骅¹, 王海霞^{3,*}

(1. 杭州海康威视数字技术股份有限公司, 杭州 310051; 2. 浙江工业大学 信息工程学院, 杭州 310023;

3. 浙江工业大学 计算机科学与技术学院, 杭州 310023)

摘 要: 掌纹因稳定性、唯一性、难复制性及易获取等特点, 是极具应用潜力的生物识别特征。针对掌纹识别中, 获取掌纹感兴趣区域 (ROI) 与增强掌纹信息普遍存在的时间成本大, 方法之间有依赖关系等问题, 提出了一种基于 U-Net 神经网络结构的掌纹图像预处理方法。通过香港理工大学掌纹库进行实验对比, 结果表明, 所提方法能消除预处理方法之间的相互影响, 实现对掌纹图像的去噪与增强处理, 并能快速、高精度提取感兴趣区域。

关键词: U-Net; 感兴趣区域 (ROI) 增强; 深度学习; 感兴趣区域 (ROI) 提取; 掌纹
中图分类号: TP391.41

文献标志码: A

文章编号: 1001-5965(2020)09-1807-10

在当今信息化时代, 生物特征识别是鉴别个人身份、保护个人信息的有效手段。生物特征识别技术, 是指利用机器视觉、图像处理等技术对生物特征采样进行身份识别。现有的主要生物识别特征包括指纹^[1]、人脸^[2]、虹膜^[3]和掌纹^[4]等。由于掌纹的纹路是由每个人的基因所决定的, 即使后天受损, 重新生长出来的纹路也保持原有形状, 因此是一种具备相当大潜力的生物识别特征。掌纹识别技术是近几年发展起来的技术, 与其他识别技术的优势在于: 掌纹区域较大, 具有较好的稳定性; 掌纹采集对硬件的要求不高, 故而设备成本较低; 并且相比于其他生物特征识别技术, 安全性较高, 采集所需时间较短, 识别的效果较好^[5]。

掌纹原始图像一般存在纹理不清晰、掌纹图像具有不定旋转角度、图像存在噪声等问题。因此有必要对设备采集获取到的掌纹原始图像进行增强处理和感兴趣区域 (ROI) 提取^[6]。现有的掌纹预处理方法普遍存在着时间成本大, 方法之间有依赖关系等问题。随着近几年神经网络的迅速

发展, 神经网络在交通车辆^[7]、步态识别^[8]、车牌识别等领域取得了很大的成功。神经网络是模拟生物神经网络进行信息处理的一种数学模型, 其目的在于模拟大脑的某些机理与机制实现一些特定的功能。它有着高度的并行结构和并行实现能力, 能够充分发挥计算机的高速运算能力, 可以很快找到最优解。同时它也可以处理一些非线性问题, 并且有着自学习功能。近些年学者们也将神经网络引入掌纹识别领域, 进一步促进了掌纹识别的发展。

本文提出神经网络在掌纹预处理部分中的图像增强与 ROI 提取的运用, 采用的神经网络模型是 U-Net 神经网络^[9], U-Net 神经网络既可以处理分割问题, 亦能降噪增强图像信息, 它能够实现每个像素从点到点的处理。本文的贡献在于将掌纹预处理部分的图像增强和 ROI 提取放在同一个神经网络模型上运行, 这大大减少了预处理的时间, 去除了掌纹预处理部分之间的相互依赖关系。

收稿日期: 2020-07-01; 录用日期: 2020-07-22; 网络出版时间: 2020-08-07 16:13

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200807.1436.004.html

基金项目: 浙江省重点研发计划 (2019C01007)

* 通信作者。E-mail: hxwang@zjut.edu.cn

引用格式: 陆展鸿, 单鲁斌, 苏立循, 等. 基于 U-Net 的掌纹图像增强与 ROI 提取[J]. 北京航空航天大学学报, 2020, 46(9): 1807-1816. LU Z H, SHAN L B, SU L X, et al. Palmprint enhancement and ROI extraction based on U-Net[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(9): 1807-1816 (in Chinese).

1 相关工作

1.1 掌纹识别

现今常用的掌纹识别方法通常分为几个大类:①基于编码特征的识别方法^[10-11]。通过将掌纹图像的特征转化为由 0、1 组成的数字编码,最后通过二进制算术运算获得图像间的相似度。②基于结构特征的识别方法^[12-13]。通过边缘检测方法提取脊线、主线等方位信息来进行掌纹识别。③基于统计特征的识别方法^[14-15]。利用图像的统计概念提取方差、标准差等特征,然后转化为用于匹配的向量。④基于子空间特征的识别方法^[16-17]。将掌纹图像视为高维矩阵或向量,并通过投影或数学变换等方式转化为低维形式进行表征和分类,随后用于图像匹配。⑤基于特征学习的识别方法^[18-19]。在现有掌纹识别的方法中,掌纹主线是重要的特征,Huang 等^[20]提出 Radon 变换掌纹的主线,再利用掌纹之间的相似度进行掌纹验证。Wu 等^[21]设计了一组定向线检测器,然后将这些检测器用于沿不同方向提取主线。通过对主要纹路的增强,可以有效地增强掌纹识别过程中对主线的提取效果,从而提高掌纹识别的准确率。

1.2 掌纹图像预处理

掌纹图像增强预处理中,最常使用的方法有小波变换(Wavelet Transform, WT)^[22]、偏微分方程图像增强方法^[23]、三维块匹配方法(BM3D)^[24]等,然而这些方法在对每张图像进行处理时往往需要较多的时间成本,而且这些方法对不同环境采集的掌纹图像不能保持良好的鲁棒性。传统的 ROI 提取方法通过获取手掌的轮廓,找到指蹼的特征点,再通过特征点的位置进行 ROI 提取。李艳等^[6]提出了一种对掌纹轮廓特征点进行提取的方法,即利用多个互相相切的外切圆和手掌边界的交点个数来判断角点的大概位置,然后通过结合手掌轮廓线近似直线的性质,从而得到手掌轮廓特征点的准确位置。但因为需要反复搜索圆心并调整半径,预处理时间长,复杂性高,不能用于实时系统。张秀峰等^[25]为了改善 ROI 提取过程中会产生偏移的问题,提出了一种新的 ROI 分割方法:先确定手掌的指蹼位置,通过两个指蹼边界点拟合成一条直线,利用这条直线为基准,矫正手掌的偏转角度并确定 ROI 区域。这些传统方法的 ROI 提取对掌纹的质量要求非常高,而神经网络的出现,很好地解决了这些问题。

1.3 神经网络

在 2015 年 U-Net 被 Ronneberger 等^[9]所提出,其在全卷积神经网络(FCN)的基础上进行改良形成的,其特点是可以较少的图像训练数据得到更好的处理结果,因其具有的特点近年来被广泛应用于医学图像处理的领域。Meng 等^[26]应用 U-Net 实现了对脑肿瘤图像的分割和增强,并与其他方法进行对比,发现有着更好的效果。Qian 等^[27]将模型应用于图像增强,也实现了较好的处理效果。神经网络在掌纹识别中也取得了较好的成绩。Gao 等^[19]结合了自上而下的无监督训练和训练样本建立 Deep Belief Net,在掌纹识别上取得优于主成分分析(PCA)、局部二元模式(LBP)的结果。Liu 和 Sun^[28]尝试使用 AlexNet 提取深度特征。Sun 等^[29]利用 CNN-F 网络进行掌纹识别,并评估不同层提取到的卷积特征。Lin 等^[30]提出了 Palm RCNN 进行掌纹识别,并公开了同济大学掌纹数据库。王海霞等^[31]改进了轻量级 U-Net,在不影响输出结果的前提下,有效的提高了计算时间。特别地,周瞳等^[32]使用卷积神经网络的方法实现掌纹的提取。这些研究都证明了神经网络在掌纹图像处理中的巨大潜力。

2 方法设计

本文方法流程图如图 1 所示,通过 U-Net 神经网络分别对掌纹图像进行图像增强与谷点分割,在该部分掌纹增强与 ROI 区域提取采取并行模式,该模式可以有效地去除图像增强与 ROI 提取相互的依赖性,确保处理方法的独立性。在 2.1 节中,对数据集的准备进行了介绍,包括有监督的图像去噪滤波、图像增强、特征点定位及数据扩增。在 2.2 节中,采用 U-Net 神经网络对掌纹图像进行增强与谷点位置预测,以及相关的后续处理。

2.1 数据集的准备

数据集的准备具体步骤包括图像去噪滤波、图像增强、以 ROI 提取为目的的特征点定位,并进行数据扩增。

首先,掌纹图像分布着不均匀的噪声,掌纹的 3 条主线常常具有对比度不清晰的问题,并且掌纹的亮暗程度不均匀,需要对图像进行滤波处理。图像滤波就是对图像进行噪声上的过滤,抑制噪声对图像的影响,并且最大化地保留图像细节特征。本文采用 BM3D 方法,这是一种非局部方法,它的原理首先是通过和相邻图像块进行匹配,然后与之进行整合形成一个三维矩阵,在 3D 空间中

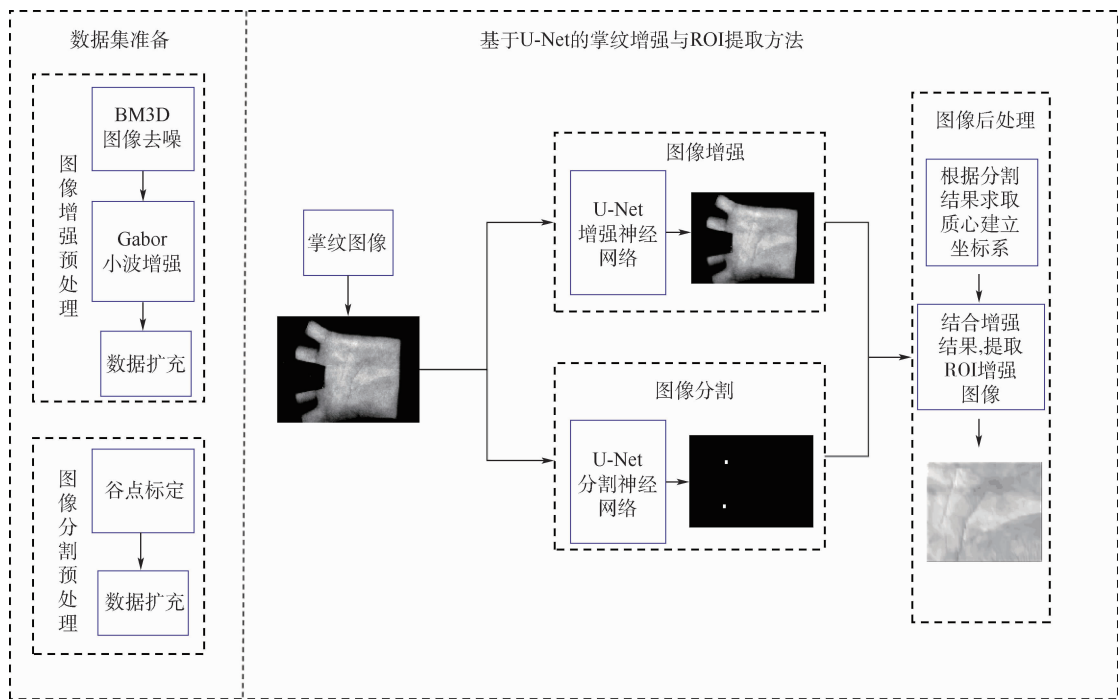


图 1 本文方法流程图
Fig.1 Proposed method flow chart

进行滤波,然后再转换为 2D。该方法分两步,实现对图像双重去噪:第一次是对原图匹配进行简单去噪,第二次去噪的权值将会以第一次去噪的结果作为基准,处理结果如图 2 所示。在实际去噪处理中,会不可避免地导致一些关键细节信息被模糊,如图 2(b)所示。掌纹的 3 条关键纹路与手掌的对比度明显降低。本文采用 Gabor 小波变换进行图像增强。Gabor 小波的优点在于对图像边缘检测敏感,从而能够有良好的尺度选择特性和方向选择,而且对光照变化不敏感,有着良好的适应性。Gabor 小波变换处理结果如图 2(c)所示。通过 BM3D 与 Gabor 小波变换,本文完成对掌纹增强数据集的处理与获取。

其次,针对掌纹 ROI 分割问题,为更精确定位 ROI 区域,本文通过对掌纹的 2 个关键指间谷点进行预测,再进行后处理定位的方式进行。数据集的标注是对关键指间谷点的标注,为尽量进行精确定位,本文采用手工标定。在标注出两个关键点之后,生成一副二值图,背景为黑,2 个谷点为白色,长度为 10 个像素的矩形,如图 2(d)所示。通过手工标注,完成对掌纹 ROI 提取数据集的获取。

由于深度学习需要较多的训练样本,而且对训练样本的处理也需要花费大量的时间。因此本文采用数据增强技术,以左右各旋转 10°来进行数据集扩充。在此次训练中,对 700 张掌纹增强图像、ROI 谷点图像进行扩充,获得 2 100 张图片。

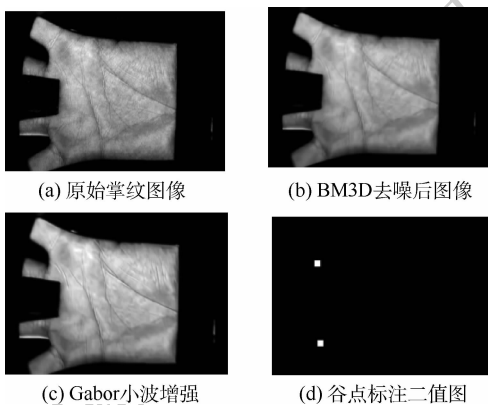


图 2 预处理步骤
Fig.2 Pre-processing steps

来防止因为数据集过少而产生的神经网络过拟合问题,获得更好的增强效果。在香港理工大学的掌纹数据库中,每一张图像的像素大小为 384 × 284。在网络训练的过程中,需要对图像进行 3 次下采样操作,每经过一次下采样处理,图像大小变为原来的一半,而 284 不能被 2 整除 3 次。所以调整图像大小为 384 × 280。

2.2 掌纹增强与 ROI 提取

2.2.1 网络结构

U-Net 神经网络通过收缩路径和扩张路径构成了一个对称的 U 型网络。主要包含下采样和上采样 2 个部分,下采样进行特征提取。下采样过程中,图像的通道数翻倍,同时特征图维度减半。应用修正线性单元(Rectified Linear Unit, ReLU)激活函数,激活函数的作用就是引进非线性因素,

这样使得神经网络可以应用于非线性的模型中去,而不是仅仅只是限制于线性函数。上采样的作用是把识别的特征还原到原图的大小上去,最终得出输出的结果。每次经过上采样层后特征图维度加倍,通道数减半,最终使图像恢复到原尺寸,在上采样的过程中同样也是应用 ReLU 激活函数。

图 3 和表 1 展示了网络结构和模型参数。本模型包含收缩路径和扩张路径。在收缩路径中,

包含 3 个收缩单元,每个单元都包含 2 个卷积模块和 1 个池化层。进行完下采样后再经过 2 个卷积模块,就可以进行上采样操作。扩张路径同样有 3 个扩张单元,前面 2 个单元包含 1 个上采样层和 2 个卷积模块。最后 1 个扩张单元中包含上采样层、两个卷积模块和 1 个 1×1 大小的卷积模块。每次使用上采样层后特征图维度加倍、通道数减半。跳跃连接增加了收缩路径和扩张路径之间的通路。

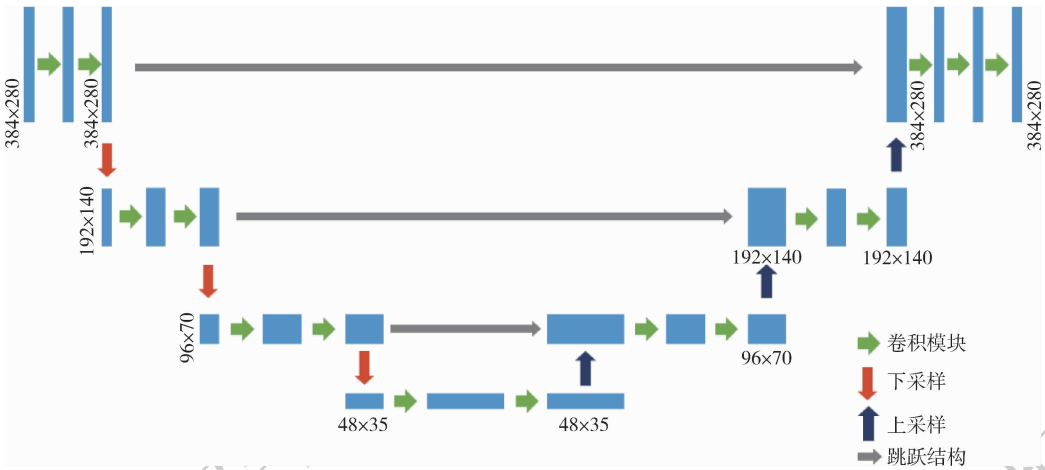


图 3 U-Net 神经网络结构
Fig. 3 U-Net neural network structure

表 1 U-Net 神经网络模型参数			
层类型	输入大小	滤波器大小	重复次数
卷积模块	$1 \times 384 \times 280$	3×3	1
卷积模块	$64 \times 384 \times 280$	3×3	1
池化层	$64 \times 384 \times 280$	2×2	1
卷积模块	$64 \times 192 \times 140$	3×3	2
池化层	$128 \times 192 \times 140$	2×2	1
卷积模块	$128 \times 96 \times 70$	3×3	2
池化层	$256 \times 96 \times 70$	2×2	1
卷积模块	$256 \times 48 \times 35$	3×3	2
上采样层	$512 \times 48 \times 35$	3×3	1
卷积模块	$512 \times 96 \times 70$	3×3	2
上采样层	$256 \times 96 \times 70$	3×3	1
卷积模块	$256 \times 192 \times 140$	3×3	2
上采样层	$128 \times 192 \times 140$	3×3	1
卷积模块	$128 \times 384 \times 280$	3×3	1
卷积模块	$64 \times 384 \times 280$	3×3	1
卷积模块	$64 \times 384 \times 280$	1×1	1

2.2.2 掌纹增强与 ROI 提取网络的能量函数设定
在深度学习中,激活函数和损失函数在神经网络中都起着重要作用,激活函数给网络引入了非线性因素,使得网络能够逼近任何的函数。常

用到的输出层激活函数主要包括: sigmoid、tanh、ReLU。不同的输出层激活函数的选择,将会产生不同的输出效果,因此需要根据各激活函数的特点进行选择。针对掌纹增强网络,本文经过实验对比分析,发现 3 种激活函数效果差异明显,当采用 ReLU 函数为激活函数,效果明显优于其他 2 种函数,因此 ReLU 选择作为掌纹增强的激活函数,它是分段函数,函数收敛速度也明显比前两者快,并且能避免 sigmoid 和 tanh 函数都存在的梯度消失的问题,因此会更加适合图像增强的处理,ReLU 函数图形如图 4(a) 所示,其计算公式为

$$g(z) = \begin{cases} z & z > 0 \\ 0 & z \leq 0 \end{cases} \quad (1)$$

机器学习过程中,需要对方法进行最小化的函数,这个函数就是损失函数。损失函数越小,表明真实值和预测值越接近,损失函数的选取要适应应用背景,没有适用所有情况的损失函数存在,常用的损失函数包含:均方误差、交叉熵误差、绝对值损失函数等。而在实际应用中也可根据自己的需求进行编写。

根据实验分析得出,本文应用的掌纹图像增强的损失函数 $loss_1$ 定义为式(2)所示,其最为接

近训练样本:

$$\text{loss}_1 = \frac{1}{N} \sum_{i=1}^N |(f_i - y_i)| \quad (2)$$

式中: N 为样本数量; f_i 为预测值; y_i 为真实值。当预测值越接近真实值时, loss_1 的返回值也越接近 0, 网络接近稳定。

针对掌纹 ROI 提取, 3 种激活函数效果差异明显, 本文采用 sigmoid 函数为激活函数。sigmoid 函数的本质是将取值为 $(-\infty, +\infty)$ 的值映射到 $(0, 1)$, 所以经常被用在二分类的问题中, sigmoid 函数图形如图 4(b) 所示, 其计算公式为

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3)$$

根据分割需求, 将本文应用的掌纹 ROI 提取的损失函数 loss_2 定义为

$$\text{loss}_2 = \frac{2 \sum_{i=1}^N (f_i \cdot y_i) + s}{\sum_{i=1}^N y_i + \sum_{i=1}^N f_i + s} \quad (4)$$

式中: s 为基值, 它的取值为 1, 它的作用是防止出现分母为 0 的状况。当预测值越接近真实值时, loss_2 的返回值也越接近 1, 网络接近稳定。

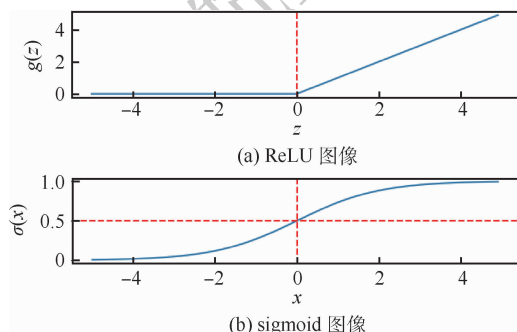
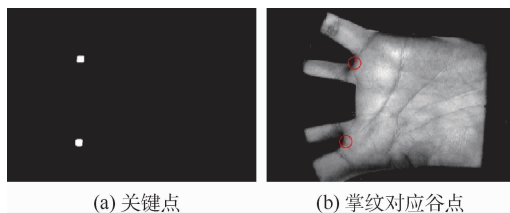


图 4 激活函数图形

Fig. 4 Graph of activation functions

2.2.3 ROI 提取后处理

通过 U-net 网络掌纹增强可直接获取结果, 而掌纹 ROI 提取只是获得中间过程图像, 对掌纹的分割还需要从指间谷点进行后处理。如图 5(a) 所示, 经过 U-Net 神经网络可得到了掌纹谷点分割结果。首先求取二值图中的 2 个关键点, 作为掌纹谷点, 如图 5(b) 掌纹图像中的圆圈标记所示。其次, 对 2 个谷点进行连接, 对连接线和相对应垂线建立新的坐标轴, 以 2 条直线交点为原点。如图 6(a) 所示, 以此坐标系为主坐标系, 2 个谷点为标准点, 调整掌纹图像角度, 对掌纹图像进行 ROI 框取, 框取的 ROI 大小为 144×128 , 如图 6(b) 所示, 最终效果图如图 6(c) 所示。



(a) 关键点 (b) 掌纹对应谷点

图 5 关键点及掌纹对应谷点

Fig. 5 Key points and correspondingly valley points on plamprint

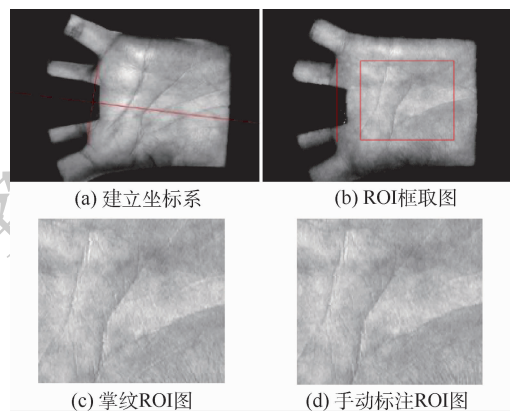


图 6 ROI 提取处理

Fig. 6 ROI extraction processing diagram

3 实验结果分析

本文实验应用的掌纹数据库是来自香港理工大学掌纹数据库^[33], 包含 7 752 个图像, 分别来自 386 个不同的手掌, 每个手掌大约采集了 20 张图像, 这 20 张图像分 2 个阶段进行采集, 这 2 个阶段间隔在 2 个月左右。本文从香港理工大学掌纹数据库中选取 1 000 张掌纹图像, 其中 700 张掌纹作为训练集进行神经网络训练, 另外选择 300 张掌纹图像组成测试集, 经过神经网络预测分别获得增强、分割图。以下将对该网络模型的性能进行评估。

3.1 掌纹图像增强处理结果

图 7 是部分原始掌纹测试集图像和网络增强处理后的 ROI 图像对比, 可以看出, 掌纹的噪声明显降低, 掌纹主要的 3 条纹路乳突纹、皱纹和屈肌线也更加的明显, 并且这种效果是普遍适用所有 300 张测试集图像的。

对掌纹图像增强效果进行量化评估, 常用峰值信噪比 (PSNR)、信息熵 (Entropy)、绝对均值误差 (AMBE) 来评估方法的性能。PSNR 的值越大, Entropy 值越大, AMBE 值越小, 则说明掌纹图像的增强效果越好。

将本文方法与 BM3D 去噪 + Gabor 小波增强、OBLNM^[34]、RKTFLT^[35] 和 SRAD^[36] 当前流行的掌纹图像增强方法的效果进行对比实验, 由于缺少真值, 本文采用 2 种方式获得真值。第 1 种

以 BM3D 去噪 + Gabor 小波增强处理过后的掌纹图像为真值,第 2 种以 OBLNM 处理过后的掌纹图像为真值。分别对 2 组真值图像添加高斯白噪声,再分别经过以上 5 种方法进行去噪增强。图 8

为第 1 种方式下掌纹图像增强效果对比。图 9 为第 2 种方式下掌纹图像增强效果对比。从表 2 和图像增强对比结果(图 8 和图 9)可以看出,本文方法要优于其他方法。

3.2 掌纹分割图像相似度分析

为了判断训练结束后的网络模型的科学性,将测试集 300 张掌纹图像进行手动标注指间谷点,并对标注好谷点后的图像进行 ROI 提取,将提取出来的掌纹感兴趣区域作为标准参考。同时采用其他 2 种方法——基于形态学角点检测的定位分割方法^[37](方法 1)与基于边缘跟踪的定位分割方法^[38](方法 2),与本文提出的基于 U-Net 神经网络输出的掌纹分割图像与标准图像进行面积重合率分析,通过将得到的谷点进一步计算成掌纹的感兴趣区域,划定位置,将掌纹图像转化为二值图像,掌纹 ROI 为 1,背景为 0,同理对其他方

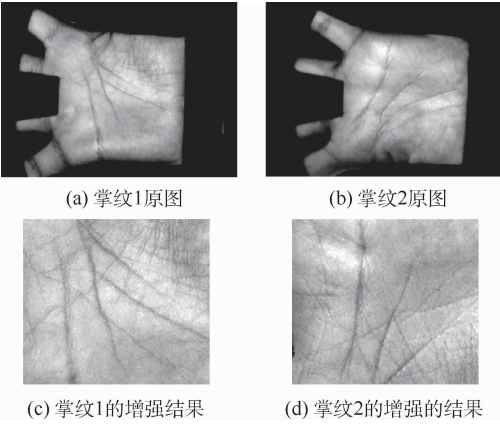


图 7 不同掌纹图像的增强结果

Fig. 7 Enhancement results of different palmprint images

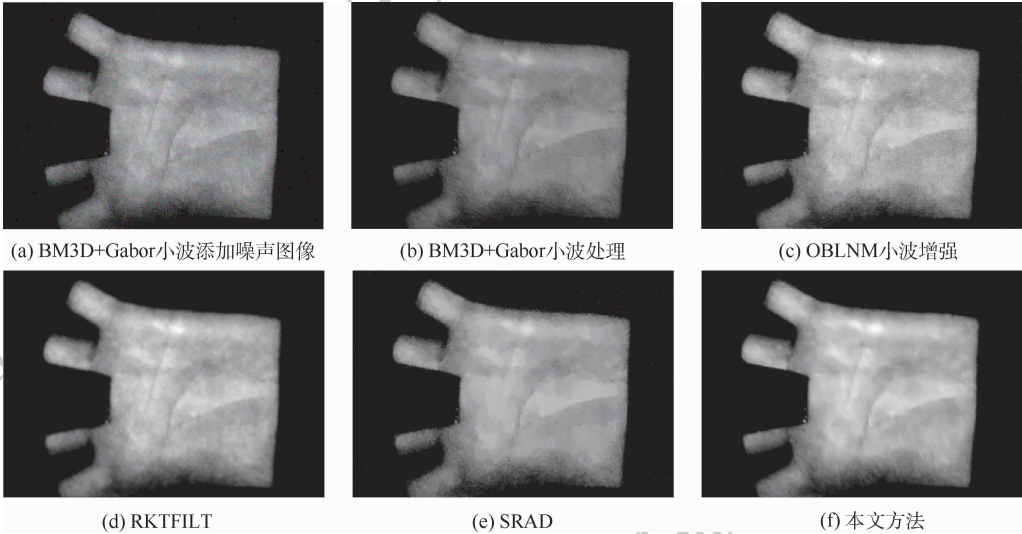


图 8 BM3D 去噪 + Gabor 小波增强后为真值的掌纹增强对比

Fig. 8 Comparison of palmprint enhancement after BM3D denoising + Gabor wavelet enhancement as truth value

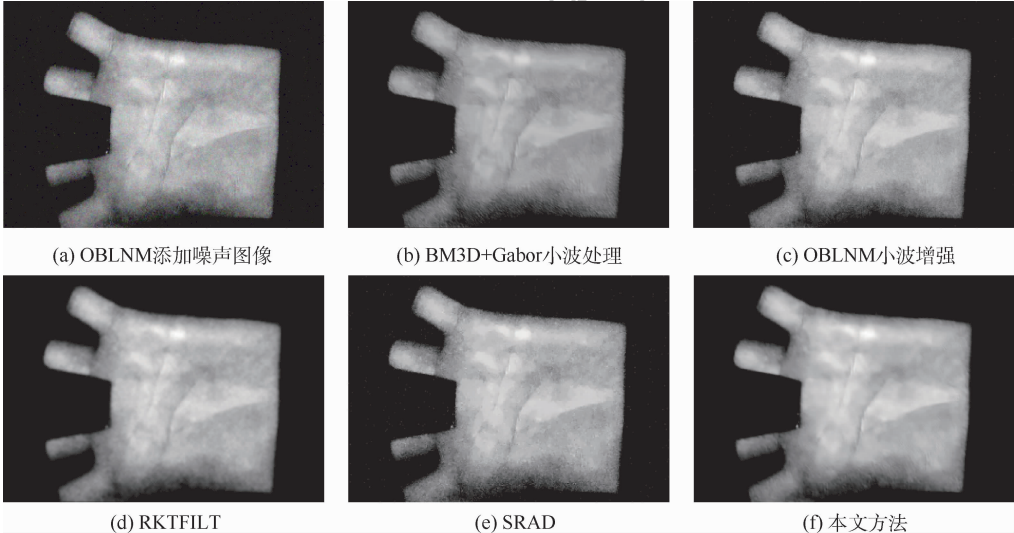


图 9 OBLNM 增强后为真值的掌纹增强对比

Fig. 9 Comparison of palmprint enhancement after OBLNM enhancement as truth value

表 2 掌纹增强效果比较结果

Table 2 Comparison results of palmprint enhancement

方法	BM3D 去噪 + Gabor 小波增强后为真值			OBLNM 增强后为真值		
	PSNR	Entropy	AMBE	PSNR	Entropy	AMBE
BM3D 去噪 + Gabor 小波增强	24.727 6	6.435 6	3.163 0	24.049 3	6.531 6	4.168 1
OBLNM	23.171 6	6.916 0	12.419 8	26.334 8	6.877 0	5.930 9
RKTFILT	23.499 4	6.820 0	15.239 3	28.022 2	6.753 7	6.806 8
SRAD	26.925 8	6.005 4	7.346 4	27.037 8	6.154 6	5.653 2
本文方法	27.120 8	7.247 6	3.165 3	27.916 9	6.920 7	3.210 1

法和人工标注也进行同样处理,最后对两者进行与操作,计算重叠部分所占比例。结果如表 3 所示,本文方法面积重合率可达 94.00%,而方法 1 的为 91.18%,方法 2 的为 88.00%。可见本文方法的结果和手动标注的真值有着较高的重合度,更加接近真实感兴趣区域,也意味着本文方法定位精度是可靠的。

为了进一步判断本文提出的掌纹分割方法准确性,将基于形态学角点检测的定位分割方法、基于谷点聚类分析的定位分割方法和本文方法与手动标注分割的掌纹感兴趣区域图像进行相似性比较,本文选取 2 个相似度指标,分别为余弦相似度(Cosin 相似度)和哈希相似度^[39]。当两者的值越接近 100% 时,表示 2 张图像越相似。结果如表 4 所示。从表中可以看出,本文方法相比于另外 2 种掌纹分割方法有着更高的相似度。实验结果表明,基于 U-Net 神经网络掌纹分割有着较好的效果,对掌纹感兴趣区域有着精确的定位。

表 3 掌纹感兴趣区域重合度分析

Table 3 Analysis of coincidence degree of palmprint ROI

方法	面积重合率/%
方法 1	91.18
方法 2	88.00
本文方法	94.00

表 4 掌纹分割图像相似度对比

Table 4 Similarity comparison of palmprint segmentation images

方法	Cosin 相似度/%	哈希相似度/%
方法 1	99.32	73.17
方法 2	99.13	71.61
本文方法	99.49	76.47

3.3 方法运行时间比较

为了验证神经网络方法处理速度更快的特点,本文对增强和分割时间进行统计。该实验以测试集 300 张掌纹图像的处理进行测试,实验环境为 Linux 操作系统,基于 tesla 和 quadro 的硬件平台。本文所提出的增强、分割方法均通过 python 编写的 U-Net 模型进行,分割后续处理部分

运用 MATLAB 程序实现,对两者的处理时间进行统计;进行比较的增强方法为 BM3D 去噪 + Gabor 小波增强方法,由 MATLAB 程序实现;分割的比较方法分别为基于形态学角点检测的定位分割方法、基于边缘跟踪的定位分割方法,由 MATLAB 程序实现。统计结果如表 5 所示,每张掌纹图像经过 BM3D 去噪 + Gabor 小波增强处理所需时间为 1 355.73 ms,基于形态学角点检测的定位分割方法进行 ROI 分割则需要 32.23 ms,完成增强分割操作总共需要 1 387.96 ms。基于边缘跟踪的定位分割方法进行 ROI 分割则需要 14.49 ms,完成增强分割操作总共需要 1 370.22 ms,而本文提出的基于 U-Net 神经网络的方法完成每张掌纹图像增强和分割仅需 90.49 ms,因此可以看出基于 U-Net 神经网络处理效率明显优于常规方法。

表 5 每张掌纹图像处理时间

Table 5 Processing time of each palmprint image

方法	时间/ms		
	增强	分割	增强和分割
方法 1	1 355.73	32.23	1 387.96
方法 2	1 355.73	14.49	1 370.22
本文方法			90.49

4 结束语

本文提出了一种基于 U-Net 神经网络的掌纹增强与分割方法。掌纹图像增强中,所提方法能有效去除掌纹图像的噪声,并达到对掌纹 3 条主要纹路增强效果,性能优于其他方法。在掌纹感兴趣区域分割中,对比手动标注谷点分割的方法,所提方法比常规掌纹分割方法有着更高的定位精度。本文还对神经网络处理掌纹图像和常规方法处理掌纹图像进行时间对比,对比发现神经网络有着更快的处理速度,因此基于 U-Net 神经网络对掌纹图像进行增强和 ROI 分割是具有高效性。

参考文献 (References)

[1] KUMAR A,KWONG C. Towards contactless, low-cost and ac-

- curate 3D fingerprint identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 681-696.
- [2] LU J W, LIONG V E, ZHOU X Z, et al. Learning compact binary face descriptor for face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(10): 2041-2056.
- [3] NGUYEN K, FOOKES C, JILLELA R, et al. Long range iris recognition: A survey[J]. Pattern Recognition, 2017, 72(12): 123-143.
- [4] XIANG B H, YU Z X, QU H B. An improved algorithm of principal line extraction and registration based on MFRAT and ICP[J]. Pattern Recognition and Artificial Intelligence, 2015, 28(3): 260-265.
- [5] WONG M, ZHANG D, KONG W K, et al. Real-time palmprint acquisition system design[J]. IEE Proceedings Vision, Image and Signal Processing, 2005, 152(5): 527-534.
- [6] 李艳, 吴贵芳, 戴高乐, 等. 一种新的掌纹轮廓特征点提取算法[J]. 微电子学与计算机, 2010, 27(5): 90-94.
LI Y, WU G F, DAI G L, et al. A new algorithm to extract contour feature points of palmprint[J]. Microelectronics and Computer, 2010, 27(5): 90-94 (in Chinese).
- [7] 卢妍辉. 智能交通系统中车辆路径识别方法[J]. 交通世界, 2018(35): 15-16.
LU Y H. Vehicle route recognition method in intelligent transportation system[J]. Transpoworld, 2018(35): 15-16 (in Chinese).
- [8] 张馨心, 姚爱琴, 孙运强, 等. 基于深度学习的步态识别算法优化研究[J]. 自动化与仪表, 2020, 35(4): 70-74.
ZHANG X X, YAO A Q, SUN Y Q, et al. Research on optimization of gait recognition algorithm based on deep learning[J]. Automation & Instrumentation, 2020, 35(4): 70-74 (in Chinese).
- [9] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation [EB/OL]. (2015-05-18) [2020-06-28]. <https://arxiv.org/abs/1505.04597>.
- [10] YUE F, ZUO W M, ZHANG D, et al. Orientation selection using modified FCM for competitive code-based palmprint recognition[J]. Pattern Recognition, 2009, 42(11): 2841-2849.
- [11] SHEN L L, WU S P, ZHENG S H, et al. Embedded palmprint recognition system using OMAP 3530[J]. Sensors, 2012, 12(2): 1482-1493.
- [12] LI C, LIU F, ZHANG Y Z. A principal palm-line extraction method for palmprint images based on diversity and contrast[C]//Proceedings of the 3rd International Congress on Image and Signal Processing. Berlin: Springer, 2010: 1772-1777.
- [13] ALI M M H, YANAWAR P, GAIKWAD A T. Study of edge detection methods based on palmprint lines[C]//Proceeding of the International Conference on Electrical, Electronics, and Optimization Techniques. Berlin: Springer, 2016: 1344-1350.
- [14] 许学斌, 张德运, 张新曼, 等. 基于离散曲波变换和支持向量机的掌纹识别方法[J]. 红外与毫米波学报, 2009, 28(6): 456-460.
XU X B, ZHANG D Y, ZHANG X M, et al. Palmprint recognition method based on discrete curvelet transform and support vector machine[J]. Journal of Infrared and Millimeter Waves, 2009, 28(6): 456-460 (in Chinese).
- [15] XU X B, LU L B, ZHANG X M, et al. Multispectral palmprint recognition using multiclass projection extreme learning machine and digital shearlet transform[J]. Neural Computing and Applications, 2016, 27(1): 143-153.
- [16] 吴婕, 任江涛. 基于子空间特征融合的两级掌纹识别算法[J]. 计算机工程与应用, 2015, 51(11): 175-178.
WU J, REN J T. Two level palmprint recognition algorithm based on subspace features[J]. Computer Engineering and Applications, 2015, 51(11): 175-178 (in Chinese).
- [17] 李新春, 马红艳, 林森. 基于子空间与纹理特征融合的掌纹识别[J]. 激光与光电子学进展, 2019, 56(7): 147-155.
LI X C, MA H Y, LI S. Palmprint recognition based on subspace and texture feature fusion[J]. Laser & Optoelectronics Progress, 2019, 56(7): 147-155 (in Chinese).
- [18] IZADPANAHKAKHK M, RAZAVI S, TAGHIPOUR-GORJIKOL-AIE M, et al. Deep region of interest and feature extraction models for palmprint verification using convolutional neural networks transfer learning[J]. Applied Sciences, 2018, 8(7): 1210.
- [19] GAO X J, LUO X L, PAN X, et al. Palmprint recognition based on deep learning[C]//Proceedings of the International Conference on Wireless Mobile and Multi-media. Berlin: Springer, 2015: 212-214.
- [20] HUANG D S, JIA W, ZHANG D. Palmprint verification based on principal lines[J]. Pattern Recognition, 2008, 41(4): 1316-1328.
- [21] WU X, ZHANG D, WANG K. Palm line extraction and matching for personal authentication[J]. IEEE Transactions on Systems Man & Cybernetics Part A: Systems & Humans, 2006, 36(5): 978-987.
- [22] 谢杰成, 张大力, 徐文立. 小波图象去噪综述[J]. 中国图象图形学报, 2002, 7(3): 3-11.
XIE J C, ZHANG D L, XU W L. Overview on wavelet image denoising[J]. Journal of Image and Graphics, 2002, 7(3): 3-11 (in Chinese).
- [23] 孔平. 基于偏微分方程的图像去噪与增强研究[D]. 南京: 南京理工大学, 2012.
KONG P. Research on image denoising and enhancement based on partial differential equation[D]. Nanjing: Nanjing University of Science and Technology, 2012 (in Chinese).
- [24] DANIELYAN A, KATKOVNIK V, EGIAZARIAN K. BM3D frames and variational image deblurring[J]. IEEE Transactions on Image Processing, 2012, 21(4): 1715-1728.
- [25] 张秀峰, 张真林, 谢红. 掌纹 ROI 分割算法的研究与实现[J]. 计算机科学, 2016, 43(S2): 170-173.
ZHANG X F, ZHANG Z L, XIE H. Research and realization of palmprint ROI segmentation algorithm[J]. Computer Science, 2016, 43(S2): 170-173 (in Chinese).
- [26] MENG Z, FAN Z Y, ZHAO Z C, et al. ENS-Unet: End-to-end noise suppression U-Net for brain tumor segmentation[C]//Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Piscataway: IEEE

- Press,2018:5886-5889.
- [27] QIAN P, LI A J, LIU M H. Latent fingerprint enhancement based on denseUNet [C] // 2019 International Conference on Biometrics (ICB). Piscataway: IEEE Press, 2019: 1-6.
- [28] LIU D, SUN D M. Contactless palmprint recognition based on convolutional neural network [C] // Proceedings of the IEEE International Conference on Signal Processing. Piscataway: IEEE Press, 2016: 1363-1367.
- [29] SUN Q L, ZHANG J X, YANG A Q, et al. Palmprint recognition with deep convolutional features [J] // Proceedings of the Chinese Conference on Image and Graphics Technologies. Berlin: Springer, 2017: 12-19.
- [30] LIN Z, CHENG Z X, YING S, et al. Palmprint and palmvein recognition based on DCNN and a new large-scale contactless palmvein dataset [J]. Symmetry, 2018, 10(4): 78.
- [31] 王海霞, 杨熙丞, 梁荣华, 等. 基于轻量级 U-Net 神经网络提取 OCT 内指纹和外指纹 [J]. 高技术通讯, 2019 (11): 1063-1072.
- WANG H X, YANG X C, LIANG R H, et al. OCT internal and external fingerprints generation based on a lightweight U-Net neural network [J]. Chinese High Technology Letters, 2019 (11): 1063-1072 (in Chinese).
- [32] 周瞳, 黄晨昕, 胡锐. 基于深度卷积网络的主掌线纹区域的提取 [J]. 工业控制计算机, 2018, 31(7): 97-98.
- ZHOU T, HUANG C X, HU R. Extracting main lines area on palm based on deep convolutional network [J]. Industrial Control Computer, 2018, 31(7): 97-98 (in Chinese).
- [33] PolyU Palmprint Database [EB/OL]. (2020-08-06) [2020-06-28]. <http://www.comp.polyu.edu.hk/~biomet->rics/.
- [34] COUPÉ P, HELLIER P, KERVRANN C. Nonlocal means-based speckle filtering for ultrasound images [J]. IEEE Transactions on Image Processing, 2009, 18(10): 2221-2229.
- [35] ROGOWSKA J, BRYANT C M, BREZINSKI M E. Cartilage thickness measurements from optical coherence tomography [J]. Journal of the Optical Society of America A: Optics, Image Science, and Vision, 2003, 20(2): 357-367.
- [36] YU Y J, ACTON S T. Speckle reducing anisotropic diffusion [J]. IEEE Transactions on Image Processing, 2002, 11(11): 1260-1270.
- [37] 刘佳雯, 陈勇. 关于掌纹图像基于形态学的角点检测的定位分割方法 [J]. 现代计算机, 2016(24): 35-38.
- LIU J W, CHEN Y. Research on go chess game image recognition algorithm [J]. Modern Computer, 2016(24): 35-38 (in Chinese).
- [38] ZHANG D, KONG W K, YOU J, et al. Online palmprint identification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(9): 1041-1050.
- [39] VENKATESAN R, KOON S M, JAKUBOWSKI M H, et al. Robust image hashing [C] // IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2000: 664-666.

作者简介:

陆展鸿 男, 学士。主要研究方向: 图像识别技术。

王海霞 女, 博士, 副教授, 博士生导师。主要研究方向: 精密测量与图像处理。

Palmprint enhancement and ROI extraction based on U-Net

LU Zhanhong¹, SHAN Lubin², SU Lixun², JIAO Yuxin¹, WANG Jiahua¹, WANG Haixia^{3,*}

(1. Hangzhou Hikvision Digital Technology Co., Ltd., Hangzhou 310051, China;

2. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China;

3. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: Palmprint is a biometric feature with great application potential due to its stability, uniqueness, hardness in copying, and easy access. In the palmprint recognition, palmprint Region of Interest (ROI) acquisition and palmprint enhancement usually have the problems of high time cost and high dependency between methods. This paper proposes a palmprint preprocessing method based on U-Net neural network structure. The experiments are carried out on palmprint database from Hong Kong Polytechnic University. The results show that the proposed method can eliminate the mutual influence between the preprocessing methods. Both denoising and enhancement of palmprint images are realized, and the region of interest can be extracted quickly and accurately; plamprint

Keywords: U-Net; Region of Interest (ROI) enhancement; deep learning; Region of Interest (ROI) extraction; plamprint

《北京航空航天大学学报》征稿简则

《北京航空航天大学学报》是北京航空航天大学主办的以航空航天科学技术为特色的综合性自然科学学术期刊(月刊)。本刊以反映航空航天领域研究成果与动态、促进学术交流、培养科技人才和推动科技成果向社会生产力转化为办刊宗旨。本刊为中国自然科学技术核心期刊,并被 Ei Compendex 等国内外权威文献检索系统收录。本刊向国内外公开发行人,为进一步提高办刊质量和所刊出文章的学术水平,特制定本简则。

1 论文作者及内容

1.1 本刊面向海内外所有学者。

1.2 主要刊载与航空航天科学技术有关的材料科学及工程、飞行器设计与制造、宇航科学与工程、信息与电子技术、控制技术和自动化工程、流体力学和动力工程、计算机科学及应用技术、可靠性工程与失效分析等领域的研究文章。航空航天科学技术民用方面以及具有航空航天工程背景的应用数学、应用物理、应用力学和工程管理等方面的文章也在本刊优先考虑之列。

2 来稿要求

2.1 论文应具有创新性、科学性、学术性和可读性。

2.2 论文为原创作品,尚未公开发表过,并且不涉及泄密问题。若发生侵权或泄密问题,一切责任由作者承担。

2.3 主题明确,数据可靠,图表清晰,逻辑严谨,文字精练,标点符号正确。

2.4 文稿撰写顺序:中文题名(一般不超过 20 个汉字),作者中文姓名、单位、所在城市、邮政编码,中文摘要(包括目的、方法、结果及结论),中文关键词(5~8 个),中图分类号,英文题名,作者英文姓名、单位、所在城市、邮政编码、国别,英文摘要,英文关键词,引言,正文,参考文献。首页下角注明基金项目名称及编号,作者信息。

2.5 作者请登录本刊网页进行在线投稿。

3 稿件的审核、录用与版权

3.1 来稿须经专家两审和主编、编委讨论后决定刊用与否。

3.2 若来稿经审查后认定不宜在本刊发表,将及时告知作者。如果在投稿满 3 个月后仍未收到本刊任何通知,作者有权改投它刊。在此之前,请勿一稿多投,否则一切后果自负。

3.3 来稿一经刊登,即赠送单行本。

3.4 来稿一经作者签字并在本刊刊出,即表明所有作者都已经认可其版权转至本刊编辑部。本刊在与国内外文献数据库或检索系统进行交流及合作时,不再征询作者意见。

邮寄地址:100191 北京市海淀区学院路 37 号 北京航空航天大学学报编辑部

办公地点:北京航空航天大学办公楼 405,407,409 房间

电 话:(010)82315594,82338922,82314839,82315426

E-mail: jbuaa@buaa.edu.cn

http://bhxb.buaa.edu.cn

http://www.buaa.edu.cn

《北京航空航天大学学报》 第五届编辑委员会

主 任 (主 编): 赵沁平

(以下按姓氏笔划为序)

副主任 (副主编): 丁希仑 王少萍 孙志梅 李秋实 李焕喜 杨嘉陵
苗俊刚 相 艳 徐立军 钱德沛 曹晋滨
编 委: 马殿富 王 琪 王 聪 邓小燕 王青云 王荣明 刘 宇
刘 红 江 洁 刘 强 闫 鹏 朱天乐 刘铁钢 齐铂金
陈万春 邹正平 苏东林 杨世春 沈成平 邱志平 宋知人
杨树斌 张晓林 杨晓奕 杨继萍 李惠峰 吴新开 张瑞丰
杨照华 宋凝芳 周 锐 林宇震 林贵平 战 强 姚仰平
胡庆雷 赵秋红 段海滨 赵巍胜 席 平 郭 宏 徐 洁
徐世杰 郭洪波 康 锐 翟 锦 熊华钢

北京航空航天大学学报

Beijing Hangkong Hangtian Daxue Xuebao

(原《北京航空学院学报》)

(月刊 1956 年创刊)

第 46 卷 第 9 期 2020 年 9 月

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS (JBUA)

(Monthly, Started in 1956)

Vol.46 No.9 September 2020

主管单位 中华人民共和国工业和信息化部
主办单位 北京航空航天大学
主 编 赵沁平
编辑出版 《北京航空航天大学学报》编辑部
邮 编 100083
地 址 北京市海淀区学院路 37 号
印 刷 北京科信印刷有限公司
发 行 北航文化传媒集团
发行范围 国内外发行
联系电话 (010) 82315594 82338922
82314839
电子信箱 jbuaa@buaa.edu.cn

Administrated by Ministry of Industry and Information
Technology of the People's Republic of China
Sponsored by Beijing University of Aeronautics
and Astronautics (BUAA)
(Beijing 100083, P. R. China)
Chief Editor ZHAO Qingping
Edited and Published by Editorial Board of JBUA
Printed by Beijing Kexin Printing Co., Ltd.
Distributed by BUAA Culture Media Group Limited
Telephone (010) 82315594 82338922
82314839
E-mail jbuaa@buaa.edu.cn
http://bhxb.buaa.edu.cn

刊 号 ISSN 1001-5965
CN 11-2625/V
国内定价 50.00 元 / 期

ISSN 1001-5965

