



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



2021-03

Vol.47 No.3

目 次

基于依存句法的图像描述文本生成	毕健旗, 刘茂福, 胡慧君, 代建华 (431)
基于多特征图像视觉显著性的视频摘要化生成	金海燕, 曹甜, 肖聪, 肖照林 (441)
基于混合网络的异源遥感图像变化检测	周圆, 李祥瑞, 杨晶 (451)
融合深度特征的输电线路金具缺陷因果分类方法	赵振兵, 张薇, 戚银城, 翟永杰, 赵文清 (461)
基于语义相关的多模态社交情感分析	胡慧君, 冯梦媛, 曹梦丽, 刘茂福 (469)
基于深度多模态特征融合的短视频分类	张丽娟, 崔天舒, 井佩光, 苏育挺 (478)
基于 FPGA 无人机影像快速低功耗高精度三维重建	李杰, 李一轩, 吴天生, 王昊榕, 梁敏 (486)
融合句嵌入的 VAACGAN 多对多语音转换	李燕萍, 曹盼, 石杨, 张燕 (500)
一种时空特征聚合的水下珊瑚礁鱼检测方法	陈智能, 史存存, 李轩涯, 贾彩燕, 黄磊 (509)
面向电力系统的多粒度隐患检测方法	徐晓华, 钱平, 王一达, 周昕悦, 徐汉麟, 徐李冰 (520)
电力系统厂站接线图拓扑关系检测技术	李昊, 管萸, 王杉, 石玮, 刘子鑫, 刘晓川 (531)
基于深度学习和图匹配的接线图检测与校核	李昊, 王杉, 耿玉杰, 王黎, 孙文昌, 苗纯源 (539)
基于记忆关联学习的小样本高光谱图像分类方法	王聪, 张锦阳, 张磊, 魏巍, 张艳宁 (549)
低延迟视频编码技术	宋利, 刘孝勇, 武国庆, 朱辰, 黄琰, 解蓉, 张文军 (558)
显著性引导的低光照人脸检测	李可夫, 钟汇才, 高兴宇, 翁超群, 陈振宇, 李勇周, 王师峥 (572)
基于对抗和迁移学习的灾害天气卫星云图分类	张敏靖, 白琮, 张敬林, 郑建炜 (585)
基于多语义线索的跨模态视频检索算法	丁洛, 李逸凡, 于成龙, 刘洋, 王轩, 漆舒汉 (596)
基于注意力机制的跨分辨率行人重识别	廖华年, 徐新 (605)
基于多尺度联合学习的行人重识别	谢彭宇, 徐新 (613)
基于块对角投影表示的人脸识别	刘保龙, 王勇, 李丹萍, 王磊 (623)
基于噪声柯西分布的社交图像标签优化与标注	练连荣, 项欣光 (632)
结合颅骨形态特征与神经网络的民族判别	孙慧杰, 赵俊莉, 郑鑫, 热孜万古丽·夏米西丁, 李奕, 周明全 (641)
融合语义信息的视频摘要生成	滑蕊, 吴心筱, 赵文天 (650)
基于无监督学习的多模态可变形配准	马腾宇, 李孜, 刘日升, 樊鑫, 罗钟铤 (658)
一种结合全局和局部相似性的小样本分割方法	刘宇轩, 孟凡满, 李宏亮, 杨嘉莹, 吴庆波, 许林峰 (665)

CONTENTS

Image captioning based on dependency syntax	BI Jianqi, LIU Maofu, HU Huijun, DAI Jianhua (431)
Video summary generation based on multi-feature image and visual saliency	JIN Haiyan, CAO Tian, XIAO Cong, XIAO Zhaolin (441)
Heterogeneous remote sensing image change detection based on hybrid network	ZHOU Yuan, LI Xiangrui, YANG Jing (451)
Causal classification method of transmission lines fitting defect combined with deep features	ZHAO Zhenbing, ZHANG Wei, QI Yincheng, ZHAI Yongjie, ZHAO Wenqing (461)
Multimodal social sentiment analysis based on semantic correlation	HU Huijun, FENG Mengyuan, CAO Mengli, LIU Maofu (469)
Deep multimodal feature fusion for micro-video classification	ZHANG Lijuan, CUI Tianshu, JING Peiguang, SU Yuting (478)
Fast, low-power and high-precision 3D reconstruction of UAV images based on FPGA	LI Jie, LI Yixuan, WU Tiansheng, WANG Haorong, LIANG Min (486)
Many-to-many voice conversion with sentence embedding based on VAACGAN	LI Yanping, CAO Pan, SHI Yang, ZHANG Yan (500)
An underwater coral reef fish detection approach based on aggregation of spatio-temporal features	CHEN Zhineng, SHI Cuncun, LI Xuanya, JIA Caiyan, HUANG Lei (509)
Multi-granularity hazard detection method for electrical power system	XU Xiaohua, QIAN Ping, WANG Yida, ZHOU Xinyue, XU Hanlin, XU Libing (520)
Topological relation detection technology of substation wiring diagram in electric power system	LI Hao, GUAN Ti, WANG Shan, SHI Wei, LIU Zixin, LIU Xiaochuan (531)
Wiring diagram detection and check based on deep learning and graph matching	LI Hao, WANG Shan, GENG Yujie, WANG Li, SUN Wenchang, MIAO Chunyuan (539)
Small sample hyperspectral image classification method based on memory association learning	WANG Cong, ZHAGN Jinyang, ZHANG Lei, WEI Wei, ZHANG Yanning (549)
Low-latency video coding techniques	SONG Li, LIU Xiaoyong, WU Guoqing, ZHU Chen, HUANG Yan, XIE Rong, ZHANG Wenjun (558)
Saliency guided low-light face detection	LI Kefu, ZHONG Huicai, GAO Xingyu, WENG Chaoqun, CHEN Zhenyu, LI Yongzhou, WANG Shizheng (572)
Classification of satellite cloud images of disaster weather based on adversarial and transfer learning	ZHANG Minjing, BAI Cong, ZHANG Jinglin, ZHENG Jianwei (585)
Cross-modal video retrieval algorithm based on multi-semantic clues	DING Luo, LI Yifan, YU Chenglong, LIU Yang, WANG Xuan, QI Shuhan (596)
Cross-resolution person re-identification based on attention mechanism	LIAO Huanian, XU Xin (605)
Multi-scale joint learning for person re-identification	XIE Pengyu, XU Xin (613)
Block-diagonal projective representation for face recognition	LIU Baolong, WANG Yong, LI Danping, WANG Lei (623)
Social image tag refinement and annotation based on noise Cauchy distribution	LIAN Lianrong, XIANG Xinguang (632)
Ethnic identification by combining features of skull morphology with neural network	SUN Huijie, ZHAO Junli, ZHENG Xin, REZIWANGULI Xiamixiding, LI Yi, ZHOU Mingquan (641)
Video summarization by learning semantic information	HUA Rui, WU Xinxiao, ZHAO Wentian (650)
Multimodal deformable registration based on unsupervised learning	MA Tengyu, LI Zi, LIU Risheng, FAN Xin, LUO Zhongxuan (658)
A few shot segmentation method combining global and local similarity	LIU Yuxuan, MENG Fanman, LI Hongliang, YANG Jiaying, WU Qingbo, XU Linfeng (665)

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

DOI: 10.13700/j.bh.1001-5965.2020.0443

基于依存句法的图像描述文本生成



毕健旗^{1,2}, 刘茂福^{1,2,*}, 胡慧君^{1,2}, 代建华³

(1. 武汉科技大学 计算机科学与技术学院, 武汉 430065;

2. 武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室, 武汉 430081;

3. 湖南师范大学 智能计算与语言信息处理湖南省重点实验室, 长沙 410081)

摘 要: 现有图像描述文本生成模型能够应用词性序列和句法树使生成的文本更符合语法规则,但文本多为简单句,在语言模型促进深度学习模型的可解释性方面研究甚少。将依存句法信息融合到深度学习模型以监督图像描述文本生成的同时,可使深度学习模型更具可解释性。图像结构注意力机制基于依存句法和图像视觉信息,用于计算图像区域间关系并得到图像区域关系特征;融合图像区域关系特征和图像区域特征,与文本词向量通过长短期记忆网络(LSTM),用于生成图像描述文本。在测试阶段,通过测试图像与训练图像集的内容关键词,计算2幅图像的内容重合度,间接提取与测试图像对应的依存句法模板;模型基于依存句法模板,生成多样的图像描述文本。实验结果验证了模型在改善图像描述文本多样性和句法复杂度方面的能力,表明模型中的依存句法信息增强了深度学习模型的可解释性。

关 键 词: 图像描述文本生成; 依存句法; 图像结构注意力; 内容重合度; 深度模型可解释性

中图分类号: TP37

文献标志码: A

文章编号: 1001-5965(2021)03-0431-10

近年来,面向跨模态的图像描述文本生成因其在计算机视觉和自然语言处理交叉领域的重要性,已引起越来越多的研究者的关注。给定一幅图像,图像描述文本生成旨在生成语法正确且可精准描述图像内容的文本。基于注意力机制的编解码框架已被广泛应用于图像描述文本生成^[1],其提取有价值的视觉特征,从而使生成的文本与图像的内容紧密相关。基于视觉文本注意力机制的编解码框架虽能较精确地发现图像视觉区域信息,但仍存在2个问题:①在编码阶段生成当前时刻图像的视觉区域信息时,仅考虑当前时刻前的文本信息,忽略了图像本身蕴含的视觉结构信息;然而此类信息难以在输入文本中发现,在缺少图

像结构信息的情况下,模型无法很好地发现图像中存在的潜在语义信息,也无法发现图像区域与文本的对应关系,进而不能生成准确的描述文本。②在解码阶段的长短期记忆网络(LSTM)中,生成文本所需的语法规则只能从其自编码机制中获得,而由于记忆局限性,LSTM更易于捕获图像特征而不能很好地学习到长距离的词与词间的依赖关系,使生成的文本多以简单句的形式存在。

为了使模型更好地发现图像结构信息,Lu等^[2]使用Faster R-CNN^[3-4]来识别图像实体,再进而根据文本信息来建立图像中实体间关系。为了使模型更好地学习语法规则,Aditya^[5]和Wang^[6]等将词性序列和句法树等语法信息作为模型的先

收稿日期: 2020-08-21; 录用日期: 2020-09-05; 网络出版时间: 2020-09-29 13:35

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200929.1427.003.html

基金项目: 国家社会科学基金重大研究计划(11&ZD189); 全军共用信息系统装备预先研究项目(31502030502)

* 通信作者。E-mail: liumaofu@wust.edu.cn

引用格式: 毕健旗, 刘茂福, 胡慧君, 等. 基于依存句法的图像描述文本生成[J]. 北京航空航天大学学报, 2021, 47(3): 431-440.
BI J Q, LIU M F, HU H J, et al. Image captioning based on dependency syntax[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 431-440 (in Chinese).

验知识,并将之与词向量共同输入到解码端参与文本生成。然而,词性序列和句法树包含的语法规则只蕴含每个词的标注信息,虽然序列或树的表示形式能提供一定的结构信息,但仍不能精准地体现文本中词与词之间的关系,即图像中实体间应存在的语法关系,如依存句法关系。因此,本文为模型引入由依存句法构建的语言模型,增强模型理解图像实体间的语法规则的能力,同时在一定程度上提高基于深度学习的图像描述文本生成模型的可解释性。

由于文本单词间可能存在长距离的依存关系,本文选择可以捕获序列内部长距离依赖的自注意力机制(Self-Attention)^[7]来对每个依存句法单元进行编码。编码阶段,在模型得到基于视觉注意力机制计算出的图像区域特征后,为了得到图像区域间的结构语法关系表示,本文提出了一种图像结构注意力方法,从依存句法中提取词与词之间的语法关系,并将其关联在与图像对应的实体上;将两者进行融合并与当前时刻的文本特征向量共同作为 LSTM 的输入,解码生成文本。

针对图像与依存句法的对应关系,在训练阶段,每幅图像与依存句法以描述文本为纽带,呈现直接对应关系;在测试阶段,为得到与测试图像匹配的依存句法模板,本文提出了一种基于内容重合度的提取方法,该方法将训练一个图像标签分类器,旨在输出给定图像相应内容关键词,根据测试图像与训练图像集的最大内容重合度建立起测试图像与训练图像的对应关系,从而间接建立测试图像与依存句法模板的映射关系。

本文提出的基于依存句法的图像描述文本生成模型训练阶段将依存句法应用在图像结构注意力中,测试阶段根据图像内容重合度建立测试图像与依存句法模板的映射关系。此外,模型生成的图像局部间结构语法关系与依存句法单元对应,生成文本的句法与提取的依存句法相吻合。

1 相关工作

编解码方法最初被广泛应用于机器翻译^[8],Vinyals 等^[9]将其迁移至图像描述文本生成领域,取得了显著的效果。Xu 等^[1]提出了一种基于注意力机制的方法来优化 LSTM 的存储能力,使模型在不同时刻关注图像不同区域,生成更准确的描述句子。Zhu 等^[10]提出了一种基于主题词的图像描述文本生成方法,使生成的句子可以包含

更多的图像焦点信息。Wang 等^[11]提出了一种基于内源性和外源性视觉信息来生成不同角度的描述句子的方法。为了使图像信息与文本信息更加契合,Liu 等^[12]提出了一种多头注意力机制来使图像中的信息与视觉文本信息对齐,使解码端可以更加准确地解析模型提供的图像与文本信息。

上述方法均致力于从图像中挖掘更多的视觉信息来填充句子,而在句子结构与句法层次的优化上稍有欠缺。为了使生成模型学习到更多的句子结构信息,Wang 等^[6]提出了一种基于句子骨架与属性词的图像描述文本生成方法,将从图像中得到的信息按照属性词与骨架的排序生成句子。Aditya 等^[5]提出了一种基于词性标注序列的图像描述文本生成方法,加快句子生成速度的同时,也增加了句子描述事物的多样性。

依存句法是一种以谓语动词为结构中心的句法分析方法,在自然语言处理领域被广泛应用。Falenska 和 Kuhn^[13]基于 BiLSTM 结构验证了依存句法内含结构信息的有效性。Li 等^[14]将半监督学习方法应用在跨领域依存句法分析中,得到了更好的中文依存句法分析效果。Wang 等^[15]按照依存句法单元构建句法树,验证了按此顺序定位自注意力机制中每个词,可以优化原模型,达到了提升模型效果的目的。综上所述,依存句法在提取句子结构信息方面显得十分有效。但目前,在图像描述文本生成任务当中,图像内容分析与生成文本的过程缺乏结构信息。因此,本文研究了基于依存句法的图像描述文本生成方法。

2 数据构建与模型方法

本文提出的基于依存句法的图像描述文本生成方法的训练阶段框架如图 1 所示。首先,输入图像经过 ResNet-101 网络^[16]得到图像的特征向量;然后,图像特征向量与词向量一同输入至视觉注意力以计算图像的局部区域特征向量表示,与经过自注意力机制的依存句法向量一同输入至图像结构注意力来计算图像的局部区域关系特征向量表示;最后,将图像局部区域特征向量与局部区域关系特征向量融合,与当前时刻的词嵌入向量一同输入到 LSTM 中生成图像描述文本。图 1 所示视觉注意力来自于基于注意力的 NIC + ATT 模型^[1]中的视觉注意力。

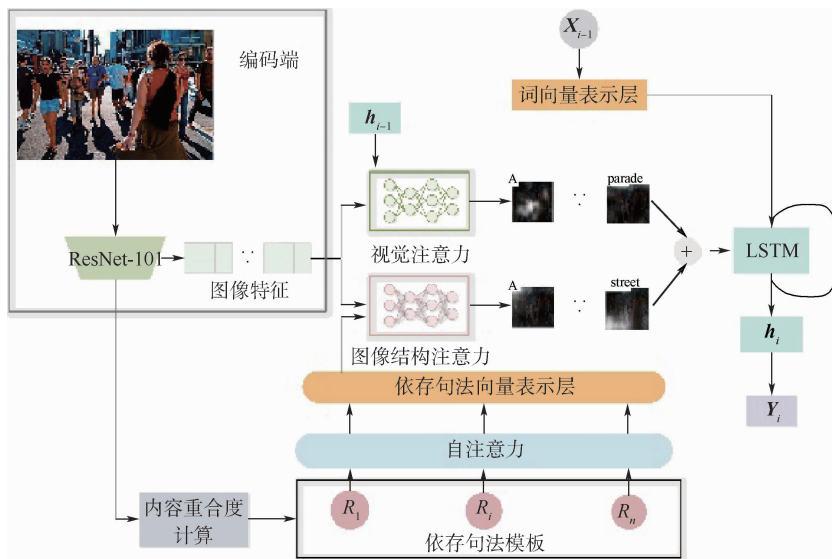


图 1 本文模型框架

Fig. 1 Framework of proposed model

2.1 依存句法库的构建

本文使用 Stanford-CoreNLP^[17] 对全部训练集文本进行依存句法分析,构成依存句法库。将每个依存关系与存在此依存关系的 2 个单词位置信息作为一个依存句法单元,其表示形式为由依存关系类型、根节点位置和自身节点位置 3 个元素构成的三元组,如下:

$$U = (r_i, l_r, i) \quad i \in [1, n] \quad (1)$$

式中: U 为依存句法单元三元组; l_r 为与文本中第 i 个单词存在依存关系的节点,在文本中对应的下标; r_i 为文本中第 l_r 个单词与第 i 个单词之间的依存关系; n 为文本长度。

数据集文本中的每个单词,依存句法单元序列使其都存在唯一的依存关系三元组与之对应,且每个句子的长度与依存句法单元序列的长度相等。

如图 2 所示,在依存句法分析中,名词“牛仔”是动词“骑马”的名词主语,因此三者构成三元组 ('nsubj', 6, 1)。此外,中文分词的结果中包含 7 个词,对应的依存句法单元有 7 个。

中文文本:牛仔在牛仔竞技中骑马。
中文分词:牛仔,在,牛仔,竞技,中,骑马。
依存句法:('ROOT', 0, 6), ('nsubj', 6, 1), ('case', 4, 2), ('compound:nn', 4, 3), ('nmod:prep', 6, 4), ('case', 4, 5), ('punct', 6, 7)

图 2 图像描述文本依存句法示例

Fig. 2 A dependency syntax example of an image caption

2.2 依存句法单元向量表示

自回归机制可以提取文本的向量表示,但不能很好地解决长距离依赖问题。依存句法体现的是句子中的词与根、词与词之间的关系,包

含长距离的语法规则。为了使计算机更好地理解依存句法单元三元组序列,本文使用自注意力机制来解决 LSTM 的自回归机制不能捕获长距离关系的问题。自注意力机制摒弃自回归机制逐个输入单词计算向量表示的方法而一次性处理整个文本,因此,自注意力机制更适合用来处理具有长距离关系的文本。

$$R = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \times V \quad (2)$$

式中: Q, K 和 V 为输入的依存句法单元序列随机嵌入向量, $Q, K, V \in \mathbf{R}^{L \times E}$, L 为依存句法长度, E 为词嵌入向量的维度; d_k 为 K 的第三维度; R 为依存句法单元序列的向量表示; softmax 为归一化函数。

2.3 图像结构注意力

一幅图像包含实体、背景等信息,且在对应的描述文本中可找到与图像中的实体对应的单词。从构成句子的语法规则角度来看,为了生成句法正确的文本,掌握单词与单词之间的关系十分重要,也就是需要发现图像中实体与实体之间的关系。图像结构注意力的目的是通过文本的依存句法中表示的词与词之间的关系,使模型学习到图像中对应每个实体之间的关系特征向量表示。具体的计算公式为

$$\begin{cases} \alpha = \text{softmax}(\text{relu}(W_i I' + W_r R)) \\ g = \text{sigmoid}(W_g R) \\ I = g \times \text{mean}(\alpha \times I') \end{cases} \quad (3)$$

式中: $I' \in \mathbf{R}^{D \times F}$ 与 $R \in \mathbf{R}^{L \times E}$ 分别为图像特征向量与依存句法单元序列向量, D 和 F 分别为图像

特征点数和维度; α 为依存句法单元向量在图像特征点上关注的权重系数; g 为依存句法单元向量在图像特征维度上关注的权重系数; W_* ($*$ $\in \{i, r, g\}$) 为模型参数; softmax、relu 和 mean 分别为归一化函数、激活函数和均值函数; I 为最终的图像结构特征向量。

2.4 图像描述文本生成

图像的描述文本由 LSTM 生成。 t 时刻输入至 LSTM 的元素由 LSTM 在 $t-1$ 时刻生成的单词 X_{t-1} 与隐藏状态 h_{t-1} 、 X_{t-1} 在 t 时刻关注的图像视觉区域特征 I_x 、依存句法单元 R 在 $t-1$ 时刻关注的图像结构特征 I_r 构成, I_{add} 为 I_x 与 I_r 的加和, h_t 为 t 时刻 LSTM 模型中的隐向量, Y_t 为模型输出的 t 时刻的词。具体的计算公式为

$$\begin{cases} I_{add} = I_x + I_r \\ h_t = \text{LSTM}(X_{t-1}, h_{t-1}, I_{add}) \\ Y_t = \text{softmax}(h_t) \end{cases} \quad (4)$$

本文提出的模型训练过程的目标为:通过最小化损失函数来学习模型的最优参数。损失函数的具体计算公式为

$$L = -\frac{1}{N} \sum_{j=1}^N \sum_{i=1}^{\text{len}(Y)} \log_2 P(Y_i) \quad (5)$$

式中: N 为训练集一个批次的描述文本的数量; $\text{len}(Y)$ 为描述文本的长度; Y_i 为第 j 个文本的第 i 个词; $P(Y_i)$ 为 Y_i 在当前文本中出现的概率。

2.5 基于内容重合度的依存句法模板提取

2.5.1 内容关键词分类器

内容关键词分类器用来对测试图像进行分类,以建立起测试图像与依存句法模板的映射关系。为得到描述文本依存句法尽可能相似的图像,首先,内容关键词标签来自于训练集描述文本中的单词,且该单词在文本对应的依存句法中带有“ROOT”关系。此外,该单词是训练集文本中出现的高频词。本文根据上述规则选择 300 个关键词作为内容关键词分类器的标签,且将该分类器视为一种图像多分类问题。具体计算公式为

$$\text{labels} = \text{softmax}(\text{MLP}_2(I)) \quad (6)$$

式中: I 为图像特征向量; MLP_2 为多层感知机模型,此处层数为 2。

2.5.2 内容重合度

内容重合度体现了测试图像与训练图像集中的每幅图像的相似程度。由于每个内容关键词代表图像中可能出现的实体或关系,基于关键词的内容重合度可以很好地提取主题类似的 2 幅

图像。

测试图像经过内容关键词分类器得到 Top-K 的标签集 C_t , 同样地, 训练集中的每幅图像也得到 Top-K 的标签集 C_{tr} 。由于每个标签的可信度均不同, 导致每个标签都会被赋予不同的权重。根据测试图像与训练图像的内容重合度 S 来确定测试图像与训练图像之间的相似关系。内容重合度的计算公式为

$$\begin{cases} S_i^x = \frac{(K - \text{ind}_t^i) + (K - \text{ind}_{tr}^i)}{2} \\ S^x = \sum_{i=1}^M S_i^x \end{cases} \quad (7)$$

式中: ind_t^i 表示测试图像的 Top-K 标签集 C_t 中的第 i 个, 同时在 C_{tr} 中出现的标签 c_i 在 C_t 中对应的下标值; ind_{tr}^i 表示标签 c_i 在训练图像的 Top-K 标签集 C_{tr} 中的下标值。在所有的标签集合 C_t 和 C_{tr} 中, 标签的下标值越大, 其对应的权重越小。 S_i^x 代表第 x 幅训练图像与测试图像相同的第 i 个标签的得分, M 为相同标签的个数, S^x 则为第 x 幅训练图像与测试图像的内容重合度。本文通过比较 S^x 的数值大小来确定训练图像集中与测试图像最相似的图像, 进而建立测试图像与依存句法模板的映射关系。以 $K=3$ 举例, 假设 $C_t = [16, 7, 8]$, $C_{tr}^1 = [1, 7, 6]$, $C_{tr}^2 = [1, 7, 8]$, 则 $S^1 = 2$, $S^2 = 3$ 。

算法 1 内容重合度计算。

输入: 测试图像 $T = (t_1, t_2, \dots, t_n)$; 训练图像 $X = (x_1, x_2, \dots, x_m)$ 。

输出: 测试图像与训练图像的重合度 S 。

步骤 1 测试图像标签集 $C_t \leftarrow \text{labels}(T)$ (见式(6))。

步骤 2 训练图像标签集 $C_{tr} \leftarrow \text{labels}(X)$ (见式(6))。

步骤 3 相同标签集 $C \leftarrow (C_t, C_{tr})$ 。

步骤 4 对应下标 $\text{ind}_t, \text{ind}_{tr} \leftarrow (C, T, X)$ 。

步骤 5 计算内容重合度 S^x (见式(7))。

算法 1 最耗时操作集中在标签识别上, 即步骤 1 和步骤 2 得到测试与训练图像标签的过程, 其次是计算内容重合度 S^x 。假设测试图像数量为 n , 训练图像数量为 m , 则算法 1 的时间复杂度为 $O(mn)$ 。

3 实验

3.1 实验设置

本文实验数据集采用 Flickr30K、Flickr8K 和

Flickr8K-CN。Flickr30K 是一个被广泛使用的公开英文数据集,共有图像 31 784 幅,每幅图像对应 5 个英文描述句子。Flickr8K 数据集中包含图像 8 091 幅,同样每幅图像对应 5 个英文描述句子。Flickr8K-CN 则为 Flickr8K 对应的中文版本。在测试阶段,使用了 7 种常见的评价指标和句子长度来验证模型生成文本的质量。7 种指标分别为: BLEU-1 ~ BLEU-4^[18]、METEOR^[19]、ROUGE-L^[20]、CIDEr^[21]。其中,前 4 个指标使用了 n-gram 算法统计生成文本与参考文本之间的覆盖率; METEOR 对 BLEU 算法进行了一定的改进,使其更好地体现出句中单词的召回率与准确率; ROUGE-L 基于最长公共子串来计算准确性与召回率; CIDEr 基于 TF-IDF 计算生成文本与参考文本的余弦相似度来衡量文本的一致性。此外,本文将模型生成的图像描述文本平均长度 Len 作为第 8 种评测指标。

本文提出的模型选择 ResNet-101 网络结构来提取图像视觉特征;词嵌入与依存句法嵌入向量均为随机初始化。训练优化器是 Adam,编码器的学习率为 0.000 1,解码器的学习率为 0.000 4,每 8 轮次损失未下降则使用 0.8 的学习率衰减系数。损失函数为交叉熵函数,模型的超参数设置如表 1 所示。

表 1 超参数设置

Table 1 Hyperparameter setting

参数	数值
图像特征向量/维	14 × 14 × 2 048
词向量/维	512
依存句法向量/维	512
LSTM 隐向量/维	512
自注意力机制头数	8
批处理大小	32

3.2 实验结果分析

表 2 ~ 表 4 分别表示了本文模型在 Flickr30K、Flickr8K 与 Flickr8K-CN 数据集上的实验结果(表中加粗数据为最优结果)。其中,NIC + ATT 是基于注意力机制的神经图像描述生成模型, AdaptAtt^[22] 是基于自适应的图像描述生成模型, NIC + WC + WA + RL^[23] 是基于图像特征动态建立与图像相关词汇表的模型, MLO/MLPF-LSTM + (BS)^[24] 是基于深度 LSTM 的模型, CACNN-GAN (ResNet-152)^[25] 是基于 GAN^[26] 的模型, NIC + DS 指在神经图像描述生成模型的基础上,编码阶段 LSTM 增加输入依存句法向量,而不添加图像结构注意力, NIC + DSSA 指在神经图像描述生成模型的基础上,增加图像结构注意力, Top-5、Top-10 分别指内容关键词分类器按概率大小取降序排名前 5 或前 10 的标签来计算内容重合度。

表 2 Flickr30K 数据集的实验结果

Table 2 Experimental results on Flickr30K dataset

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Len
NIC + ATT(Baseline)	62.84	39.00	25.07	17.52	17.98	44.57	30.18	11.06
AdaptAtt	60.69	41.80	25.92	18.63	19.71	45.61	33.36	
NIC + WC + WA + RL				24.50	21.50	51.60	58.40	
MLO/MLPF-LSTM + (BS)	66.20	47.20	33.10	23.00	19.60			
CACNN-GAN(ResNet-152)	69.30	49.90	35.80	25.90	22.30			
NIC + DS(Top-5)	57.09	39.35	28.66	20.73	20.81	48.24	49.78	17.58
NIC + DSSA(Top-5)	58.62	40.46	29.81	22.62	20.96	49.98	51.74	17.56
NIC + DS(Top-10)	59.76	44.53	31.48	24.75	21.31	51.36	50.91	18.43
NIC + DSSA(Top-10)	61.81	47.33	33.97	26.06	23.57	52.81	52.48	18.62

表 3 Flickr8K 数据集的实验结果

Table 3 Experimental results on Flickr8K dataset

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Len
NIC + ATT(Baseline)	60.32	37.88	24.66	16.33	18.48	46.16	34.99	11.17
NIC + DS(Top-10)	57.76	41.16	30.70	27.78	19.54	48.81	36.69	14.47
NIC + DSSA(Top-10)	59.45	45.86	36.05	29.36	21.92	50.06	40.24	15.72

表 4 Flickr8K-CN 数据集的实验结果

Table 4 Experimental results on Flickr8K-CN dataset

模型	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	Len
NIC + ATT(Baseline)	59.16	36.30	22.73	16.02	16.87	43.59	31.09	10.82
NIC + DS(Top-10)	56.28	40.03	29.42	25.61	17.48	46.21	34.16	13.45
NIC + DSSA(Top-10)	58.72	46.86	33.05	28.16	20.57	49.10	38.48	14.36

由表 2 可知,本文模型 NIC + DSSA 除 BLEU-1 指标外,其他 7 种指标均优于 NIC + ATT 模型,尤其在 BLEU-4 与 CIDEr 指标上分别改善了 8.54% 和 22.30%。由此可见,依存句法的使用可以提高模型对文本与图像结构信息的捕获能力。在 CIDEr 指标上,NIC + DSSA 模型略低于 NIC + WC + WA + RL 模型,由于 CIDEr 指标计算得分时考虑到参考文本与模型生成文本的长度差,由文本平均长度 Len 可知,本文模型 NIC + DSSA 生成的句子较长,所以在 CIDEr 指标上不占优势。CACNN-GAN 模型使用了 ResNet-152 网络提取图像特征,但由于在生成阶段不能很好地考虑到长距离的依赖关系,在 BLEU-4 指标上仍低于本文提出的模型。此外,由表 3 和表 4 的数据可知,本文提出的模型在中英文数据集上具有一定的泛化能力。

在 3 个实验数据集中,NIC + DS 模型与 NIC + DSSA 模型在生成句子的平均长度接近的情况下,NIC + DSSA 模型在各个指标上均好于 NIC + DS 模型,在 BLEU-4 指标上提高近 1.5%。针对图像结构注意力,本文选取 Flickr30K 数据集中的一幅测试图像构建出 2 种模型对应的图像注意力在测试图像上的分布图,如图 3 所示。

由 NIC + DSSA 模型生成的图像注意力分布图与描述文本可知,除主句中的实体外,添加了图像结构注意力的 NIC + DSSA 模型捕捉到了其他图像实体间关系(<< “young woman”, “talk”, “man” >, “converse”, < “two other men” >>), 且使用“while”将主句与从句连接起来;而 NIC + DS 模型生成的文本更注重图像中某个实体细节描述,如“black and white striped dress”,而忽略了图像实体间的关系,不能很好地在句式上体现出依存句法添加至模型后的优势。为了控制其他变量的影响,两者均输入相同的依存句法模板。

给定测试图像,本文模型应尽可能地提供符合图像内容实体之间实际句法关系的依存句法模板。在内容关键词分类器中,若 K 值不同,模型为测试图像选择的依存句法模板可能不同。因此,本文研究了 K 取值对 BLEU-4 指标的影响,如图 4 所示。

由图 4 可知,在 K 值为 10 时,模型在实验中得到了最好的图像描述文本生成效果,BLEU-4 值最高。 K 取值较小时,分类器选择的图像描述标

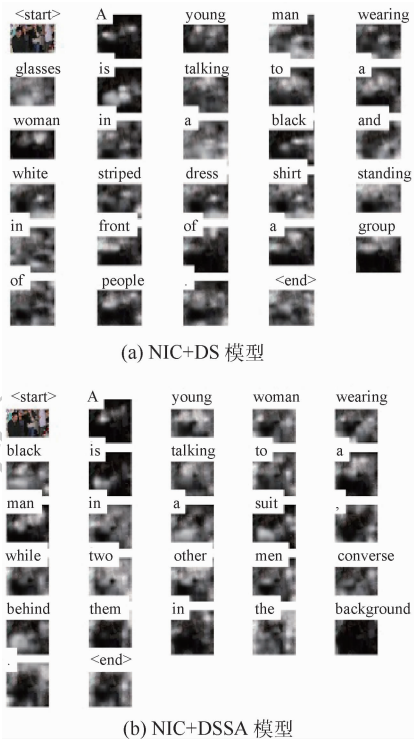


图 3 生成文本效果对比

Fig. 3 Comparison of generated captions

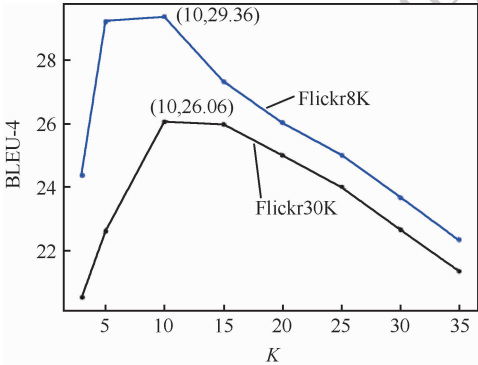


图 4 不同 K 值选取对实验结果的影响

Fig. 4 Experimental results affected by different K values

签较少,不能完整地概括一幅图像; K 取值较大时,标签会同图像主题偏离而降低生成图像描述文本的准确性。无论 K 值过大还是过小,都会导致“假重合”情况发生,故本文 K 值为 10。

当 K 值为 10 时,每个关键词标签可以较准确地描绘图像的不同侧面,当 2 幅图像的内容重合度较高时,描述文本中实体或事件高度相似。图 5 给出了当 K 值为 10 时,计算 2 幅图像内容重合度的例子(此例内容重合度最大为 55)。

根据内容重合度,测试图像可以在训练图像集中匹配到与自身内容最相似的图像,由于在训练图像集中,每幅图像都有 5 个参考文本与之对应,因此本文模型可以将对应的 5 个依存句法模板依次输入到模型中,指导模型生成

图像描述文本,在一定程度上增加了生成句子的多样性。

如图 6 所示,本文提出的 NIC + DSSA 模型可以根据不同的依存句法模板生成不同的图像描述文本,且生成的文本(5)与参考文本(5)在句式上保持一致。

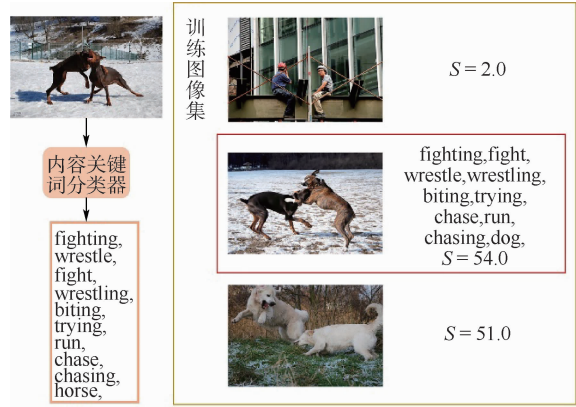


图 5 K 取 10 时分类效果

Fig. 5 Results of classification when K is 10

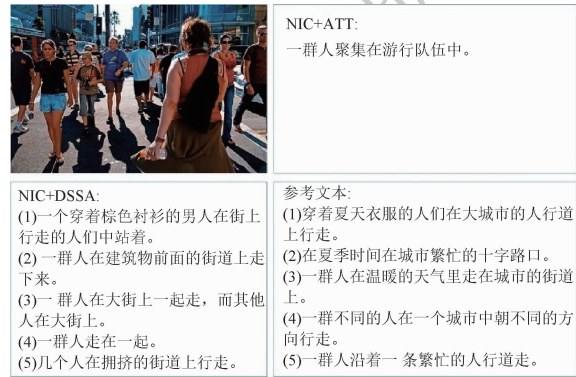


图 6 模型生成的文本多样性示例

Fig. 6 Example of diversity of captions generated by model

3.3 深度学习模型可解释性分析

本节将进一步细粒度地分析本文模型如何根据输入图像与依存句法信息生成描述文本,使深度学习模型更具可解释性。

由图 7(a)可知,NIC + ATT 模型在生成文本句的主体时,图像关注的区域较分散,不能很好地区分每个单词或词组对应图像区域的异同。而 NIC + DSSA 模型由于加入了依存句法信息中的结构关系,生成的每个单词或词组关注的图像区域可以清晰地分辨出差异,并具有一定的连续性,如图 7(b)所示。

在图 7 所示的 2 个生成文本中,NIC + ATT 模型在生成图像注意力时,特别在词组“are gathered in”对应的图像区域中,关注的图像区域居中、分布较广且高度相似,因此,不能很好地捕获图像区

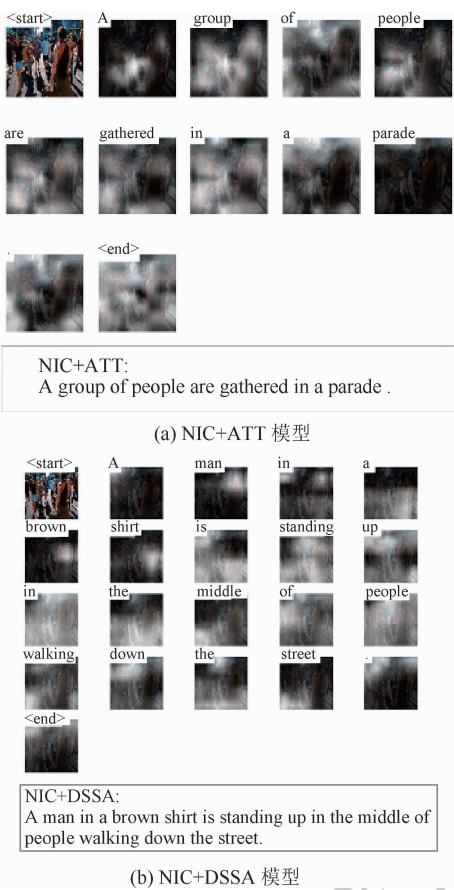


图 7 图像注意力对比

Fig. 7 Comparison of image attention

域内部实体间关系,即在此图中仅识别出“parade”;而 NIC + DSSA 模型在生成图像注意力时,不同的图像区域间存在一定的差异性,而仅在描述事物的名词词组“brown shirt”、动词词组“standing up”或介词词组“the middle of”对应的图像区域分布较类似,符合人类分析图像的过程。

如图 8 所示,根据相似图像,模型将得到 5 个依存句法模板。在图 8 显示的图 7 文本对应的依存句法模板中,第 1 个依存句法单元代表生成的文本在位置 8 上的单词“standing”应为该文本句的词根“ROOT”;第 2 个依存句法单元代表生成的

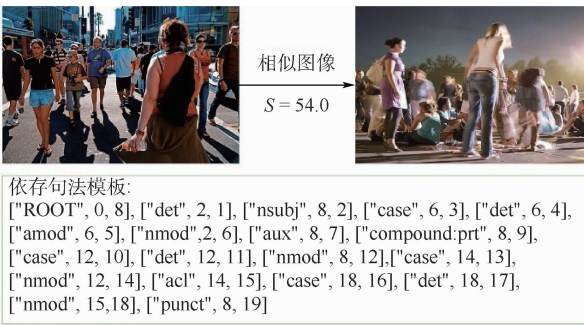


图 8 相似图像和依存句法模板

Fig. 8 Similar image and dependency syntactic template

文本在位置 1 上的单词“A”应为在位置 2 上的单词“man”的限定词“det”;第 3 个依存句法单元则代表生成的句子在位置 2 上的单词“man”应为在位置 8 上的单词“standing”的名词主语“nsubj”。由图 8 可知,输入的依存句法通过图像结构注意

力间接指导了描述文本生成,且将图像的区域间结构关系与文本中的依存关系对应。图 9 给出了本文模型 NIC + DSSA 在 Flickr8K 数据集部分图像上的实验结果,体现了依存句法在模型生成图像描述文本中的指导作用。



图 9 基于依存句法的图像描述文本生成示例

Fig. 9 Examples of image captioning based on dependency syntax

为分析模型生成文本时子句的使用情况,统计了测试阶段模型在 Flickr30K 数据集上生成描述文本连接词的使用情况,如表 5 所示。本文模型由于依存句法信息的加入,在生成连接词方面表现能力较突出,很好地解释了其生成文本较长这一现象。

表 5 描述文本中连接词数量统计	
Table 5 Statistics of conjunction numbers in captions	
模型	连接词数量
NIC	1
NIC + DSSA	66
参考文本	58

4 结 论

本文提出了一种基于依存句法的图像描述文本生成模型,将依存句法作为语法规则输入至图

像结构注意力中,使模型在生成图像描述文本的过程中考虑到依存句法提供的词与词之间的依赖关系。

- 1) 验证了通过使用图像结构注意力加强模型对图像中实体间关系的学习理解能力。
- 2) 提出了一种通过提取图像内容关键词来计算图像内容重合度的方法,建立测试图像与训练图像集的映射关系,进而将测试图像与依存句法模板匹配。
- 3) 在 Flickr30K、Flickr8K 和 Flickr8K-CN 三个数据集上进行的大量实验表明,本文提出的模型在增加文本多样性和句法复杂度 2 个方面均有较好的表现。

此外,本文还通过模型生成图像描述文本的实例分析,解释了依存句法在本文提出的模型中如何指导文本生成及关注图像结构特征,增强了深度学习模型的可解释性。

在未来的工作中,仍有2个问题需要持续关注:①依存句法表示问题,目前使用的自注意力机制虽然可以较好地得到长距离的依赖关系表示,但其可解释性不高;②依存句法模板获取问题,本文提出的内容重合度计算方法旨在根据“ROOT”中心词识别内容相似的图像,但忽略了图像内容的深层含义如图像主题^[27]等,有时无法得到最优的依存句法模板。

参考文献 (References)

- [1] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention [C] // Proceedings of the 32nd International Conference on Machine Learning, 2015:2048-2057.
- [2] LU C, KRISHNA R, BERNSTEIN M S, et al. Visual relationship detection with language priors [C] // Proceedings of the 14th European Conference on Computer Vision, 2016:852-869.
- [3] 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 28, 2015:91-99.
- [4] GUO Y, CHENG Z, NIE L, et al. Quantifying and alleviating the language prior problem in visual question answering [C] // Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2019:75-84.
- [5] ADITYA D, JYOTI A, LIWEI W, et al. Fast, diverse and accurate image captioning guided by part-of-speech [C] // Proceedings of the 32nd IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019:10695-10704.
- [6] WANG Y, LIN Z, SHEN X, et al. Skeleton key: Image captioning by skeleton-attribute decomposition [C] // Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017:7272-7281.
- [7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C] // Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017:5998-6008.
- [8] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [C] // Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014:3104-3112.
- [9] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(4):652-663.
- [10] ZHU Z, XUE Z, YUAN Z. Topic-guided attention for image captioning [C] // Proceedings of the 25th IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2018:2615-2619.
- [11] WANG T, HU H, HE C. Image caption with endogenous-exogenous attention [J]. Neural Processing Letters, 2019, 50(1):431-443.
- [12] LIU F, LIU Y, REN X, et al. Aligning visual regions and textual concepts: Learning fine-grained image representations for image captioning [EB/OL]. (2019-05-15) [2020-08-01]. <https://arxiv.org/abs/1905.06139v1>.
- [13] FALENSKA A, KUHN J. The (non-) utility of structural features in BiLSTM-based dependency parsers [C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019:117-128.
- [14] LI Z, PENG X, ZHANG M, et al. Semi-supervised domain adaptation for dependency parsing [C] // Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019:2386-2395.
- [15] WANG X, TU Z, WANG L, et al. Self-attention with structural position representations [EB/OL]. (2019-09-01) [2020-08-01]. <https://arxiv.org/abs/1909.00383>.
- [16] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition [C] // Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016:770-778.
- [17] CHRISTOPHER D M, MIHAI S, JOHN B, et al. The stanford-CoreNLP natural language processing toolkit [C] // Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics, 2014:55-60.
- [18] PAPINENI K, ROUKOS S, WARD T, et al. BLEU: A method for automatic evaluation of machine translation [C] // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002:311-318.
- [19] CHOUETI T K, ESCUDIER B, POWLES T, et al. Cabozantinib versus everolimus in advanced renal cell carcinoma (METEOR): Final results from a randomised, open-label, phase 3 trial [J]. The Lancet Oncology, 2016, 17(7):917-927.
- [20] LIN C Y. Rouge: A package for automatic evaluation of summaries [C] // Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 2004:74-81.
- [21] VEDANTAM R, ZITNICK C L, PARIKH D. CIDER: Consensus-based image description evaluation [C] // Proceedings of the 28th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015:4566-4575.
- [22] LU J, XIONG C, PARIKH D, et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning [C] // Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017:3242-3250.
- [23] FAN Z, WEI Z, HUANG X, et al. Bridging by word: Image grounded vocabulary construction for visual captioning [C] // Proceedings of the International Conference on the Association for the Advance of Artificial Intelligence, 2019:6514-6524.
- [24] 汤鹏杰, 王瀚漓, 许恺晟. LSTM 逐层多目标优化及多层概率融合的图像描述 [J]. 自动化学报, 2018, 44(7):1237-1249.
- TANG P J, WANG H L, XU K S. Multi-objective layer-wise optimization and multi-level probability fusion for image description generation using LSTM [J]. Acta Automatica Sinica, 2018, 44(7):1237-1249 (in Chinese).
- [25] 薛子育, 郭沛宇, 祝晓斌, 等. 一种基于生成式对抗网络的图

像描述方法[J]. 软件学报, 2018, 29(2): 30-43.

XUE Z Y, GUO P Y, ZHU X B, et al. Image description method based on generative adversarial networks[J]. Journal of Chinese Information Processing, 2018, 29(2): 30-43 (in Chinese).

[26] SALVARIS M, DEAN D, TOK W H, et al. Generative adversarial networks[EB/OL]. (2014-06-10) [2020-08-01]. <https://arxiv.org/abs/1406.2661>.

[27] LIU M, HU H, LI L, et al. Chinese image caption generation via visual attention and topic modeling[J/OL]. IEEE Transactions on Cybernetics, 2020 (2020-06-22) [2020-08-01]. <https://ieeexplore.ieee.org/document/9122435>.

作者简介:

毕健旗 男, 硕士研究生。主要研究方向: 自然语言处理。

刘茂福 男, 博士, 教授, 博士生导师。主要研究方向: 自然语言处理、图像分析与理解。

胡慧君 女, 博士, 副教授, 硕士生导师。主要研究方向: 智能信息处理、图像分析与理解。

代建华 男, 博士, 教授, 博士生导师。主要研究方向: 人工智能、智能信息处理。

Image captioning based on dependency syntax

BI Jianqi^{1,2}, LIU Maofu^{1,2,*}, HU Huijun^{1,2}, DAI Jianhua³

- (1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;
2. Hubei Provincial Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430081, China;
3. Hunan Provincial Key Laboratory of Intelligent Computing and Language Information Processing, Hunan Normal University, Changsha 410081, China)

Abstract: Current image captioning model can automatically apply the part-of-speech sequences and syntactic trees to make the generated text in line with grammar. However, the above-mentioned models generally generate the simple sentences. There is no groundbreaking work in language models promoting the interpretability of deep learning models. The dependency syntax is integrated into the deep learning model to supervise the image captioning, which can make deep learning models more interpretable. An image structure attention mechanism, which recognizes the relationship between image regions based on the dependency syntax, is applied to compute the visual relations and obtain the features. The fusion of image region relation features and image region features, and the word embedding are employed into Long Short-Term Memory (LSTM) to generate the image captions. In testing, the content keywords of the testing and training image datasets are produced due to the content overlap of two images, and the dependency syntax template corresponding to the test image can be indirectly extracted. According to the dependency syntax template, the diverse descriptions can be generated. Experiment results verify the capacity of the proposed model to improve the diversity of generated captions and syntax complexity and indicate that the dependency syntax can enhance the interpretability of deep learning model.

Keywords: image captioning; dependency syntax; image structure attention; content overlap; interpretability of deep learning model

Received: 2020-08-21; **Accepted:** 2020-09-05; **Published online:** 2020-09-29 13:35

URL: kns.cnki.net/kcms/detail/11.2625.V.20200929.1427.003.html

Foundation items: Major Projects of National Social Science Foundation of China (11&ZD189); Pre-research Foundation of Whole Army Shared Information System Equipment (31502030502)

* **Corresponding author.** E-mail: liumaofu@wust.edu.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0479

基于多特征图像视觉显著性的视频摘要化生成

金海燕^{1,2}, 曹甜¹, 肖聪¹, 肖照林^{1,2,*}

安理工大学 计算机科学与工程学院, 西安 710048; 2. 陕西省网络计算与安全技术重点实验室, 西安 710048)

摘 要: 如何高效提取视频内容即视频摘要化, 一直是计算机视觉领域研究的热点。

简单通过图像颜色、纹理等特征进行检测已无法有效、完整地获取视频摘要。基于视觉注意力金字塔模型, 提出了一种改进的可变比例及双对比度计算的中心-环绕视频摘要化方法。首先, 以超像素方法对视频图像序列进行像素块划分以加速图像计算; 然后, 检测不同颜色背景下的图像对比度特征差异并进行融合; 最后, 结合光流运动信息, 合并静态图像与动态图像显著性结果提取视频关键帧, 在提取关键帧时, 利用感知哈希函数进行相似性判断完成视频摘要化生成。在 Segtrack V2、ViSal 及 OVP 数据集上进行仿真实验, 结果表明: 所提方法可以有效提取图像感兴趣区域, 得到以关键帧图像序列表示的视频摘要。

关 键 词: 视频摘要化; 视觉注意力金字塔; 视频显著性; 关键帧提取; 相似性判断

中图分类号: TP391.41

文献标志码: A

文章编号: 1001-5965(2021)03-0441-10

如何以最为直观且快速的方式查阅视频数据, 完成像文字搜索一般的浏览效率是计算机视觉与图像处理领域的研究热点。视频摘要是对原始视频内容的高度浓缩, 其将重要且具有代表性的视频内容以一种简洁的形式呈现出来, 方便用户对视频的浏览和管理^[1]。

2017年, 刘全等^[2]使用带视觉注意力机制的循环神经网络(RNN)改进了传统的深度Q网络模型, 提出了一种较为完善的深度强化学习模型。2018年, 郎洪等^[3]提出一种鲁棒主成分分析(RPCA)优化方法, 为了快速筛选与追踪前景目标, 以基于帧差欧氏距离方法设计显著性目标帧快速提取算法。2019年, 张芳等^[4]为准确检测复杂背景下的显著区域, 提出了一种全卷积神经网络与低秩稀疏分解相结合的显著性检测方法, 结合利用全卷积神经网络学习得到的高层语义先验知识, 检测图像中的显著区域。李庆武等^[5]针对现

有显著性检测算法检测目标类型单一、通用性差的问题, 提出了一种基于无监督栈式降噪自编码网络的显著性检测算法。2020年, 陈炳才等^[6]提出了一种融合边界连通性与局部对比性的图像显著性检测算法, 得到的显著图更接近于真值图。

由于现有视频数据量过于庞大, 占用内存资源较多, 在浏览时比较困难。本文基于多特征图像和视觉注意力金字塔模型, 提出了一种改进的可变比例及双对比度计算的中心-环绕视频摘要化方法, 通过提取的关键帧快速理解视频的主要内容。本文方法不仅易于实现, 还改善了传统方法的提取效果。在 Segtrack V2、ViSal 及 OVP 数据集上进行仿真实验, 验证了本文方法的有效性。

1 相关工作

1.1 视觉显著性

在某一场景中, 能够令人眼所引起注意的区

收稿日期: 2020-08-31; 录用日期: 2020-10-27; 网络出版时间: 2020-11-03 09:44

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201102.1036.002.html

基金项目: 陕西省技术创新引导计划(2020CGXNG-026); 陕西省自然科学基金基础研究计划(2019JM-221)

*通信作者: E-mail: xiaozhaolin@xaut.edu.cn

引用格式: 金海燕, 曹甜, 肖聪, 等. 基于多特征图像视觉显著性的视频摘要化生成[J]. 北京航空航天大学学报, 2021, 47(3): 441-450. JIN H Y, CAO T, XIAO C, et al. Video summary generation based on multi-feature image and visual saliency[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 441-450 (in Chinese).

域就是该场景中最显著的区域,场景其余部分则可能不会被人眼所注意或考虑在内。通过不同方式满足人眼的机制并遵循人眼视觉习惯所进行的检测,即为视觉显著区域的检测。

2017年,Ablavatski等^[7]设计了一种改进的基于注意力的体系结构,用于多对象识别;Qu等^[8]为了解决僵化的描述问题,提出了一种神经和概率框架,将卷积神经网络(CNN)与循环神经网络相结合,以产生端到端的图像字幕。2019年,Liu和Yang^[9]提出了一种“前景-中心背景”显著区域检测模型,提高了显著性检测的性能。

1.2 视频摘要化

现有的大部分视频摘要都为静态视频摘要。2016年,Li等^[10]开发了一种用于在Internet上搜索结果的多媒体新闻摘要的新颖方法,可发现与查询相关的新闻信息中的基本主题,并将每个主题中的新闻事件穿线以生成与查询相关的简要概述。2018年,Hu和Li^[11]通过融合基于多个特征和图像质量的全局重要性和局部重要性来生成动态视频摘要。Meng等^[12]选择在不同视图之间代表视频的视觉元素,使用质心共正则化方法的多视图稀疏字典选择,优化了每个视图中的代表性选择,并通过将它们针对共识选择进行正则化来强制将视图特定的选择相似。

1.3 关键帧提取中的图像特征

颜色特征作为较为常见和易于获得的信息为特征提取时所广泛采用。除了颜色以外,图像的纹理也是视频摘要化的常见特征。作为一个图像或物体表面具有的固有特性,纹理特征通过空间中的某种形式的颜色变化而产生不同的图案,并通过对图像进行量化产生特征结果。LBP(Local Binary Pattern)算子是最常见的原始特征算子之一,所提取出的特征即为图像中局部位置的纹理特征。

2 显著性检测

2.1 超像素分割预处理

本文采用SLIC方式^[13]对图像进行分割,将图像中的前景区域和背景区域进行合理的划分。将不属于前景物体的部分划分为一部分,将有可能属于前景的区域归属于另一部分。被超像素划分出的区域之间则代表了前景与背景区域之间的界限。设视频图像序列 $F = \{f_1, f_2, f_3\}$,得到每一帧都使用SLIC后的超像素图像序列为 Y , $Y = \{y_1^k, y_2^k, \dots, y_n^k\}$, k 为期望的超像素数量,对所有通道分量下的图像进行超像素分割。

2.2 双对比度金字塔显著性检测

计算RG、BY颜色分量及亮度特征分量。其中,亮度分量 $\text{Intensity} = (R + B + G)/3$,红-绿颜色分量 $\text{RG} = R - G$,黄-蓝颜色分量 $\text{BY} = B - (G + R)/2$ 。对已进行超像素分割的图像序列,为每一个图像特征通道构建中心金字塔与环绕金字塔的尺度空间。设每一个图像的中心金字塔为 C , $C = \{C_0, C_1, C_2, \dots, C_N\}$,每一个图像的环绕金字塔为 s , $s = \{S_0, S_1, S_2, \dots, S_N\}$ 。本文 N 的值为5,即金字塔的层数为5。同样,每一个中心金字塔图像 C_k 都有一个与之对应的环绕金字塔图像 S_k 。其中,环绕金字塔图像 S_k 是由中心图像经过高斯平滑处理所得。高斯平滑因子为

$$\mu = \sqrt{\mu_s^2 - \mu_c^2} \quad (1)$$

式中: μ_s 为环绕金字塔图像 S_k 的平滑因子值; μ_c 为用于得到中心金字塔图像 C_k 对应的值。因此,算法可以自由动态地调整平滑因子的值大小,产生更多的灵活性来适应系统,从而在之后的计算中得到更好的效果。

对每一个图像特征通道计算中心-环绕对比度。设中心金字塔中第 i 层图像为 C_i ,环绕金字塔中第 i 层图像为 S_i ,由式(2)、式(3)分别得到正-负对比度 Contrast_+ 和负-正对比度 Contrast_- 。

$$\text{Contrast}_+ = C_i - S_i \quad (2)$$

$$\text{Contrast}_- = S_i - C_i \quad (3)$$

同时,将数值中所有小于0的特征值统一设为0。在得到对比度结果后,将所有特征图按照正-负对比度与负-正对比度2类进行加法求和。

2.3 光流信息提取

设像素点 $p(x, y)$ 在某个时间段 t 内移动的距离为 (dx, dy) ,原像素点表示为 $p(x, y, t)$,变化移动的点表示为 $p(x + dx, y + dy, t + dt)$,考虑亮度恒定的情况,得到

$$p(x, y, t) = p(x + dx, y + dy, t + dt) \quad (4)$$

根据亮度恒定得到的公式以灰度值进行空间坐标位置求导,得到泰勒级数展开公式,进一步得

$$\frac{\partial p}{\partial x} \cdot \frac{dx}{dt} + \frac{\partial p}{\partial y} \cdot \frac{dy}{dt} + \frac{\partial p}{\partial t} \cdot \frac{dt}{dt} = 0 \quad (5)$$

$dx/dt, dy/dt$ 分别为 X, Y 轴方向点以灰度信息来描述的值的速率,设为 U, V ;而 $\partial p/\partial x, \partial p/\partial y, \partial p/\partial t$ 分别为二维图像上 X, Y, t 方向的偏导,设为 p_x, p_y, p_t ,由此得到光流法公式:

$$p_x U + p_y V + p_t = 0 \quad (6)$$

p_x, p_y, p_t 即为根据图像序列在 (x, y, t) 上的差分。由于式(6)中存在2个未知变量的情况,在求得式(6)后,需要多加一个约束式来得到光流场计算结果。这里使用Lucas-Kanade光流法方式来增加约束。

2.4 前后帧图像抑制法的光流检测

在动态检测时,可能会由于经常引入不相关的背景或线条轮廓内容,对检测产生较大影响,且在融合时并不能完全削减这部分噪声。以颜色直方图方式统计前后图像内颜色的占比,设当前帧的颜色直方图为 $Hist_i$,前一帧为 $Hist_{i-1}$,若相邻 2 帧相似度高,则 2 帧之间的相似性可以描述为

$$D = \sum_{j=1}^N \min(Hist_i^j, Hist_{i-1}^j) \tag{7}$$

式中: j 为颜色分量; N 为总颜色量级数目; $j \in N$; D 为 2 帧的相似性表示。

因此,构建加权函数 W 如下:

$$W = \frac{\sin\left(2D\pi - \frac{3\pi}{2}\right)}{2} \tag{8}$$

最终融合时,按式(9)进行:

$$M = W_i M_{i-1} + (1 - W_i) M_i \tag{9}$$

式中: M_{i-1} 表示前一帧图像的运动显著性结果; M_i 表示当前帧的运动显著性结果; M 为最终运动显著图。

图 1 展示了经过调整前与调整后的动态显著

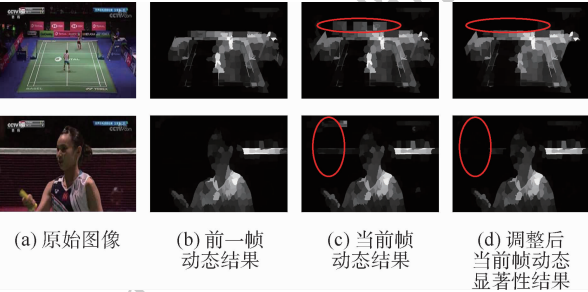


图 1 动态显著图调整效果前后对比

Fig. 1 Effect comparison of dynamic saliency map before and after adjustment

图效果,并用椭圆形标出了调整后产生改进的区域。

2.5 显著性图像计算及自适应融合

设 M_i 为运动显著结果, J_i 为静态显著结果,通过 dif_i 将权重系数归一化至 $[0,1]$,如下:

$$dif_i = \max(M_i) - \text{median}(J_i) \tag{10}$$

并通过式(11)完成融合:

$$Saliency_i = [(1 - dif_i) J_i + dif_i \cdot M_i] \sigma + (1 - \sigma) dif_i \cdot J_i \tag{11}$$

式中: σ 为比例系数,本文设为 0.4。

显著结果自适应融合如图 2 所示。

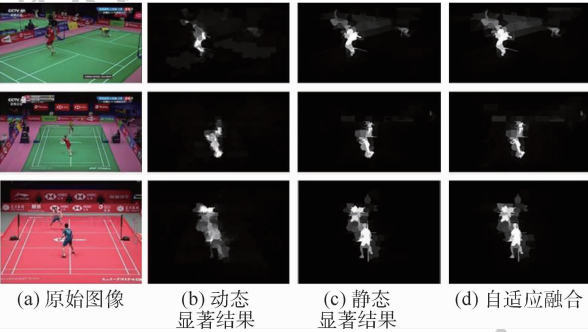


图 2 显著结果自适应融合

Fig. 2 Adaptive fusion of saliency results

3 关键帧提取

在关键帧提取前,对连续视频图像序列进行显著性目标检测的目的是为了能够预先完成显著前景目标的提取,同时提升关键帧提取阶段的检测准确率与效果。本节主要方法内容和整体技术框架如图 3 所示。

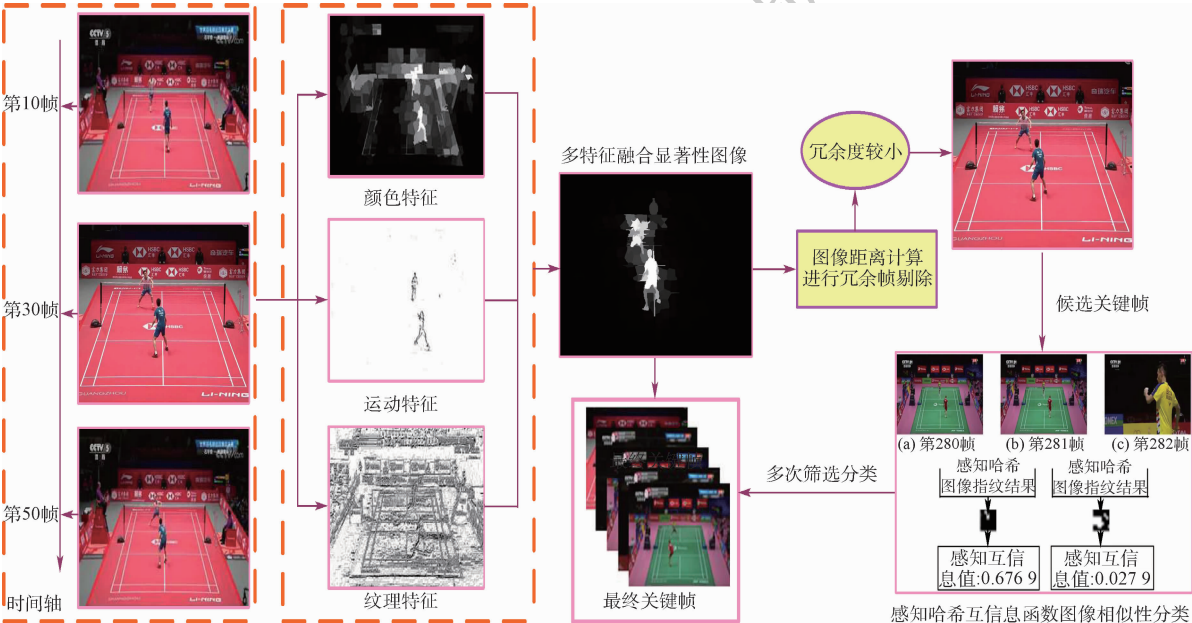


图 3 关键帧提取主要方法内容和整体技术框架

Fig. 3 Main method content and overall technical framework of key frame extraction

3.1 LBP 纹理特征提取

通过尺寸为 3×3 的基准检测窗口进行 0、1 标记,形成二进制的特征值结果,如下:

$$\text{LBP}_1(x_i, y_i) = \sum_{j=1}^8 s(g(j) - g(i)) 2^k \quad (12)$$

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & \text{其他} \end{cases} \quad (13)$$

式中: $g(j)$ 为基准窗口内第 j 个点的灰度; $g(i)$ 为中心点对应的值。

根据旋转不变原则,最终的特征值为

$$\text{LBP}_{\min} = \min(\text{LBP}_x | x = 0, 1, 2, \dots, j-1) \quad (14)$$

3.2 显著性检测效果增强及颜色特征提取

取显著性检测结果与原始视频中同一帧图像进行对应相乘。先得到原图像 R, G, B 分量, 设为 f_R, f_G, f_B , 设当前帧显著性结果图像为 F , 将结果 F 分别与 f_R, f_G, f_B 对应相乘, 得到增强的显著性结果 F' , 如图 4 所示。

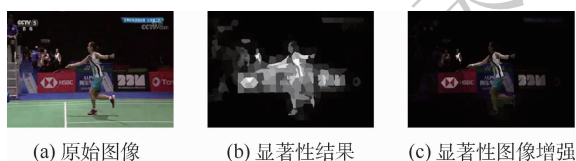


图 4 显著性检测效果增强结果

Fig. 4 Enhancement results of saliency detection effect

将图像从 RGB 空间转化为 HSV 空间, 再分别提取 H, S, V 三个分量结果, 将 H, S, V 三通道按照 16:4:4 的等级进行量化; 之后将所有的颜色分量按比例等级融合形成特征矢量, 如下:

$$L = HQ_s Q_v + SQ_v + V \quad (15)$$

式中: Q_s 和 Q_v 为量化的等级。

将颜色空间进行量化, 并将量化后的颜色根据式(15)进行融合并映射, 映射的像素值范围为 $[0, 255]$, 得到相乘图颜色信息。

3.3 哈希函数相似性计算

基于内容的感知哈希 (Perceptual Hash) 函数是一种用于获取图像哈希值, 并用其来描述的特征相似性表示方法。根据值的对比计算, 可以得到 2 幅图像基于汉明距离的相似性程度结果。

调整图像分辨率大小统一至 32×32 范围, 并转换多通道图像为单通道, 完成离散余弦变换 (DCT)。此步是为了将图像具有的位置特征转移至频域当中, 并能够对图像进行良好的压缩, 且保持无损转换。基于 DCT 的对称变换方式, 待图像转为频域下的特征编码后, 反方向 DCT 得到原先的特征信息, 具体的 DCT 变换如式(16)、式(17)所示。

$$F(u, v) = c(u) c(v) \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} f(i, j) \cdot \cos\left[\frac{(i+0.5)\pi}{N} u\right] \cos\left[\frac{(j+0.5)\pi}{N} v\right] \quad (16)$$

$$c(u) = \begin{cases} \sqrt{\frac{1}{N}} & u = 0 \\ \sqrt{\frac{2}{N}} & u \neq 0 \end{cases} \quad (17)$$

式中: $F(u, v)$ 表示 DCT 变换结果; N 为像素点的数量; $c(u) c(v)$ 为 DCT 变换中的正交变换矩阵; $f(i, j)$ 为 DCT 变换前原始像素点的值。

在得到 32×32 大小的图像区域后, 只取图像矩阵中左上角位置中大小为 8×8 区域的像素点坐标值矩阵, 此部分区域可以表示整幅图像当中频率信息最低的区域。按式(18)计算图像区域内 64 个像素点的平均像素大小 $\bar{p}'$:

$$\bar{p}' = \left(\sum_{i=1}^n p_i \right) / n \quad (18)$$

式中: p_i 为像素点大小。

计算比较像素点与 $\bar{p}'$ 的大小得到完整的 pHash 值, 即图像的感知哈希值, 并计算 2 个不同图像间 pHash 值的汉明距离。

3.4 基于哈希函数互信息计算

互信息是指在 2 个不同的个体中, 相互之间包含对方信息内容数量多少的描述, 属于信息论理论范畴, 其基于熵的概念来进行 2 个物体的计算, 公式如下:

$$H(X) = - \sum_{i=1}^n p(x_i) I(x_i) = - \sum_{i=1}^n p(x_i) \ln p(x_i) \quad (19)$$

$$I(x_i) = - \sum_{i=1}^n \ln p(x_i) = \sum_{i=1}^n \ln \left(\frac{1}{p(x_i)} \right) \quad (20)$$

式中: $p(x_i)$ 为基于事件 x_i 的概率数值; $I(x_i)$ 为具体内容信息的量。则 2 个随机变量间熵对应的联合信息期望值可表示为

$$I(x_i, y_j) = \sum_{i=1}^n \ln \left(\frac{p(x_i, y_j)}{p(x_i) p(y_j)} \right) \quad (21)$$

若 (x, y) 对应的联合分布用 $p(x, y)$ 表示, 对应的边缘分布用 $p(x), p(y)$ 表示, 联合与乘积分布 $p(x, y), p(x) p(y)$ 相对熵的结果即为 $I(x, y)$, 即互信息:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \ln \frac{p(x, y)}{p(x) p(y)} \quad (22)$$

根据式(20), 得到对于图像中的互信息为

$$I(X, Y) = H(X) + H(Y) - H(XY) \quad (23)$$

本文基于感知哈希函数进行图像相似性计算, 结合连续前后帧图像的 pHash 图像, 进行相似

性计算。

图3中展示了2018年羽毛球汤姆斯杯丹麦对阵马来西亚比赛视频中第280帧、第281帧、第282帧图像中的感知哈希图像结果。其中,第280帧和第281帧属于同一镜头下的相似场景图像,而第281帧及第282帧图像则发生了跳变,图中展示了这连续3帧图像的互信息值变化。

3.5 摘要化生成

将得到的所有图像的纹理信息、所有显著增强图像的颜色信息及光流法得到的光流信息进行融合,将每个特征图像对应的图像矩定义为矩阵 A 、 B 、 C 并按列拼接,形成融合的特征向量矩阵 $[A \ B \ C]$ 。计算前后2个特征矩阵之间的欧氏距离,其距离定义为其中所有元素间欧氏距离的累加和。根据欧氏距离计算特征矩阵之间的相似性,以距离的平均值进行冗余帧剔除,小于距离平均值的帧被舍弃。

根据自适应感知哈希互信息值进行相似关键帧分类。第1步,得到自适应 MI 值。计算所有连续相邻图像帧的互信息值,得到互信息值集合 $MIS = \{MI_1, MI_2, \dots, MI_{N-1}\}$, 设前 m 个数量的 $I(X, Y)$ 平均值结果为 $\overline{MI}_1$, 方差为 E_1 , 后 $N - m - 1$ 个 $I(X, Y)$ 的平均值结果为 $\overline{MI}_2$, 方差为 E_2 , 则

$$\overline{MI}_1 = \frac{\sum_{i=1}^m MI_i}{m} \quad (24)$$

$$\overline{MI}_2 = \frac{\sum_{i=m+1}^{N-1} MI_i}{N - 1 - m} \quad (25)$$

$$E_1 = \frac{\sum_{i=1}^m (MI_i - \overline{MI}_1)^2}{m} \quad (26)$$

$$E_2 = \frac{\sum_{i=m+1}^{N-1} (MI_i - \overline{MI}_2)^2}{N - 1 - m} \quad (27)$$

假设满足式(26)的最小 $I(X, Y)$ 值为筛选分界值:

$$\omega = mE_1^2 + (N - 1 - m)E_2^2 \quad (28)$$

小于此值的图像帧分为一类,否则另归为新的一类,得到划分后的关键帧类别序列 $f^2 = \{\text{Cluster}_1, \text{Cluster}_2, \dots, \text{Cluster}_n\}$ 。

设已划分的2个相邻关键帧集合 Cluster_1 与 Cluster_2 的互信息值为 $MI(\text{Cluster}_i, \text{Cluster}_{i+1})$, 则 $MI(\text{Cluster}_i, \text{Cluster}_{i+1}) =$

$$\frac{\sum_{x=1}^{n_1} \sum_{y=n_1+1}^{n_2} MI(\text{Cluster}_i, \text{Cluster}_{i+1})}{n_1 n_2} \quad (29)$$

式中: n_1 和 n_2 分别为关键帧类别 Cluster_i 和 Cluster_{i+1} 中图像的数量; MI 为结果,即当前关键帧类别集合与关键帧类别集合的图像互信息值,定义为2个集合中所有图像与除本身之外的所有其他图像的互信息值和的平均值。

根据特定阈值进行关键帧分类结果合并。将2个集合的互信息值与阈值进行比较,合并规则与相邻图像间的筛选方式相同,最终得到经合并后一定数量的分类集合 $f^3 = \{\text{Cluster}_1, \text{Cluster}_2, \dots, \text{Cluster}_n\}$ 。最终挑选每一集合中与集合内其余图像 MI 相比最大的图像帧为当前集合的代表。

4 实验及分析

4.1 显著性数据集

在显著性检测的过程中,实验所使用的图像数据来源于 Segtrack V2、MSRA^[3] 及 ViSal 数据集。Segtrack 是一种视频对象分割数据集,同时也作为显著性检测的数据集,Segtrack V2 是 Segtrack 数据集的扩大版本,主要包含人类奔跑、鸟类、猎豹、羚羊等动物的运动视频;MSRA 则包含约5 000张图像,囊括了各类场景图像,且都包含了真值图像 ground truth;ViSal 则同样也是用于目标检测的数据集,且其中的数据都是视频形式。

4.2 视频摘要数据集

本文所使用的数据集包括了 YouTube、OVP (Open Video Project) 等公共视频数据集。以运动视频为研究对象,采集了网上大量的室内羽毛球比赛运动场景视频。YouTube 数据集格式为 MPEG, 视频分辨率大小为 352×240 ;OVP 数据集的格式包括 FLV 和 AVI 两种,类型包括新闻、动画、广告、电视剧、比赛视频等不同的场景。室内羽毛球比赛场景视频则包括2018年世界羽联年终总决赛男单决赛、2018年羽毛球汤姆斯杯丹麦对阵马来西亚、2019年世锦赛第二轮戴资颖对阵菲迪亚尼等的比赛视频,皆为 MP4、 1280×720 分辨格式。在运动视频的用户摘要确定过程中,选择10位同学进行人工筛选,在预先不告知内容重点的情况下,得到10位同学的关键帧摘要结果,并对这10种结果中相似的视频帧计算帧数平均值得到最终的用户摘要结果,确保用户摘要的客观及合理性。

本文对多特征融入后的距离计算剔除冗余帧的过程中,使用了欧氏距离的平均值作为剔除的阈值。但有时完全通过平均值无法剔除足够数量的冗余帧,仍然会保留一部分无用的帧,该部分图像既会对之后的检测和判断产生干扰,也会降低整体的计算时间效率。根据数据集类型的不同,

在平均值的基础上加入了微调因子 α , 对 YouTube、OVP 等数据集设置 $\alpha = 2.5$, 而对比赛视频数据集设置为 $\alpha = 20$ 。

仿真实验中, 人工阈值 thresh 可以控制最终生成的摘要所包含图像数量的多少。 thresh 越高, 所产生的摘要图像数量越少; 反之越多。经过实验分析, 对 YouTube、OVP 等数据集设置 $\text{thresh} = 0.5$, 对运动视频数据集设置为 $\text{thresh} = 1.2$, 使得结果达到了最佳。

4.3 显著性检测

本文方法主要与 MR^[14]、SF^[15]、FT^[16]、GS^[17] 方法进行对比, 实验结果如图 5 所示。

MR 方法根据图像元素与给定种子或查询的相关性定义图像元素的显著性, 将图像表示为以超像素为节点的闭环图, 以提取背景区域和前景显著对象; SF 方法采用基于对比度的显著性估计抽象出不必要的细节, 从元素对比中得出显著性度量, 以将前景和背景分开; FT 方法可输出具有清晰定义的显著对象边界的全分辨率显著图, 通过保留原始图像中更多的频率内容, 以保留这些边界; GS 方法使用背景先验来计算图像的测地线显著性区域。

从图 5 可知, MR 方法无法有效区分物体中心和图像之间的类别差异; SF 方法对于背景明亮的物体同样也会检测出来, 抗干扰性较差; FT 方法可以快速地检测出图中不同物体间的频谱差,

但无法增加具体的细节信息, 只能体现对比度信息; GS 方法则与之前方法相比效果有所提升, 但仍然会有较大的噪声干扰存在。本文方法不但能够准确定位目标, 对于前后背景对比度差异较小的图像也能较好地地区分出前景目标。本文以 Seg-track V2 和 MSRA 为数据展示显著性的 F-measure 柱状结果差别, 如图 6 所示。

4.4 实验结果及分析

本文选用不同类别视频, 如 OVP 包含的新闻或比赛场景等, 并将结果与现有的几种经典视频摘要或关键帧提取算法 (如 OV^[18]、VSUMM^[19]、STIMO^[20]、SD^[21]、KBKS^[22] 等) 进行对比。图 7 为使用公共摘要数据集中的“v20.flv”动画视频及“v101.flv”新闻视频得出的摘要化结果。

从图 7 可以看到, 在“v20.flv”和“v101.flv”视频中, OV 算法在左侧视频中的第 1、4 帧产生了冗余, 右侧视频中第 5 帧图像出现了冗余, 同时在左侧视频中的第 2、5、6、8 帧出现了检测不准确的情况, 检测结果较差; 而在 VSUMM 算法中, 左右 2 段视频中误检的情况比 OV 算法相比较少, 但仍有冗余的情况; 在 STIMO 算法中, 左侧视频有 8 帧图像命中了用户摘要结果, 比 OV、VSUMM 两种算法的结果要更好一些, 但检测出的图片数量较多, 也包含了一些冗余的图像; 而在 SD 算法与 KBKS 算法中, 得到的结果基本都包含了用户摘要, 只有少量误检的情况, 但得到的摘要数量少于

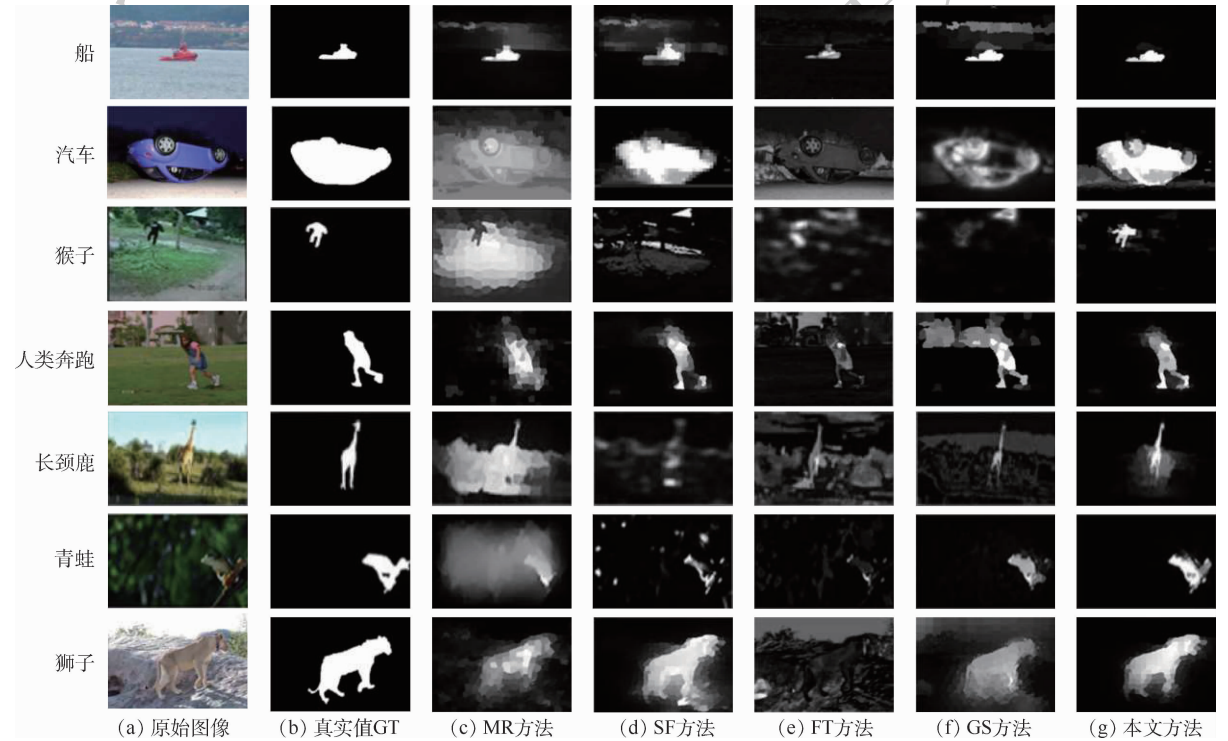


图 5 数据集在不同方法上的显著性图比较

Fig. 5 Comparison of saliency maps of datasets among different methods

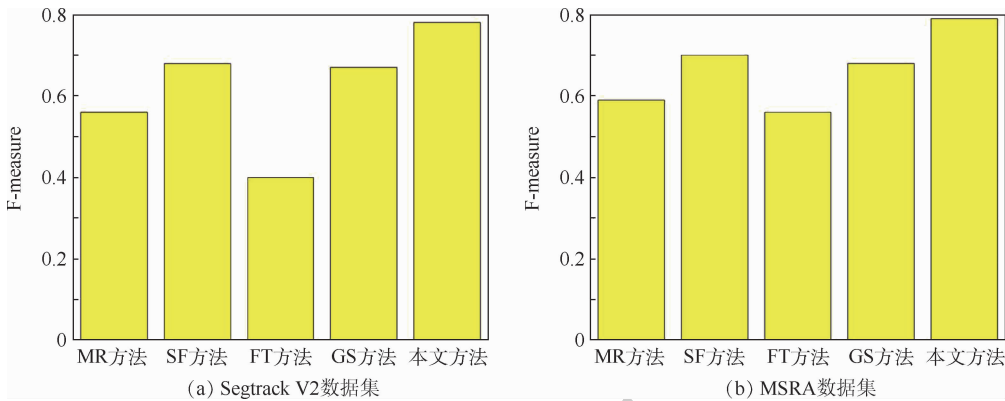


图 6 F-measure 在不同数据集上的情况
Fig.6 F-measure on different datasets

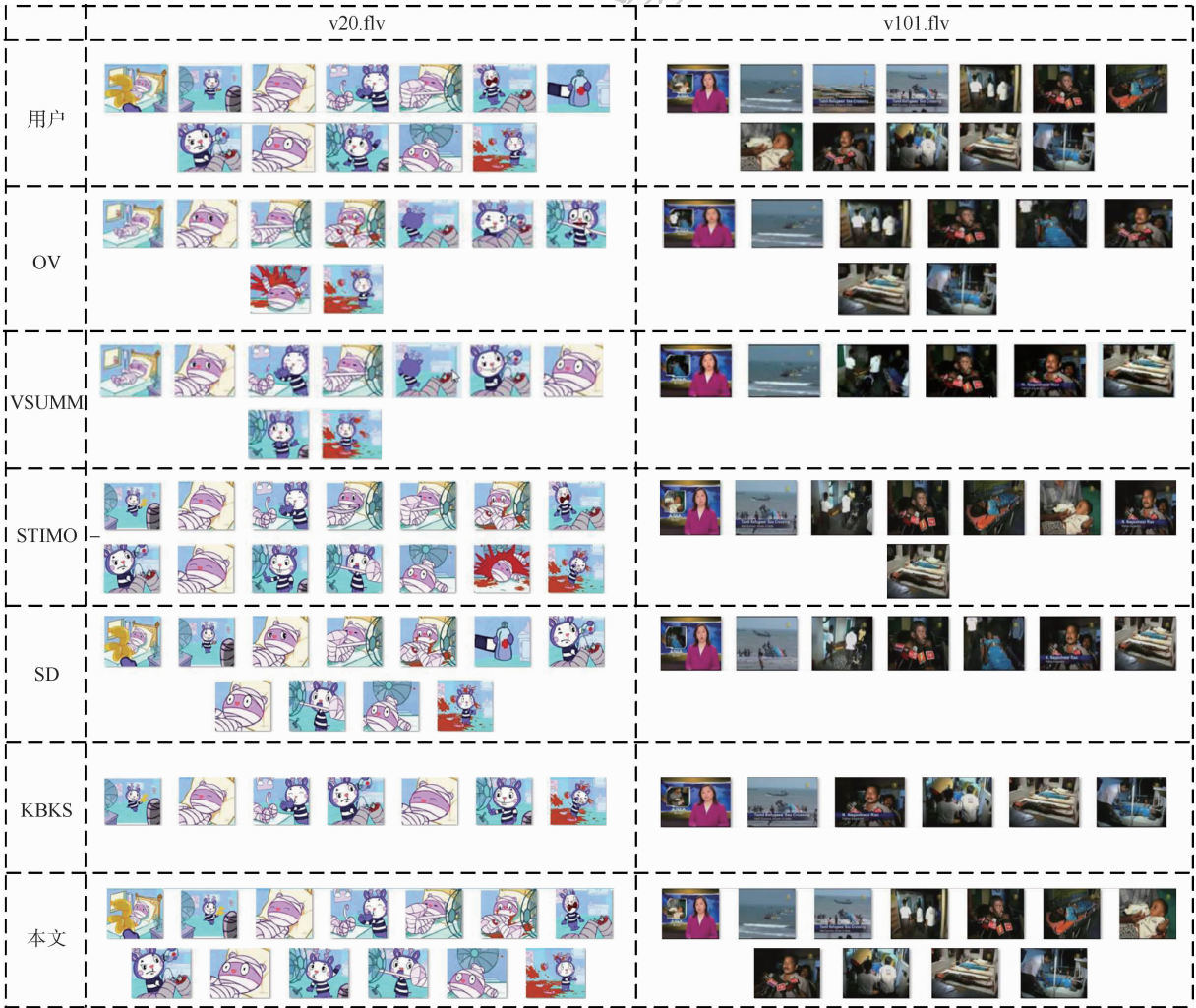


图 7 视频“v20. flv”及“v101. flv”在不同摘要算法下的结果
Fig.7 Results of video “v20. flv” and “v101. flv” under different summarization algorithms

真实用户标注结果,无法完整地描述视频的主要内容;在本文中左侧视频产生了 13 帧摘要结果,其中 11 帧命中了用户摘要,同时只产生了 2 帧冗余,在右侧视频中,得到 11 帧摘要结果,其中 11 帧命中摘要,只漏掉了用户摘要中的第 3 帧,表现出了良好的摘要化结果。

在图 8 中,以 2018 年世界羽联年终总决赛男

单决赛及 2018 年羽毛球汤姆斯杯丹麦对阵马来西亚两段视频为准,进行结果分析。由于电视中的运动比赛视频分辨率较高,且会出现镜头停留时间较长的情况,一般方法较容易产生检测冗余的现象。在 OV 算法与 VSUMM 算法中,都得到了大于用户摘要数量的关键帧,其中对于镜头产生移动,但基本内容没有产生变化的图像,2 种算

法都得到了不同数量的相似场景冗余帧。

在 OV 算法与 VSUMM 算法中,左侧视频的第 1、2 帧及右侧的第 4、5、8 帧都得到了不同数量的相似场景冗余帧;STIMO 算法也是相同;SD 算法与 KBKS 算法则出现了更多;在本文方法中,左侧视频虽然得到的摘要结果数量不足,但 9 帧的结果中都命中用户摘要,只有漏检的情况发生,而在右侧视频中得到了足够数量的图

像结果,用户摘要共 13 帧,其中有 11 帧命中用户结果,说明了本文方法的高准确性及低误检率。

为了更加直观地描述不同方法在公开数据集及室内羽毛球比赛运动场景视频数据上的表现,使用准确率、错误率、漏检率、精度、召回率和 F-measure 等指标,分别对实验数据视频依次对比,如表 1 所示。

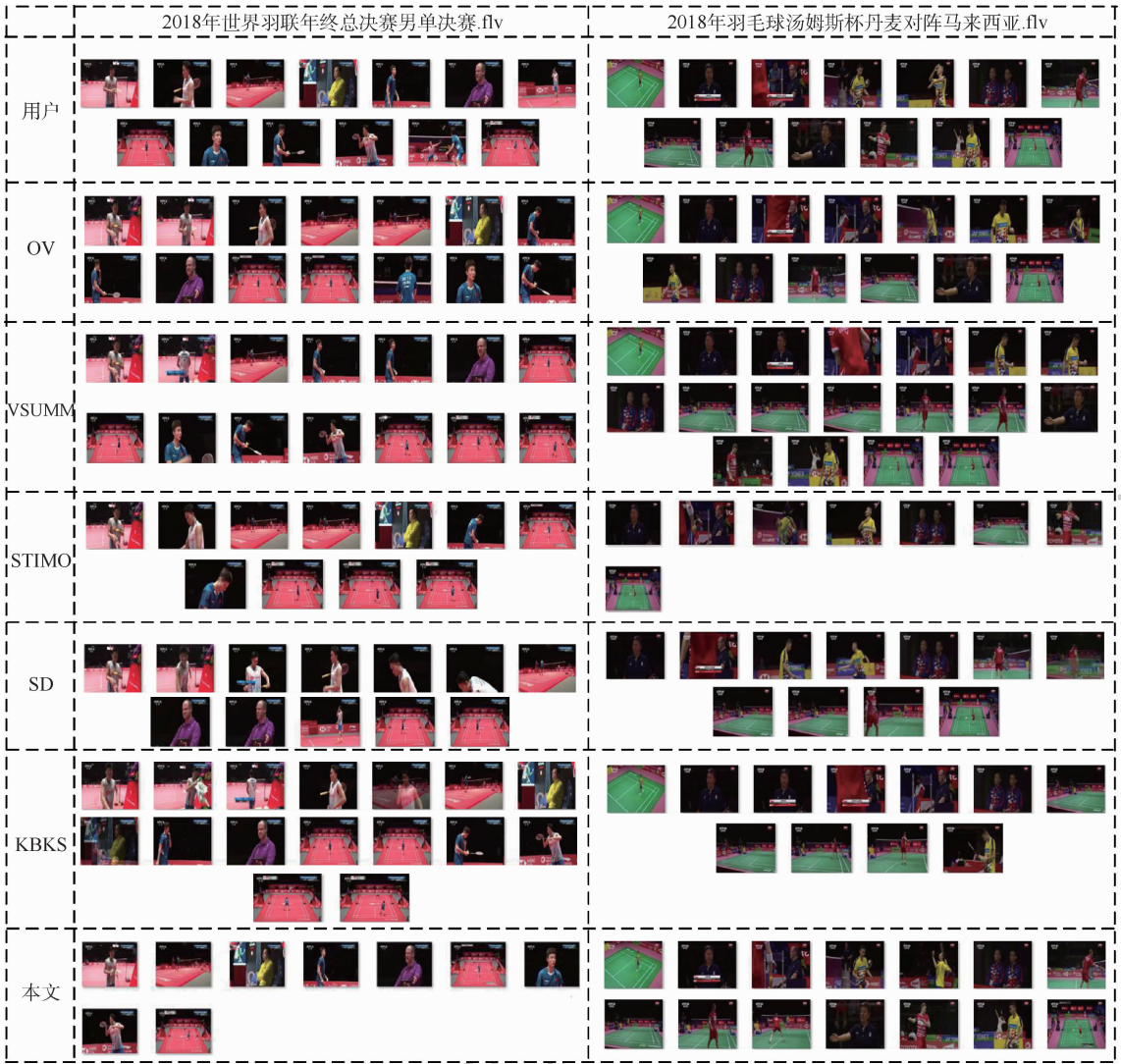


图 8 运动视频在不同摘要算法下的结果

Fig. 8 Results of sports video under different summarization glgorithms

表 1 运动视频在不同摘要算法下的对比

Table 1 Comparison of sports videos under various summarization algorithms						
算法	准确率	错误率	漏检率	精度	召回率	F-measure
OV	0.58	0.08	0.42	0.88	0.58	0.7
VSUMM	0.42	0.08	0.58	0.83	0.42	0.56
STIMO	0.67	0.08	0.33	0.89	0.67	0.76
SD	0.33	0.25	0.67	0.57	0.33	0.42
KBKS	0.5	0.08	0.5	0.86	0.5	0.63
本文	0.92	0	0.08	0.92	0.92	0.92

5 结 论

本文以视觉显著性模型为基础,在中心-环绕比金字塔检测模型中融合超像素分割以加速大分辨率图像视频数据的计算,同时选择双对比度方式计算、提取图像中更多的特征信息;在计算运动显著图时,通过前后帧结果抑制的方式生成更佳的动态结果,并使用自适应方式融合得到效果良好的显著结果,良好的显著性结果也为在视频摘

要与关键帧选择的过程提供了更多的图像信息。

1) 在摘要化生成中,结合了显著性检测及感知哈希互信息方式进行提取。在显著性检测的基础上,利用图像的纹理及颜色等信息,对显著性检测的结果图像进行了二次特征提取并完成相邻图像之间的相似性判断,进行多次冗余帧的剔除及相似图像的类别划分,从而得到属于不同类别的、能够最大程度描述视频的结果。

2) 在公共数据集及本文的运动视频数据集上进行了效果对比,验证了本文方法对视频摘要化生成的良好效果和较优表现。

参考文献 (References)

- [1] 唐铭谦. 基于对象的监控视频摘要算法研究[D]. 西安:西安电子科技大学,2018:1-3.
TANG M Q. The research of surveillance video synopsis algorithm based on objects[D]. Xi'an: Xidian University, 2018: 1-3 (in Chinese).
- [2] 刘全, 翟建伟, 钟珊, 等. 一种基于视觉注意力机制的深度循环 Q 网络模型[J]. 计算机学报, 2017, 40(6): 1353-1366.
LIU Q, ZHAI J W, ZHONG S, et al. A deep recurrent Q-network based on visual attention mechanism [J]. Chinese Journal of Computers, 2017, 40(6): 1353-1366 (in Chinese).
- [3] 郎洪, 丁朔, 陆健, 等. 复杂场景下的交通视频显著性前景目标提取[J]. 中国图象图形学报, 2018, 24(1): 50-63.
LANG H, DING S, LU J, et al. Traffic video significance foreground target extraction in complex scenes[J]. Journal of Image and Graphics, 2018, 24(1): 50-63 (in Chinese).
- [4] 张芳, 王萌, 肖志涛, 等. 基于全卷积神经网络与低秩稀疏分解的显著性检测[J]. 自动化学报, 2019, 45(11): 2149-2158.
ZHANG F, WANG M, XIAO Z T, et al. Saliency detection via full convolution neural network and low rank sparse decomposition[J]. Acta Automatica Sinica, 2019, 45(11): 2149-2158 (in Chinese).
- [5] 李庆武, 马云鹏, 周亚琴, 等. 基于无监督栈式降噪自编码网络的显著性检测算法[J]. 电子学报, 2019, 47(4): 871-879.
LI Q W, MA Y P, ZHOU Y Q, et al. Saliency detection based on unsupervised SDAE network [J]. Acta Electronica Sinica, 2019, 47(4): 871-879 (in Chinese).
- [6] 陈炳才, 陶鑫, 陈慧, 等. 融合边界连通性与局部对比性的图像显著性检测[J]. 计算机学报, 2020, 43(1): 16-28.
CHEN B C, TAO X, CHEN H, et al. Saliency detection via fusion of boundary connectivity and local contrast [J]. Chinese Journal of Computers, 2020, 43(1): 16-28 (in Chinese).
- [7] ABLAVATSKI A, LU S, CAI J. Enriched deep recurrent visual attention model for multiple object recognition [C] // Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2017: 971-978.
- [8] QU S, XI Y, DING S. Visual attention based on long-short term memory model for image caption generation [C] // Proceedings of Chinese Control and Decision Conference (CCDC). Piscataway: IEEE Press, 2017: 4789-4794.
- [9] LIU G H, YANG J Y. Exploiting color volume and color difference for salient region detection [J]. IEEE Transactions on Image Processing, 2019, 28(1): 6-16.
- [10] LI Z, TANG J, WANG X, et al. Multimedia news summarization in search [J]. ACM Transactions on Intelligent Systems and Technology, 2016, 7(3): 1-20.
- [11] HU T L, LI Z C. Video summarization via exploring the global and local importance [J]. Multimedia Tools and Applications, 2018, 77(17): 22083-22098.
- [12] MENG J, WANG S, WANG H, et al. Video summarization via multi-view representative selection [C] // Proceedings of IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2017: 1189-1198.
- [13] ACHANTA R, SHAJI A, SMITH K, et al. SLIC superpixels compared to state-of-the-art superpixel methods [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(11): 2274-2282.
- [14] YANG C, ZHANG L, LU H, et al. Saliency detection via graph-based manifold ranking [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2013: 3166-3173.
- [15] PERAZZI F, KRAHENBUHL P, PRITICH Y, et al. Saliency filters: Contrast based filtering for salient region detection [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2012: 733-740.
- [16] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2009: 1597-1604.
- [17] WEI Y C, WEN F, ZHU W, et al. Geodesic saliency using background priors [C] // Proceedings of European Conference on Computer Vision (ECCV). Berlin: Springer, 2012: 29-42.
- [18] DEMENTHON D, KOBLA V, DOERMANN D. Video summarization by curve simplification [C] // Proceedings of ACM International Conference on Multimedia. New York: ACM Press, 1998: 211-218.
- [19] DE AVILA S E F, LOPES A P B, DA LUZ A, et al. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method [J]. Pattern Recognition Letters, 2011, 32(1): 56-68.
- [20] FURINI M, GERACI F, MONTANGERO M, et al. STIMO: Still and moving video storyboard for the web scenario [J]. Multimedia Tools and Applications, 2010, 46(1): 47-69.
- [21] CONG Y, YUAN J, LUO J. Towards scalable summarization of consumer videos via sparse dictionary selection [J]. IEEE Transactions on Multimedia, 2012, 14(1): 66-75.
- [22] GUAN G, WANG Z, LU S, et al. Keypoint based keyframe selection [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 23(4): 729-734.

作者简介:

金海燕 女, 博士, 教授, 博士生导师, CCF 会员。主要研究方向: 计算机视觉、图像处理、智能信息处理等。

曹甜 女,硕士研究生。主要研究方向:计算机视觉、图像处理等。

肖聪 男,硕士研究生。主要研究方向:计算机视觉、图像处理等。

肖照林 男,博士,副教授,硕士生导师,CCF 会员。主要研究方向:计算机视觉、计算摄影学等。

Video summary generation based on multi-feature image and visual saliency

JIN Haiyan^{1,2}, CAO Tian¹, XIAO Cong¹, XIAO Zhaolin^{1,2,*}

(1. College of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China;
2. Shaanxi Key Laboratory for Network Computing and Security Technology, Xi'an 710048, China)

Abstract: How to extract video content efficiently, that is, video summarization, is a research hotspot in the field of computer vision. Video summary cannot be obtained effectively and completely by simply detecting the image color, texture and other features. Based on the visual attention pyramid model, this paper proposes an improved center-surround video summarization method with variable ratio and double contrast calculation. First, the video image sequence is divided into pixel blocks by superpixel method to speed up image calculation. Then, the contrast feature difference under different color backgrounds is detected and fused. Finally, combined with the optical flow motion information, the static and dynamic saliency results are merged to extract the key frames of the video. When extracting the key frames, the perceived Hash function is used to perform similarity judgment to complete the video summary generation. Simulation experiments are carried out on Segtrack V2, ViSal and OVP datasets. The experimental results show that the proposed method can be used to effectively extract the area of interest, and finally obtain the video summary expressed by the sequence of key frame images.

Keywords: video summarization; visual attention pyramid; visual saliency; key frame extraction; similarity judgment

Received: 2020-08-31; Accepted: 2020-10-27; Published online: 2020-11-03 09:44
URL: kns.cnki.net/kcms/detail/11.2625.V.20201102.1036.002.html
Foundation items: Technology Innovation Leading Program of Shaanxi (2020CGXNG-026); Natural Science Basic Research Program of Shaanxi (2019JM-221)
* Corresponding author. E-mail: xiaozhaolin@xaut.edu.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0455

基于混合网络的异源遥感图像变化检测

周圆^{1,*}, 李祥瑞², 杨晶¹

天津大学 电气自动化与信息工程学院, 天津 300072; 2. 天津大学 国际工程师工程学院, 天津 300072)

摘 要: 为了更加准确地进行异源遥感图像的变化检测任务,提出了一种基于混合网络的异源遥感图像变化检测算法。利用伪孪生网络提取异源遥感图像块间空间维度的变化特征,利用早期融合网络提取异源遥感图像块间光谱维度的变化特征,将2支网络提取的特征进行融合,并将融合后的变化特征输入到 sigmoid 层进行二分类检测。同时,在伪孪生网络中加入对比损失函数,通过优化对比损失函数,使得在特征空间中,未变化图像对的空间特征差异更小,变化图像对的空间特征差异更大,以提升网络的区分能力和收敛速度。

关 键 词: 变化检测; 异源; 伪孪生网络; 早期融合网络; 对比损失函数

中图分类号: TP751.1

文献标志码: A

文章编号: 1001-5965(2021)03-0451-10

遥感图像变化检测是指提取不同时相同一区域的遥感图像之间变化区域的过程。变化检测是遥感领域重要的研究方向,作为监测地表状况的关键技术,在城市扩张、环境监测、灾害评估等众多领域发挥着巨大作用。

根据待检测的2幅图像是否由同一传感器获取,可将变化检测算法分为同源遥感图像变化检测和异源遥感图像变化检测2种类型。随着遥感技术的快速发展,卫星传感器已由过去的单一型发展成了如今多样化的传感器。在某些特定的紧急情况下,如灾害管理和灾后评估等,由于卫星运行需要一定的周期,仅有的遥感图像可能是通过不同的卫星传感器在目标日期获得的,这种情况下,需要快速、灵活、准确的方法来处理不同传感器获取的图像。

Alberga^[1]提出了一种利用光谱不变测度来描述多时相图像像素空间排列的差异的方法,该方法考虑了局部照度差异,不需要考虑光谱的维数,类似于比较一般单通道图像上的局部块的纹理特征。Fernandez-Prieto 和 Marconcini^[2]提出了

一种对标签描述的变化子集进行选择变化检测的参数模型,依赖于图像建模的联合概率,独立于图像的光谱特征。文献[3-4]提出了一种基于核的多源遥感数据框架,将多时相图像在光谱通道上拼接,利用核典型相关分析(KCCA)^[3]学习非线性特征变换,提高变化检测精度^[4]。基于目标的变化检测算法是一种综合图像分割和分类的集成方法^[5],适用于有显著检测目标^[6-8]的情况,如专门对建筑物的变迁检测、道路的整改检测及河流的变化检测等。Mercier 等^[9]提出了一种基于条件 copulas 的异源遥感图像变化检测算法,为了得到2幅图像的统计关系,建立了一个手工制作的变更训练区域模型。

传统的变化检测算法往往基于手工提取特征,表征效果并不理想。深度神经网络可以提取图像更深层次的非线性特征^[10],具有更好的表征性,因此出现了许多基于深度学习的变化检测算法^[11]。文献[12]提出了一种基于叠加去噪自编码(SDAEs)的变化检测算法,利用人工或现有算法获取粗糙的差异图,从差异图选取的不变特征

收稿日期: 2020-08-24; 录用日期: 2020-09-25; 网络出版时间: 2020-10-15 09:45

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201014.1906.001.html

* 通信作者。E-mail: zhouyuan@tju.edu.cn

引用格式: 周圆, 李祥瑞, 杨晶. 基于混合网络的异源遥感图像变化检测[J]. 北京航空航天大学学报, 2021, 47(3): 451-460.
ZHOU Y, LI X R, YANG J. Heterogeneous remote sensing image change detection based on hybrid network[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 451-460 (in Chinese).

对网络进行优化。文献[13]提出了一种基于孪生卷积神经网络(Siamese CNN)用来检测建筑物和树木变化的框架,先对不同传感器获得的图像进行预处理,将其转化为同一模态后再输入到孪生网络中提取变化信息。Liu 等^[14]提出了一种新的对称卷积耦合网络(SCCN),应用一种新的目标函数进行网络中参数的更新。Hughes 等^[15]利用伪孪生卷积神经网络(Pseudo-Siamese CNN)来识别异源图像中相应的小块。Niu 等^[16]使用条件生成对抗网络(cGAN)将异源遥感图像转换到一个空间,使其信息具有更一致的表示。

图像作为一种信号的载体,某个像素的特性必然和其邻域像素有着很大的联系,因此其具有较强的空间信息特性,同时多光谱遥感图像具有丰富的光谱特征。如果能够同时利用图像的空间特征和光谱特征,必然能为变化检测任务带来效果的提升,然而现有方法大多单单利用图像的空间特征或光谱特征。基于上述考虑,本文设计了一种可以同时提取图像的空间特征和光谱特征的混合卷积神经网络(HCNN)模型,实验表明了所

提算法的有效性。

1 混合网络总体结构

混合网络的总体结构如图 1 所示。其主要分为 2 支网络:分支 1 网络是伪孪生网络,利用 2 个权重不共享的 2D 卷积神经网络(2D-CNN)提取异源遥感图像块的空间特征,并将其输入到循环神经网络(RNN)学习图像块间的空间特征的变化特征;分支 2 网络是早期融合网络,由于多光谱图像具有丰富的光谱信息,为了充分利用光谱信息,先将 2 个立体图像块在光谱通道上拼接,获得具有变化信息的更多光谱通道的图像块,利用 3D 卷积神经网络(3D-CNN)学习图像块的光谱特征的变化特征。将 2 支网络输出的特征进行融合,得到综合考虑空间特征和光谱特征的变化特征,将此特征输入到 sigmoid 层进行二分类,判断该特征是否发生了变化。在伪孪生网络中加入对比损失函数,使得在特征空间中,未变化图像对的空间特征差异更小,变化图像对的空间特征差异更大,有利于提升网络的区分能力。

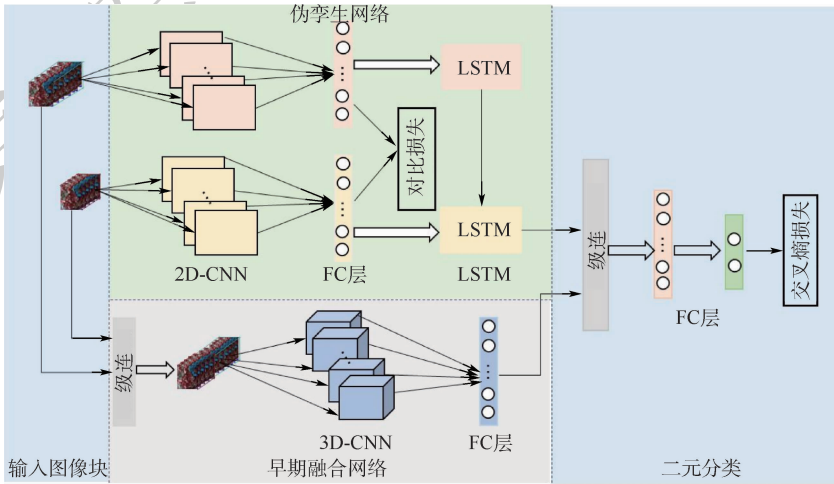


图 1 基于混合网络的异源遥感图像变化检测

Fig. 1 Change detection of heterogeneous remote sensing image based on hybrid network

2 基于伪孪生网络的空间变化特征提取

本节介绍伪孪生网络的结构,如图 1 所示。伪孪生网络包括 2D-CNN 和 RNN 两部分。利用 2D-CNN 分别提取 2 张异源遥感图像块的空间特征,再利用 RNN 提取 2 个空间特征间的变化特征。

2.1 异源遥感图像块的空间特征提取

CNN 可以有效地提取图像的空间特征,笔者认为相对简单的网络架构更适合于本文任务,主要原因如下:①变化检测的目的仅仅是区分 2 个类(变化和未变化),这比计算机视觉中一般的图

像识别问题(如涉及 1 000 个类别的 ImageNet 分类)所需的模型复杂度要低得多。②输入的图像块的空间尺寸较小,自然需要的网络层数较少。因此,本节采用的 CNN 网络是自行设计的针对变化检测的网络,结构简图如图 1 所示,网络具体配置将在 5.2 节介绍。

2.2 空间特征间的变化特征提取

变化检测任务针对不同时间在同一地点拍摄的图像进行分析,得到其中发生变化的位置,这 2 张图像间存在时间相关性,属于时间序列数据。RNN 能够处理有时间依赖的、有先后顺序的输入,上一个时刻的网络状态信息将会作用于下一个时

刻的网络状态,从而使网络表现出动态的时间特征。本文采用 RNN 对不同时刻图像块之间的依赖关系进行建模。LSTM^[17] 是一种特殊的 RNN 结

构。本文将 2D-CNN 提取的不同时刻异源遥感图像块的空间特征输入到 LSTM 中,提取 2 种空间特征间的变化特征。具体实现方式如图 2 所示。

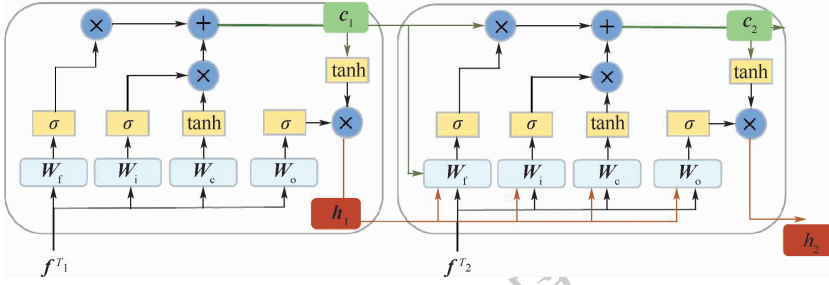


图2 变化检测任务中使用 LSTM 模型

Fig. 2 LSTM model used in change detection task

LSTM 中包含一系列时间的单元状态,当前 t 时刻 LSTM 的隐藏层输出 h_t 可按照如下方式计算:

$$h_t = o_t \times \tanh(c_t) \quad (1)$$

式中: $\tanh(\cdot)$ 为双曲正切函数; c_t 为记忆单元; o_t 为用来控制单元状态有多少信息输出到 LSTM 当前状态的输出门,其更新公式为

$$o_t = \sigma(W_o^T \times f^{T_t} + U_o^T \times h_{t-1} + b_o) \quad (2)$$

其中: W_o 和 U_o 分别为输入的权重矩阵和上一时刻输出的权重矩阵; b_o 为偏置; σ 表示 sigmoid 激活函数; f^{T_t} 表示跨传感器图像块的空间特征; T_t 为图像块获取的时间参数; T 为权重矩阵的时间参数。

记忆单元 c_t 通过是否部分丢弃当前记忆内容并添加记忆单元 $\tilde{c}_t$ 的新内容来更新,计算方式为

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t \quad (3)$$

式中: f_t 表示遗忘门; i_t 表示输入门; $\otimes$ 表示逐像素相乘。

新的记忆单元 $\tilde{c}_t$ 为

$$\tilde{c}_t = \tanh(W_c^T \times f^{T_t} + U_c^T \times h_{t-1} + b_c) \quad (4)$$

式中: W_c 和 U_c 分别为输入的权重矩阵和上一时刻输出的权重矩阵; b_c 为偏置。

i_t 代表输入门,将新的输入信息选择性地输入到记忆单元中, f_t 代表遗忘门,将记忆单元中的信息选择性地遗忘。这 2 个门的计算方式如下:

$$i_t = \sigma(W_i^T \times f^{T_t} + U_i^T \times h_{t-1} + V_i^T \times c_{t-1} + b_i) \quad (5)$$

$$f_t = \sigma(W_f^T \times f^{T_t} + U_f^T \times h_{t-1} + V_f^T \times c_{t-1} + b_f) \quad (6)$$

式中: W_i 、 U_i 和 V_i 分别为输入门的输入的权重矩阵、上一时刻输出的权重矩阵和上一时刻单元状态的输出权重; b_i 、 b_f 为偏置; W_f 、 U_f 和 V_f 分别为遗忘门的输入的权重矩阵、上一时刻输出的权重矩阵和上一时刻单元状态的输

出权重。

3 基于早期融合网络的光谱变化特征提取

本节介绍早期融合网络的结构,如图 1 所示。早期融合网络包括 3D-CNN 和全连接(FC)层两部分。先将 2 个立体图像块在光谱通道上拼接,获得具有更多光谱通道数的一个图像块,利用 3D-CNN 能够提取该图像块中光谱维度上的特征的变化。

在高/多光谱图像分类任务中,3D 卷积已被证实能够学习立体图像块的空间-光谱特征的局部信号变化^[18]。因此,在本节网络中利用 3D 卷积构建卷积层,提取图像块的光谱特征的局部信号变化,为后续的分类提供重要的判别信息。图 3 展示了 2D 卷积和 3D 卷积的主要区别,3D 卷积要比 2D 卷积多了一个运算维度。

传统的 3D 卷积定义如下:

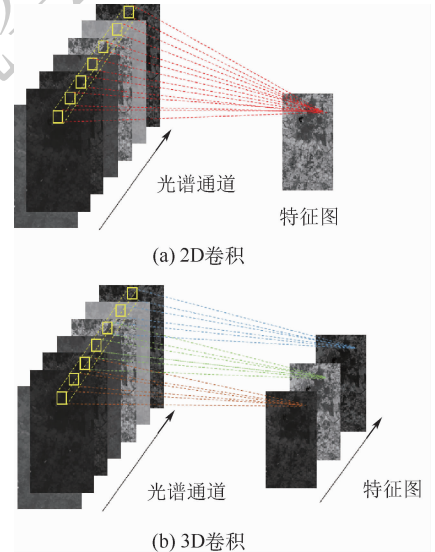


图3 2D 卷积和 3D 卷积操作的比较

Fig. 3 Comparison of 2D convolution and 3D convolution operations

$$v_{ij}^{xyz} = f \left(\sum_m \sum_{h=0}^{H_l-1} \sum_{w=0}^{W_l-1} \sum_{r=0}^{R_l-1} k_{ijm}^{hwr} v_{(l-1)m}^{(x+h)(y+w)(z+r)} + b_{ij} \right) \quad (7)$$

式中: v_{ij}^{xyz} 为特征图在 (x, y, z) 位置处的输出; f 为激活函数; H_l 、 W_l 和 R_l 分别为 3D 卷积核的长度、宽度和深度; l 为网络的层数; j 为某网络层中卷积核的个数; k_{ijm}^{hwr} 为连接到前一层中第 m 个特征块的卷积核中位置为 (h, w, r) 对应的值; b_{ij} 为偏置。

在本节提出的模型中, 每个特征块被独立对待。因此, 在上述公式中, 将 m 设置为 1, 则 3D 卷积可以被表示为

$$v_{ij}^{xyz} = f \left(\sum_{h=0}^{H_{l-1}} \sum_{w=0}^{W_{l-1}} \sum_{r=0}^{R_{l-1}} k_{ij}^{hwr} v_{(l-1)i}^{(x+h)(y+w)(z+r)} + b_{ij} \right) \quad (8)$$

式中: i 为前一层的特征块的个数; v_{ij}^{xyz} 为位置 (x, y, z) 处的输出, 通过计算第 $l-1$ 层的第 i 个特征块 $v_{(l-1)i}^{(x+h)(y+w)(z+r)}$ 和第 l 层的第 j 个核 k_{ij}^{hwr} 的卷积得到。

4 损失函数

混合网络的损失函数主要由 2 部分组成: 二元交叉熵损失函数和对比损失函数。变化检测任务属于二元分类问题, 则根据传统的二元分类网络, 选择二元交叉熵损失函数, 伪孪生网络中的 2 支子网络权重不共享, 对不同时刻提取的空间特征之间添加了额外的限制条件, 即对比损失函数^[19]。通过减小对比损失函数, 使特征空间中未变化图像对的空间特征差异更小, 变化图像对的空间特征差异更大。

令 $(\mathbf{X}^{T_1}, \mathbf{X}^{T_2})$ 表示输入图像对, y 表示类标签。 $y=0$ 表示图像对是变化的, $y=1$ 表示图像对是未变化的。二元交叉熵损失函数和对比损失函数分别用 L_{ce} 和 L_{con} 表示:

$$L_{ce} = y \ln \hat{y} + (1 - y) \ln (1 - \hat{y}) \quad (9)$$

$$L_{con} = y \|f^{T_1} - f^{T_2}\|_2^2 + (1 - y) (\max(0, m - \|f^{T_1} - f^{T_2}\|_2))^2 \quad (10)$$

式中: $\hat{y}$ 表示 sigmoid 层的输出; m 表示边缘, 通常设置 $m=1$; f^{T_1} 和 f^{T_2} 表示不同时刻获取图像块的空间特征。

整体的损失函数如下:

$$L = L_{ce} + \lambda L_{con} \quad (11)$$

式中: λ 为用来平衡分类损失和对比损失的权重。

2 种损失函数的结合使网络的训练更加稳定, 使异源遥感图像变化检测的结果更加准确。

5 实验结果与分析

5.1 实验数据集

1) Bastrop 复杂火灾数据集

数据集^[20]包含 3 组由 3 种不同类型的传感

器获得的图像对, 已经完成几何配准等预处理。图 4(a)~(d) 显示了待检测的 4 张图像 (烧毁区域用蓝色标出), 空间大小为 1534×808 , 具有不同的光谱波段。

① L5T1 vs. L5T2 数据。图 4(a)、(b) 为同一传感器 Landsat 5 (L5) 分别在 2011 年 8 月 26 日 (T1) 和 2011 年 9 月 11 日 (T2) 获取, 区别是夏季 16 天的时间间隔内的烧伤疤痕。图 4(e) 为参考的真实变化图像。

② L5T1 vs. ALI2 数据。图 4(c) 为另一个传感器 Landsat Advanced Land Imager (ALI) 在 2011 年 9 月 12 日获取。变化仍然是火灾导致的烧伤疤痕。图 4(e) 为参考的真实变化图像。

③ L5T1 vs. L8T2 数据。图 4(d) 为另一个传感器 Landsat 8 (L8) 在 2013 年 6 月 28 日获取。除了火灾导致的烧伤疤痕以外, 还涉及到自然因素导致的变化, 这些变化与火灾导致的变化光谱类别相同。图 4(f) 为参考的真实变化图像。

2) Onera 变化检测数据集

本文使用数据集^[21]中的香港图像对和孟买图像对进行变化检测, 主要变化是城市增长。图 5(a)、(b) 分别在 2016 年 9 月 27 日和 2018 年 3 月 23 日由同一传感器获取的香港图像对。图 5(c) 为参考图像, 图像大小为 693×538 , 这里选取了 RGB + IR 四个光谱通道。图 5(d)、(e) 分别在

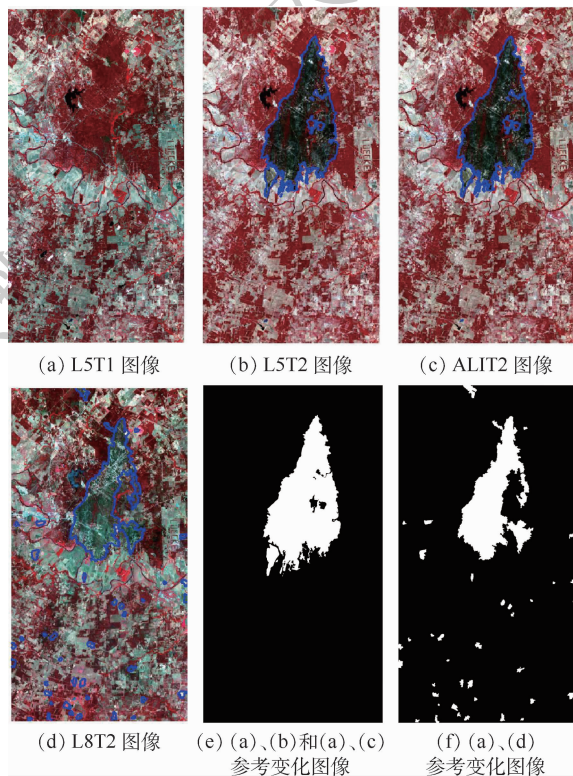


图 4 Bastrop 复杂火灾数据集

Fig. 4 Bastrop complex fire dataset

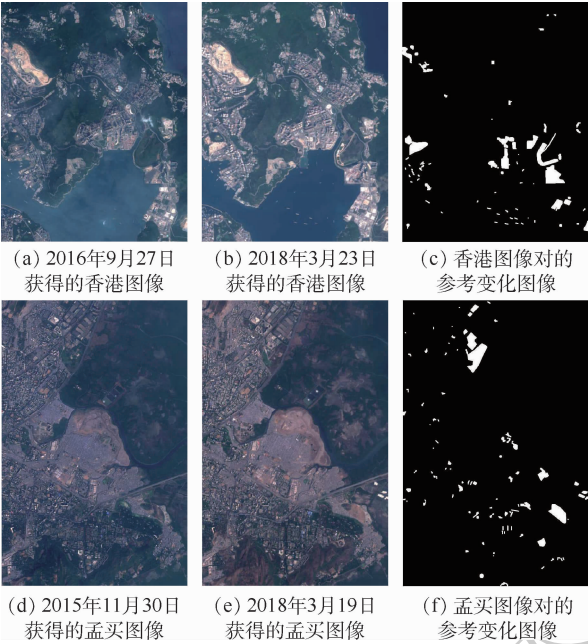


图 5 Onera 变化检测数据集
Fig.5 Onera change detection dataset

2015 年 11 月 30 日和 2018 年 3 月 19 日同一传感器获取的孟买图像对。图 5(f) 为参考图像,图像大小为 858×557 ,这里选取了 RGB + IR 四个光谱通道。

5.2 实验设置

1) 训练集。从变化和未变化的区域中分别各随机选取 900 个像素点对,以这些像素点为中心,与周围邻近像素点构成大小为 $9 \times 9 \times c_1(c_2)$ 的图像块, c_1 和 c_2 分别为时间 T_1 和 T_2 获取的图像的光谱通道数。

2) HCNN 网络架构。图 6 展示了网络配置的示意图。 T_1 和 T_2 图像块通过一些卷积层,其中卷积核的大小为 3×3 ,而不是 5×5 或者 7×7 。这样做的好处是提高网络的非线性,使得网络更加具有判别性。卷积的步长为 1,填充为 0。所有的卷积和 FC 层都用 ReLU 作为激活函数,除了最后一层 FC 层利用 sigmoid 作为激活函数。

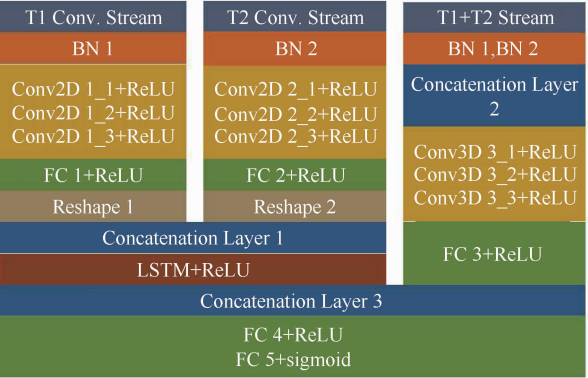


图 6 HCNN 的网络配置示意图

Fig.6 Schematic diagram of HCNN network configuration

3) 网络训练。本节提出的混合网络基于 Keras 深度学习框架进行训练,并利用 Adam 作为优化器。初始的学习率设置为 0.01。所有的网络权重都使用 he_normal 初始化器进行初始化,该初始化器从正态分布中抽取样本,使用 sigmoid 函数作为输出层的激活函数,进行二元分类。本节网络的训练硬件平台是单个有 12GB 内存的 NVIDIA GeForce GTX TITAN。

5.3 实验指标

遥感图像变化检测结果分析常采用定性和定量相结合的方式。从定性的视觉图像分析,可以大致观测出检测结果图和实际地物变化参考图是否存在较大差异。从定量的评价指标分析,可以根据精确的数字结果来比较不同算法的优劣。

本文中 TP 和 TN 分别表示被正确检测的变化像素个数和被正确检测的未变化像素个数,FP 表示虚检数,指参考图中未发生变化但在实际的变化检测结果图发生变化的像素个数,FN 表示漏检数,指参考图中发生变化但在实际的变化检测结果中未发生变化的像素个数。由上述指标已经扩展得到了很多指标,其中,以总体分类精度 (Overall Accuracy, OA) 和 Kappa 系数应用最为广泛。这 2 个值越大,变化检测效果越好。

OA 值等于被正确检测的像素数与总像素数的比值,具体计算方式如下:

$$OA = \frac{TP + TN}{TP + TN + FN + FP} \tag{12}$$

AA 值等于各个类别中分类正确像素数与总像素数的比值的平均值,具体计算方式如下:

$$AA = \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \tag{13}$$

Kappa 系数是另外一种表征分类精度的方法,等于采用某算法分类与完全随机的分类产生错误减少的比例,反映出变化检测结果和参考图的一致性,可用如下公式表示:

$$Kappa = \frac{OA - PRE}{1 - PRE} \tag{14}$$

式中:PRE 为预期分割的比例,是偶然期望的比例,按照如下方式计算:

$$PRE = \frac{(TP + FN)(TP + FP) + (FP + TN)(FN + TN)}{(TP + TN + FP + FN)^2} \tag{15}$$

5.4 不同算法的对比

1) Bastrop 复杂火灾数据集上的实验

为了验证本文提出的 HCNN 模型在异源传感器获得的遥感图像变化检测任务上的优秀表

现,在 Bastrop 复杂火灾数据集上将其与使用广泛的变化检测算法进行了对比,包括传统和基于神经网络的算法:MAD + CVA^[4]、KCCA + CVA^[20]、DCCA + CVA^[22]、DCCAE + CVA^[23]、早期融合网络(Early-Fusion network, EF)^[24]、伪孪生网络(Pseudo-Siamese network, PS)^[15]。为了加快处理进程,以 5 个像素点作为提取步长,并利用投票机制来预测图像中所有像素点的标签。每个像素点的标签由其被覆盖的所有图像块在网络的输出按权重进行投票。具体的投票权重分布符合以该像素为中心的二维高斯分布,即某个像素离该像素点越近,则权重就更大。

利用上述算法在 OA 和 Kappa 系数指标下得到的针对 Bastrop 复杂火灾数据集的结果分别如表 1 和图 7 所示。这 3 组实验的变化检测任务对算法的挑战性是越来越大的,原因在于:L5T1 vs. L5T2 实验中采用的是同一传感器在不同时刻所获取的图像对,图像对中所发生的变化主要是火灾的烧伤疤痕;L5T1 vs. ALIT2 实验中采用的是不同传感器在不同时刻所获取的图像对,图像对中所发生的变化依然是火灾的烧伤疤痕,异源遥感图像增加了变化检测的困难程度;L5T1 vs. L8T2 实验中不仅采用了不同传感器在不同时刻所获取的图像,而且图像中的变化除了火灾导致的烧伤疤痕以外,还涉及到自然因素导致的变化。从 OA 值和 Kappa 系数可以看出,实验指标在 3 组实验中整体上是逐渐降低的,但是笔者提出的 HCNN 模型都能达到最好的效果,充分证明了本文提出的网络结构的有效性。

图 8 ~ 图 10 分别为在 L5T1 vs. L5T2 实验、L5T1 vs. ALIT2 实验和 L5T1 vs. L8T2 实验中不同算法得到的变化检测结果,从直观上观察不同算法的效果(白色表示变化区域,黑色表示未变化区域)。

表 1 Bastrop 复杂火灾数据集的变化检测结果

算法	OA / %		
	L5T1 vs. L5T2	L5T1 vs. ALIT2	L5T1 vs. L8T2
MAD + CVA	87.233	79.991	77.075
KCCA + CVA	92.614	90.175	91.671
DCCA + CVA	95.556	94.676	93.943
DCCAE + CVA	96.598	95.713	94.308
EF	97.358	97.115	94.314
PS	98.382	98.136	94.526
HCNN	99.286	98.764	95.581

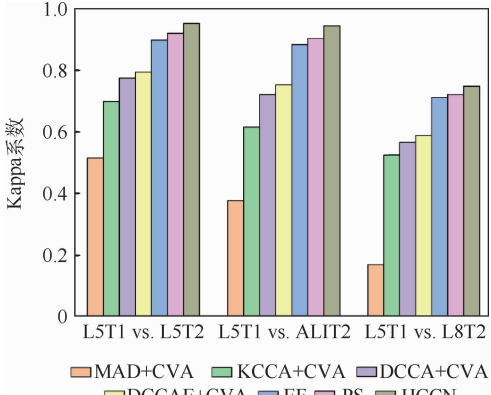


图 7 Bastrop 复杂火灾数据集的变化检测结果
数据分析(Kappa 系数)

Fig. 7 Change detection result data analysis of
Bastrop complex fire dataset (Kappa coefficient)

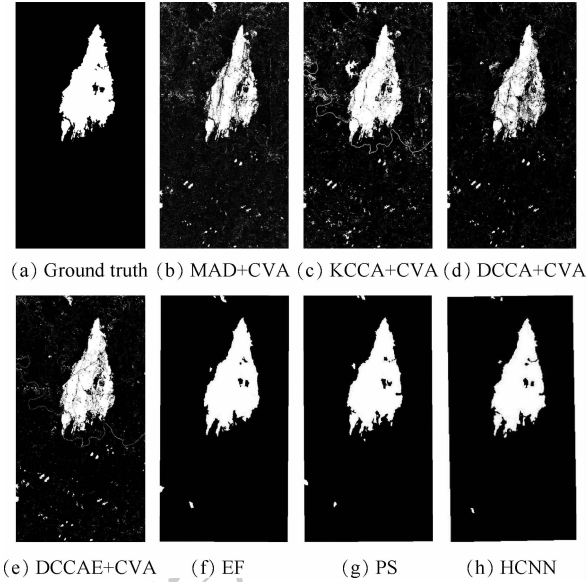


图 8 L5T1 vs. L5T2 实验中的变化检测结果

Fig. 8 Results of change detection in L5T1 vs. L5T2 experiment

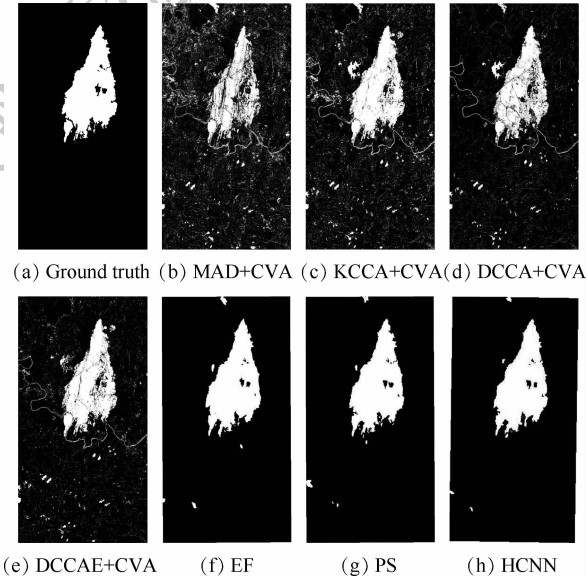


图 9 L5T1 vs. ALIT2 实验中的变化检测结果

Fig. 9 Results of change detection in L5T1 vs. ALIT2
experiment

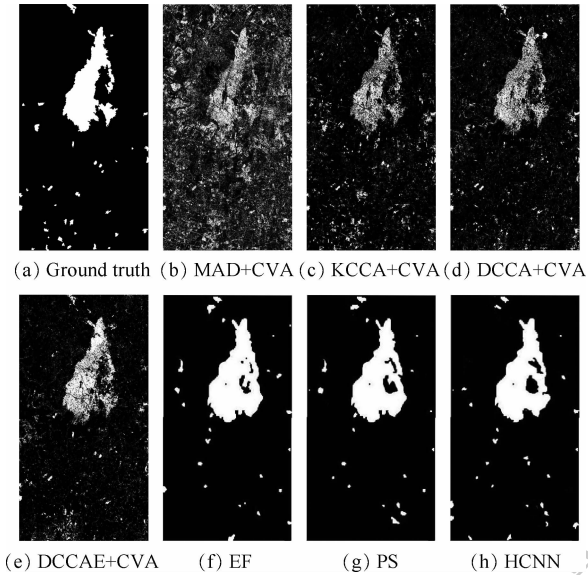


图 10 L5T1 vs. L8T2 实验中的变化检测结果
Fig. 10 Results of change detection in L5T1 vs. L8T2 experiment

从 3 组实验的结果可以看出,2 种传统算法 MAD + CVA 和 KCCA + CVA 在 3 组实验上的效果都较差,存在较多的虚检和漏检的像素点,并且虚检和漏检的情况也是越来越严重的,原因是:传统算法只能提取图像的低维特征,表征能力不强,无法准确地区分不同的变化,所以存在很多误检的情况;基于 CNN 的深度特征映射算法 DCCA + CVA 和 DCCAE + CVA 算法在前 2 组实验中效果一般,虚检和漏检的情况相对于传统的 MAD + CVA 和 KCCA + CVA 算法有了改善,然而在第 3 组实验中误检的情况仍然很严重,这是由于此类算法在特征映射时只考虑了图像的光谱特征,忽略了图像空间特征,在变化较难检测的第 3 组实验中没有达到较好的效果;基于 CNN 的分类算法 EF 和 PS 效果较好,通过提取图像块的空间特征,充分利用邻域信息,能够有效滤除噪声等干扰。

HCNN 具备了 EF 和 PS 网络充分利用空间特征的优点,同时融合深度和低级的空间-光谱-时间特征,有利于提升网络的判别能力,通过优化本节设计的损失函数,在特征空间中,未变化的图像对特征差异更小,变化的图像特征差异更大,网络具有更强的区分能力,从而在 3 组实验中都能检测出最正确的变化区域。

2) Onera 变化检测数据集上的实验

在 Onera 变化检测数据集上验证 HCNN 模型在同一传感器获得的遥感图像变化检测任务上也有不错的性能。

表 2 和表 3 分别展示了孪生网络 (Siamese network, S)^[13]、PS^[15]、EF^[24] 及本节提出的 HCNN

模型在香港图像对和孟买图像对上的实验结果。图 11 和图 12 分别展示了上述算法在香港图像对和孟买图像对上得到的变化检测结果(白色表示变化区域,黑色表示未变化区域)。可以看出,检测得到的变化地图与地面真实情况大致相同,HCNN 相比对图像空间特征进行利用的其他 3 种算法,融合深度和低级的空间-光谱-时间特征,同时本节设计的损失函数使得网络区分能力更强,可以看到,HCNN 获得的结果图相比于其他结果图而言,漏检的变化像素较少。

表 2 在香港图像对上使用不同算法获得的变化检测定量结果

Table 2 Quantitative results of change detection obtained by different algorithms on Hong Kong image pairs			
算法	OA/%	AA/%	Kappa 系数
S	96.542	68.511	0.424
PS	97.900	76.090	0.631
EF	97.134	74.943	0.598
HCNN	97.617	81.200	0.656

表 3 在孟买图像对上使用不同算法获得的变化检测定量结果

Table 3 Quantitative results of change detection obtained by different algorithms on Mumbai image pairs			
算法	OA/%	AA/%	Kappa 系数
S	95.928	69.293	0.457
PS	96.325	80.217	0.635
EF	95.449	80.599	0.628
HCNN	97.293	82.583	0.673

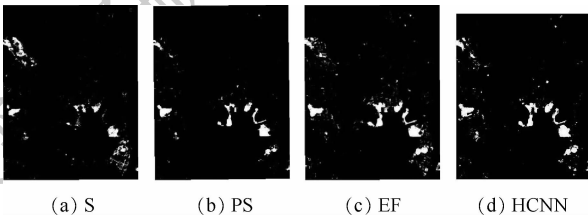


图 11 在香港图像对上使用不同算法获得的变化检测结果
Fig. 11 Change detection results on Hong Kong image pair obtained by different algorithms

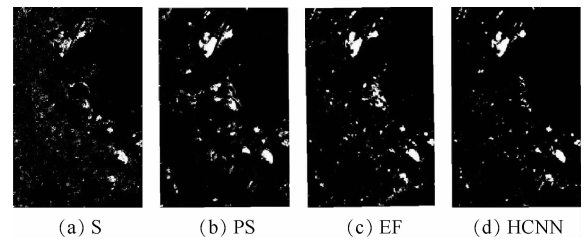


图 12 在孟买图像对上使用不同算法获得的变化检测结果
Fig. 12 Change detection results on Mumbai image pair obtained by different algorithms

5.5 3D 卷积的影响

为了分析早期融合网络中使用 2D 卷积和 3D 卷积对检测精度的影响,在 L5T1 vs. L8T2 实验中,对 EF 网络与 HCNN 网络分别使用 2D 卷积和 3D 卷积进行运算,实验结果指标如表 4 所示。可以看出,在 EF 网络和 HCNN 网络中使用 3D 卷积操作,检测精度更高,这是因为 3D 卷积能够提取图像块的光谱特征的局部信号变化,为后续的分类提供重要的判别信息。

表 4 L5T1 vs. L8T2 实验中 EF 网络与 HCNN 网络使用 2D 卷积和 3D 卷积对检测精度的影响

Table 4 Effect of EF and HCNN networks with 2D convolution and 3D convolution on detection accuracy in L5T1 vs. L8T2 experiment

算法	OA/%	Kappa 系数
EF(2D 卷积)	94.314	0.711
EF(3D 卷积)	94.938	0.723
HCNN(2D 卷积)	95.258	0.731
HCNN(3D 卷积)	95.581	0.759

5.6 对比损失的影响

为了验证对比损失函数对网络训练的良性作用,对比了 L5T1 vs. ALIT2 实验中在 PS 网络与 HCNN 网络训练中是否添加额外的对比损失函数的变化检测结果指标,如表 5 所示。总体而言,在 PS 网络和 HCNN 网络中添加了对比损失后,OA 和 Kappa 系数值有所提高。这是因为:对比损失函数能够使得在特征空间中,未变化的图像对特征差异更小,变化的图像特征差异更大,有利于提升网络的区分能力。

此外,对比损失函数会增加网络的收敛速度。如图 13 所示,可以发现,在使用对比损失函数后,无论是 PS 网络还是 HCNN 网络,对比损失函数都更快地下降,网络的收敛速度有了较大的提升。

表 5 L5T1 vs. ALIT2 实验中 PS 网络与 HCNN 网络是否添加对比损失函数对检测精度的影响

Table 5 Effect of PS and HCNN networks with and without contrast loss function on detection accuracy in L5T1 vs. ALIT2 experiment

算法	OA/%	Kappa 系数
PS(未添加对比损失函数)	98.136	0.903
HCNN(未添加对比损失函数)	98.66	0.932
PS(添加对比损失函数)	98.356	0.917
HCNN(添加对比损失函数)	98.764	0.943

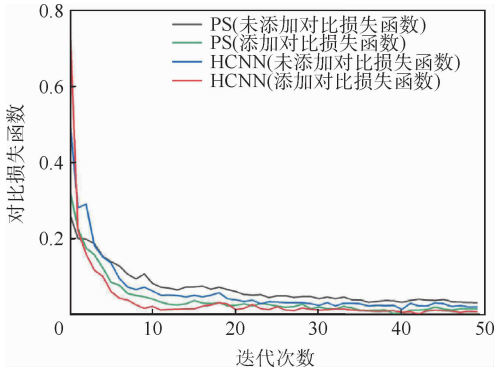


图 13 PS 网络与 HCNN 网络是否添加对比损失函数的二元交叉熵损失随着网络训练的变化曲线

Fig. 13 Binary cross entropy loss curves of PS and HCNN networks with and without contrast loss function binary in network training

6 结 论

- 1) 本文算法可实现较为优异的异源遥感图像变化检测性能,如在 Bastrop 复杂火灾数据集的实验中,所提出的 HCNN 算法的 OA 值在 3 组实验中分别达到 99.286%、98.764%、95.581%,均高于其他的检测算法。
- 2) 本文算法在同传感器获取的遥感图像变化检测任务中也有非常好的效果,Onera 变化检测数据集上的实验结果证明了这一点。
- 3) 3D 卷积的使用可以更加有效地提取图像块光谱特征的局部信号变化,为后续的特征融合提供了重要的判别信息。
- 4) 设计的对比损失函数可以使得特征空间中未变化的图像对特征差异更小,变化的图像特征差异更大,提升了网络的区分能力和收敛速度。

参考文献 (References)

[1] ALBERGA V. Similarity measures of remotely sensed multi-sensor images for change detection applications[J]. Remote Sensing, 2009, 1 (3) : 122-143.

[2] FERNANDEZ-PRIETO D, MARCONCINI M. A novel partially supervised approach to targeted change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2011, 49 (12) : 5016-5038.

[3] LAI P L, FYFE C. Kernel and nonlinear canonical correlation analysis[J]. International Journal of Neural Systems, 2000, 10 (5) : 365-377.

[4] VOLPI M, DE MORSIER F, CAMPS-VALLS G, et al. Multi-sensor change detection based on nonlinear canonical correlations[C] // 2013 IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE Press, 2013: 1944-1947.

[5] ZHU Z, WOODCOCK C E. Object-based cloud and cloud shadow detection in Landsat imagery[J]. Remote Sensing of Envi-

- ronment,2012,118:83-94.
- [6] JIAN M, LAM K, DONG J, et al. Visual-patch-attention-aware saliency detection[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2015, 45(8):1575-1586.
- [7] JIAN M, QI Q, DONG J, et al. Integrating QDWD with pattern distinctness and local contrast for underwater saliency detection[J]. Journal of Visual Communication and Image Representation, 2018, 53:31-41.
- [8] JIAN M, ZHANG W, YU H, et al. Saliency detection based on directional patches extraction and principal local color contrast[J]. Journal of Visual Communication and Image Representation, 2018, 57:1-11.
- [9] MERCIER G, MOSER G, SERPICO S B. Conditional copulas for change detection in heterogeneous remote sensing images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2008, 46(5):1428-1441.
- [10] GAO F, LIU X, DONG J, et al. Change detection in SAR images based on deep Semi-NMF and SVD networks[J]. Remote Sensing, 2017, 9(5):435.
- [11] MOU L, BRUZZONE L, ZHU X X. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2019, 57(2):924-935.
- [12] ZHANG P, GONG M, SU L, et al. Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2016, 116:24-41.
- [13] ZHANG Z, VOSSelman G, GERKE M, et al. Change detection between multimodal remote sensing data using Siamese CNN [EB/OL]. (2018-07-25) [2020-08-01]. <https://arxiv.org/abs/1807.09562>.
- [14] LIU J, GONG M, QIN K, et al. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images[J]. IEEE Transactions on Neural Networks and Learning Systems, 2016, 29(3):545-559.
- [15] HUGHES L H, SCHMITT M, MOU L, et al. Identifying corresponding patches in SAR and optical images with a pseudo-Siamese CNN[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 15(5):784-788.
- [16] NIU X, GONG M, ZHAN T, et al. A conditional adversarial network for change detection in heterogeneous images[J]. IEEE Geoscience and Remote Sensing Letters, 2018, 16(1):45-49.
- [17] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.
- [18] LI Y, ZHANG H, SHEN Q. Spectral-spatial classification of hyperspectral imagery with 3D convolutional neural network[J]. Remote Sensing, 2017, 9(1):67.
- [19] CHOPRA S, HADSELL R, LECUN Y. Learning a similarity metric discriminatively, with application to face verification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2005, 1:539-546.
- [20] DAUDT R C, LE SAUX B, BOULCH A, et al. Urban change detection for multispectral earth observation using convolutional neural networks[C]//2018 IEEE International Geoscience and Remote Sensing Symposium. Piscataway: IEEE Press, 2018:2115-2118.
- [21] VOLPI M, CAMPS-VALLS G, TUIA D. Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2015, 107:50-63.
- [22] ANDREW G, ARORA R, BILMES J, et al. Deep canonical correlation analysis[C]//International Conference on International Conference on Machine Learning, 2013, 28:1247-1255.
- [23] WANG W, ARORA R, LIVESCU K, et al. On deep multi-view representation learning[C]//International Conference on Machine Learning, 2015:1083-1092.
- [24] ZAGORUYKO S, KOMODAKIS N. Learning to compare image patches via convolutional neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015:4353-4361.

作者简介:

周圆,女,博士,教授,博士生导师。主要研究方向:机器学习、计算机视觉、3D图像/视频编码与传输、无线传感器网络、网络视频通信等。

Heterogeneous remote sensing image change detection
based on hybrid network

ZHOU Yuan^{1,*}, LI Xiangrui², YANG Jing¹

(1. School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China;
2. Tianjin International Engineering Institute, Tianjin University, Tianjin 300072, China)

Abstract: In order to more quickly and accurately perform the change detection task of heterogeneous remote sensing images, this paper presents a heterogeneous remote sensing image change detection algorithm based on a hybrid network. The algorithm uses a pseudo-siamese network to extract change features between the heterogeneous image blocks in spatial dimension, and uses an early fusion network to extract change features between the heterogeneous image blocks in spectral dimension. The features extracted from the two networks are fused and the fused features are input to the sigmoid layer for binary classification to determine whether the feature has changed. In addition, the contrast loss function is added to the pseudo-siamese network, so that in the features space, the spatial features of the unchanged image pair are closer, and the spatial features of the changed image pair are farther away, which is conducive to improving the network's distinction ability and convergence speed.

Keywords: change detection; heterogeneous; pseudo-siamese network; early fusion network; contrast loss function

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

DOI: 10.13700/j.bh.1001-5965.2020.0456

融合深度特征的输电线路金具缺陷因果分类方法

赵振兵^{1,*}, 张薇¹, 戚银城¹, 翟永杰², 赵文清²

¹华北电力大学 电气与工程学院, 保定 071003; ²华北电力大学 控制与计算机工程学院, 保定 071003)

摘

要: 针对输电线路金具缺陷样本不足和缺陷目标形态多样化, 仅仅利用深度学习

模型导致金具缺陷分类准确率较低的问题, 提出了一种结合深度网络和逻辑回归模型的因果分类方法。首先, 通过样本扩充方法获得数量丰富化和角度多样化的数据集; 然后, 基于微调后的 VGG16 模型提取深度特征并进行特征处理, 以构建符合因果关系学习的输入特征集; 最后, 通过全局混杂平衡进行金具缺陷特征与标签之间的因果关系学习, 构建符合金具特点的因果逻辑回归模型, 完成金具缺陷分类。为了证明所提方法的有效性, 利用无人机实际采集的 4 类金具缺陷图片分别进行了实验, 所使用的训练样本和测试样本数量较原始数据集提升了 5 倍左右。实验结果表明: 所提方法可以实现对输电线路金具缺陷的精准分类, 其中, 防震锤相交和变形分类准确率分别达到了 0.929 9 和 0.911 8, 屏蔽环锈蚀和均压环损坏分类准确率分别达到了 0.956 7 和 0.966 9。

关键词: 输电线路金具缺陷; 因果关系学习; 深度特征; 逻辑回归模型; VGG

中图分类号: TM726; TH165+.3

文献标志码: A

文章编号: 1001-5965(2021)03-0461-08

金具作为输电线路十分重要且分布广泛的金属附件, 主要起到支持、固定、防护和接续线路各个部件及维持整个线路稳定的作用^[1]。由于金具长期处于复杂恶劣的野外环境中, 极易发生多种故障, 如变形、破损、锈蚀和裂纹等, 对输电线路上的金具进行定期检修可以最大限度地减弱电路损坏事故的发生, 保障整个电力系统的安全正常运行。传统的金具巡检主要通过人工检修, 耗时费力的同时检修效率较低。目前, 电力系统^[2-3]已经广泛采用安全系数较高的直升机和无人机进行输电线路金具等重要部件的巡检工作, 但由于拍摄角度的多变性和金具所处背景的复杂性, 必须利用目前鲁棒性较强的图像处理算法, 并结合金具先验知识处理其缺陷分类问题, 通过对获得的金具图像的处理可以判断出金具是否

正常, 以及产生何种缺陷隐患, 这能够很大程度地提高输电线路自动化巡检的效率。

目前, 国内外众多学者对电力系统中输电线路金具缺陷进行了成因分析。文献[4]分析了输电线路电力金具腐蚀和磨损的原因, 并提出相应的建议和改进方法。文献[5]采用图像融合和处理等技术对金具图像进行缺陷状态评估, 为电力金具缺陷分类提供了有力的理论支撑。文献[6]通过分析多种线夹缺陷原因, 提出了针对其缺陷的修补方法和运维建议, 以保证线路的安全运行。以金具缺陷的成因分析作为金具缺陷分类研究的理论依据, 其研究可大致分成 2 类, 即基于传统的分类算法研究^[7-10]和基于深度学习的分类算法研究^[11-12], 但 2 类算法不能很好地适应于缺陷样本数量不足和缺陷目标形态多样化的金具缺陷分

收稿日期: 2020-08-24; 录用日期: 2020-08-28; 网络出版时间: 2020-09-11 16:40

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200911.1032.002.html

基金项目: 国家自然科学基金(61871182, 61773160); 北京市自然科学基金(4192055); 河北省自然科学基金(F2020502009); 中央高校基本科研业务费专项资金(2018MS095, 2020YJ006); 模式识别国家重点实验室开放课题基金(201900051)

* 通信作者: E-mail: zhaozhenbing@ncepu.edu.cn

引用格式: 赵振兵, 张薇, 戚银城, 等. 融合深度特征的输电线路金具缺陷因果分类方法[J]. 北京航空航天大学学报, 2021, 47(3): 461-468. ZHAO Z B, ZHANG W, QI Y C, et al. Causal classification method of transmission lines fitting defect combined with deep features[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 461-468 (in Chinese).

类,具有一定的局限性。

实现金具缺陷准确分类是实现输电线路自动化巡检的基础,输电线路金具缺陷分类由于其样本数量和缺陷状态的特殊性,仅仅使用深度学习模型进行缺陷分类有很大的局限性。针对金具缺陷数据集的特殊性,本文设计了一种融合深度特征与标签因果关系学习的金具缺陷因果分类方法,采用 VGG16 模型^[13]进行金具缺陷深度特征提取,通过全局混杂平衡进行其特征与标签的因果关系学习,实现了输电线路巡检图像中金具缺陷图片的精准分类。

1 相关工作

目前,已有的输电线路金具缺陷分类应用研究取得一定进展,主要集中在防震锤、线夹和均压环等部件,其中大部分方法以传统分类方法为主。文献[7]提出一种利用图像匹配和形态学运算方法对间隔棒框架断裂故障进行了识别与分类。文献[8]通过提取防震锤和线夹等边缘图像中图形基元并施加结构约束,有效实现了障碍物的识别。文献[9]通过结合形态学处理和 RGB 彩色模型,实现了防震锤锈蚀缺陷识别。文献[10]提出了利用基于类 Haar 特征和级联 Ada Boost 分类器对防震锤的目标基元进行缺陷判别的方法,有效降低了防震锤缺陷误判率。该类分类方法可扩展性差,对背景复杂、目标形态多变的金具缺陷分类有很大的局限性。基于深度模型的金具缺陷分类方法较少,文献[11]提出了通过设置锈蚀特征色值分离出目标和背景并结合深度模型进行输电线路锈蚀识别的方法。文献[12]通过样本扩充和调整卷积神经网络(Convolutional Neural Networks, CNN)中卷积核大小,有效提高了防震锤、均压环缺陷识别准确率。该类方法主要适用于样本充足且目标形态单一的金具分类研究。针对缺陷样本过少且缺陷目标形态多样化的金具缺陷数据集,直接利用深度学习模型难以充分学习到金具缺陷特征导致其分类效果不佳。

在后深度学习时代,仅仅利用深度学习模型进行金具缺陷分类已不能满足需求,因为深度学习模型主要依赖特征分布的统计特性,进而学习输入与输出之间复杂的相关性表示^[14],导致模型并未充分学习到金具缺陷特征。Judea 和 Ghosh^[15]也提出,深度学习主要对数据进行曲线拟合训练,研究数据的相关关系而非具有稳定性的因果关系。Bottou 团队提出^[16],机器学习非常适合学习数据集中隐藏的因果关系,并且基于因

果关系学习建模能够带来更稳定、泛化性能更好的模型。借鉴 Bottou 团队的思想,通过对特征与标签进行因果关系学习建模,并结合机器学习中经典的逻辑回归分类模型,可以帮助模型充分学习到金具缺陷特征,从而提高其缺陷分类准确率。

2 金具缺陷因果分类方法

本文研究的主要任务是在输电线路航拍图像中对正常金具和该类金具缺陷 2 种状态进行分类,由于已有数据集中仅包含防震锤、均压环和屏蔽环 3 类金具的 4 类缺陷样本,分别选择这 3 类金具为研究对象。基于单类金具缺陷数据集,选取速度和精度均衡的 VGG16 模型提取该类金具不同状态的深度特征,并基于因果逻辑回归模型进行金具缺陷分类。本文方法的框架如图 1 所示。首先,采用简单且性能较好的水平和旋转操作进行样本扩充;其次,将扩充后的样本输入 VGG16 模型进行微调训练,在 FC7 层提取金具样本的深度特征向量,并进行主成分分析法(Principal Component Analysis, PCA)降维;最后,基于全局混杂平衡原理进行特征与标签的因果关系学习,并结合因果逻辑回归模型完成对输电线路金具缺陷和正常样本更为准确的分类。

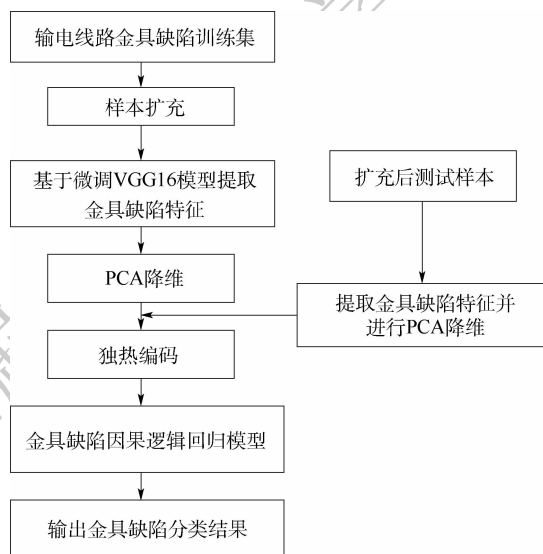


图1 本文方法总体流程

Fig. 1 Overall flowchart of proposed method

2.1 金具缺陷样本扩充

输电线路金具缺陷数据集具有所处背景复杂、拍摄视角多、样本数量失衡和类内差异大等特点。图 2 展示了本文研究的 3 类主要金具及其存在的主要缺陷状态。可以看出,屏蔽环目标形态单一,但受无人机拍摄时光照和天气的影响,屏蔽环锈蚀样本成像也有所不同,均压环正常和损坏



图 2 三类主要金具及其缺陷样本示例

Fig. 2 Examples of three types of main fitting and their defect samples

样本形态多样化,防震锤目标形态差异较小。此外,通过对金具缺陷数据集的样本数量进行统计,发现 3 种金具正常和缺陷状态下均有不同程度的数量失衡。因此,输电线路金具缺陷数据集的一系列特点决定了实现金具正常和缺陷样本准确分类具有一定的难度。

使用适当的样本扩充方法可以获得丰富多样的训练样本,有效防止模型训练过拟合^[17],一定程度上可以降低样本角度、形态、样本不平衡对金具缺陷分类的影响。本文结合输电线路金具缺陷数据集特点,使用水平翻转和 15°、45°和 75°三个

角度对金具训练样本进行旋转等,以获得角度真实多样化和样本数量丰富化的训练集,并将扩充后的正常和缺陷样本作为训练样本,输入到微调后的 VGG16 模型进行训练并提取深度特征向量。

由于金具测试集中缺陷样本过少,直接利用原始缺陷样本进行分类器测试,会严重影响最终分类器的精度,采用与训练样本相同的扩充方法对测试集中的缺陷样本进行扩充,并基于扩充后的测试样本进行深度特征提取和处理。

2.2 深度特征提取

输电线路金具缺陷分类相较于一般的分类任务难度更大,主要体现在 2 个方面:①同一类别的金具由于本身制造工艺、规格不同导致形态多样化,不同形态的同类金具的缺陷也具有多样性;②同一类别的金具正常和缺陷形态差异较小。因此,要想实现对金具缺陷样本的准确分类,必须对所提取的金具缺陷特征有很高的要求。文献[18]提出利用预训练模型进行特征提取,获得的特征通常具有较好的表征能力。VGG16 模型作为性能稳定且泛化性很强的卷积神经网络,如今被大量地应用在图像的特征提取问题中。因此,本文选择 VGG16 模型作为特征提取器,并通过微调训练符合金具缺陷数据集特点的模型,进而提取出更具表征能力的深度特征。本文选择 FC7 层输出的高维度特征作为所提取的金具深度特征向量,VGG16 模型结构如图 3 所示。

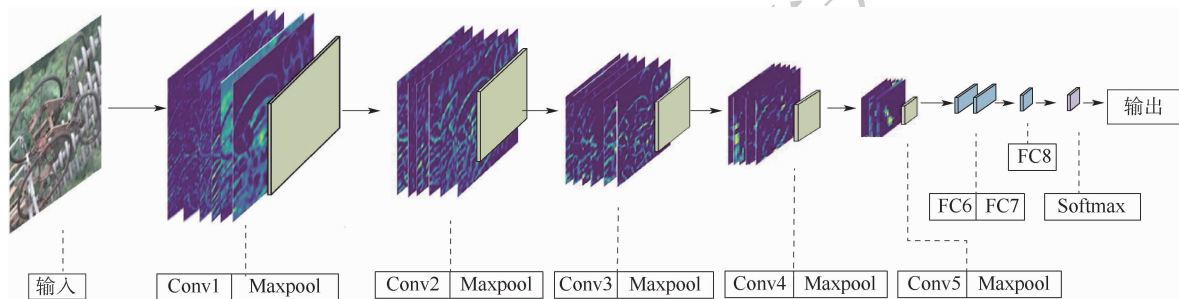


图 3 VGG16 模型结构

Fig. 3 Structure of VGG16 model

当深度模型提取的特征维度过高时,往往会造成维度灾难并影响后面的分类器性能^[19]。为了降低数据维度过高带来的复杂度,本文选择较为经典的 PCA^[20]对所提取的 4 096 维特征进行降维处理。多次实验表明,当特征维度降为 64 维时,金具缺陷分类器具有较好的性能,并且线性变换后的特征能有效表达原来特征,最大化保持金具的内在特征信息。图 4 为金具缺陷数据集深度特征提取处理图。

2.3 结合因果关系学习的金具缺陷分类方法

传统的因果逻辑回归模型^[21]是机器学习经典的分类模型,但并非严格意义上的因果模型,其回归系数仅代表特征和标签之间的高度相关性而非因果性。本文为了实现输入特征与输出标签之间的因果关系学习作为指导信息进行金具缺陷分类,受文献[22]中用于消除不可知数据选择偏差的因果学习方法的启发,首先,使用独热编码(One-hot Encoding)^[23]将金具缺陷特征矩阵转化

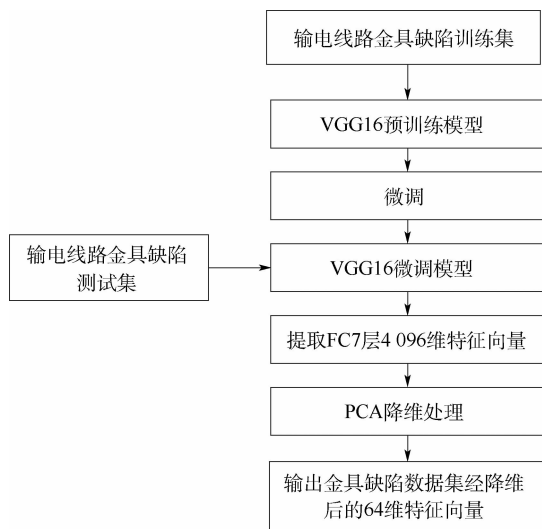


图4 金具缺陷数据集深度特征提取处理图

Fig. 4 Extraction and processing of deep features of fitting defect dataset

成二值化形式,其特征值仅包含0和1两种形式,通过二值化特征识别处理特征和混杂特征,进而将训练样本分成处理组和对照组,并利用全局混杂平衡方法完成金具缺陷特征与标签之间的因果关系学习,从而构建因果正则化项;其次,根据输电线路金具缺陷数据集的样本不平衡情况,通过引入结合金具缺陷数据集特点类别权重并结合原始样本权重,提高金具缺陷的类别权重值进而提高其分类精度;最后,将因果正则化项与因果逻辑回归模型相结合,通过联合优化权重系数 $\mathbf{W}$ 和回归分布系数 β 训练金具缺陷分类模型,损失函数为

$$L = L_w + \lambda_1 L_B + \lambda_2 \left(\sum_{k=1}^m W_k - 1 \right)^2 + \lambda_3 \|\mathbf{W}\|_2^2 + \lambda_4 \|\beta\|_1 \quad (1)$$

式中: L_w 为因果逻辑回归模型的加权损失,具体见式(3); L_B 为用于计算样本中 n 维特征的全局混杂平衡损失之和,具体见式(4); λ_1 为因果逻辑回归模型的因果正则化项系数; λ_2 为权重集 $\mathbf{W}$ 的归一化系数,避免模型训练过程中权重为0; λ_3 为权重 $\mathbf{W}$ 的正则化项参数,用于减小权重方差以保证训练的稳定性; λ_4 为因果分布 β 的正则化项参数,用于避免模型训练过拟合。

由于因果逻辑回归模型通常仅仅考虑样本权重 W_i ,对于样本数量不平衡的金具缺陷数据集而言,通过引入与标签相关的类别权重 W_c ,提高金具缺陷类别的权重,相对降低金具正常类别的权重,可以在缺陷样本数量有限的情况下,提高分类模型对金具缺陷样本的准确率。本文根据训练集中正常和缺陷样本数量情况,通过式(2)计算出其类别权重,并与原始样本权重相乘作为新的权

重,用于优化训练金具缺陷分类模型。

$$W_c = \frac{n_samples}{n_classes \cdot n_count(y)} \quad (2)$$

式中: $n_samples$ 为训练样本数量; $n_classes$ 为训练样本类别数量; $n_count(y)$ 为不同类别样本出现次数。

因果逻辑回归模型的加权损失函数如式(3)所示,改进后模型通过学习全局样本加权降低模型的复杂度,并且回归系数不再仅仅代表特征与标签的高度相关性,而是对标签具有预测能力和因果含义。

$$L_w = \sum_{i=1}^m W_i \ln(1 + \exp((1 - 2Y_i)(x_i \beta))) \quad (3)$$

式中: m 为训练集样本数量; x_i 和 Y_i 分别为第 i 个样本对应的特征和标签; β 为该样本对应的因果回归系数;整体为因果逻辑回归模型的加权损失函数。

因果逻辑回归损失函数在原有损失函数基础上增加了利用全局混杂平衡方法进行因果关系学习的正则化项,其主要思想为:将样本中的 n 维特征依次设置为处理特征进行混杂平衡^[24]处理,完成 n 维特征与对应标签的因果关系学习,其平衡损失函数如式(4)所示。具体地,混杂平衡处理方法是:将样本中第 j 维特征设置为处理特征,其余为混杂特征,训练集中出现该处理特征的样本集为处理组,未出现处理特征的样本集为对照组,再通过优化处理组和对照组平均值的二阶矩调整训练权重完成该特征与标签的因果关系学习。

$$L_B = \sum_{j=1}^n \left\| \frac{\mathbf{X}_{-j}^T (\mathbf{W} \odot \mathbf{I}_j)}{\mathbf{W}^T \cdot \mathbf{I}_j} - \frac{\mathbf{X}_{-j}^T (\mathbf{W} \odot (1 - \mathbf{I}_j))}{\mathbf{W}^T \cdot (1 - \mathbf{I}_j)} \right\|_2^2 \quad (4)$$

式中: n 为训练集中单个样本的特征维度;训练集整体为 $m \times n$ 维的特征矩阵; $\mathbf{W}$ 为训练集样本的权重集; $\mathbf{X}_{-j}$ 为输入特征矩阵 $\mathbf{X}$ 中第 j 列特征设为0时的混杂特征矩阵;矩阵 $\mathbf{I}$ 为样本特征是否接受处理的指标矩阵;整体为 n 维特征进行混杂平衡处理的损失值。

基于因果逻辑回归模型的金具缺陷分类方法流程如下:

输入: 金具训练集特征矩阵 $\mathbf{X}$ 、对应标签 $\mathbf{Y}$ 及基础参数。

输出: 金具缺陷样本分类准确率和召回率等。

步骤1 输入金具训练集特征矩阵 $\mathbf{X}$,通过 $\mathbf{X}$ 计算预测指标矩阵 $\mathbf{I}$ 。

步骤2 初始化因果系数和样本权重 W_i ,计算类别权重 W_c 和新权重 $\mathbf{W}$,计算优化函数的初

始值。

步骤 3 迭代次数依次增加,使用迭代优化方法^[25]进行损失函数优化,直至达到最优状态,得到权重 W 和因果系数 β 。

步骤 4 输入金具测试集特征矩阵,利用训练好的因果模型进行测试。

步骤 5 输出金具测试集缺陷分类准确率和召回率等。

3 实验结果与分析

考虑到模型的分类效果和建库的可操作性,通过搜集基于人工经验标注的 3 437 张输电线路航拍图像,经过裁剪等操作获取因果分类模型所需要的单类金具缺陷数据集,构建的数据集主要包括 4 类,即防震锤变形数据集、防震锤相交数据集、屏蔽环锈蚀数据集和均压环损坏数据集。其中,每个小型缺陷数据集仅包含正常和缺陷样本。特别地,防震锤相交数据集包含其正常样本和相交样本,防震锤变形数据集包含其正常和变形样本。由于已有的航拍数据集金具缺陷样本数量较少且图片质量较差,为了提高模型的泛化能力,将单类金具中正常和缺陷样本均进行随机打乱处理,并按照 8:2 的比例设置训练集和测试集,测试集中绝不包含训练样本。

此外,为了证明本文方法泛化能力强,选取经典的 Cifar-10 分类数据集中猫、狗 2 类图片构建子集,作为实验所需小型公共数据集,由于 Cifar-10 数据集图片大小为 32×32 ,部分图片清晰度低且质量较差,为了与金具缺陷数据集样本数量大体保持一致,2 类样本数量分别选取 1 000 张,同样按照 8:2 的比例设置训练集和测试集。

本文实验的硬件环境为 Linux Ubuntu16. 04 和具有 8 GB 内存的 NVIDIA GeForce TITAN Xp,本文方法基于 MATLAB 平台实现,其他对比方法使用 Tensorflow 平台。经过模型微调,实验参数设置如下:批处理大小设置为 24,学习率设置为 10^{-3} ,迭代次数设置为 3 000,正则化参数 $\lambda_1 = 0.1, \lambda_2 = \lambda_3 = 1, \lambda_4 = 0.001$ 。为了验证本文方法的有效性,使用扩充后的数据集进行了金具缺陷分类对比实验。

3.1 不同分类器下金具缺陷分类实验结果及分析

为了验证本文方法的有效性,在相同的网络优化参数设置的基础上,选择的对比方案如下:①方案 1。采用基于图像像素提取的因果逻辑回归模型进行金具缺陷分类实验。②方案 2。采用基于深度卷积特征提取的 VGG16 + Softmax 模型

进行金具缺陷分类实验。③方案 3。采用本文方法进行金具缺陷分类实验。

考虑到金具缺陷样本不平衡对分类器的影响,采用了常用的二分类模型评价指标,即准确率、召回率和 F1 分数。其中,准确率代表分类器预测正确的样本数量占总体样本数量的比例,召回率代表分类器预测的正类中实际正实例占有正实例的比例,F1 分数是统计学中用来衡量二分类模型精确度的指标,代表模型精确率和召回率的调和平均数。

表 1~表 4 分别展示了防震锤变形、防震锤相交、屏蔽环锈蚀和均压环损坏 4 类金具缺陷分类的实验结果。结果表明,基于 VGG16 深度模型比传统的因果逻辑回归模型进行金具缺陷分类的准确率提高了 5%~8% 左右,说明逐层级深入提取图像语义特征比传统的像素提取方式更能学习到金具正常和缺陷图像的语义特征,从而提高了模型的分类准确率。

表 1 防震锤相交分类实验结果

Table 1 Classification experimental results of shockproof hammer intersection

方案	准确率	召回率	F1 分数
方案 1	0.803 7	0.763 2	0.805 6
方案 2	0.887 8	0.807 1	0.884 6
方案 3	0.929 9	0.894 7	0.891 4

表 2 防震锤变形分类实验结果

Table 2 Classification experimental results of shockproof hammer deformation

方案	准确率	召回率	F1 分数
方案 1	0.808 8	0.388 9	0.518 5
方案 2	0.860 3	0.472 2	0.641 5
方案 3	0.911 8	0.666 7	0.800 1

表 3 屏蔽环锈蚀分类实验结果

Table 3 Classification experimental results of shielding ring corrosion

方案	准确率	召回率	F1 分数
方案 1	0.839 5	0.861 1	0.826 7
方案 2	0.882 7	0.847 2	0.865 2
方案 3	0.956 7	0.902 7	0.948 9

表 4 均压环损坏分类实验结果

Table 4 Classification experimental results of grading ring damage

方案	准确率	召回率	F1 分数
方案 1	0.842 5	0.834 2	0.843 3
方案 2	0.914 4	0.887 6	0.914 5
方案 3	0.966 9	0.956 5	0.972 3

本文提出的结合金具正常、缺陷深度卷积特征和标签的因果关系学习的分类方法,相较于传统的因果逻辑回归模型的分类方法准确率提升了12%左右,较基于VGG16深度模型的分类方法提高了4%~7%左右。此外,本文方法分类的召回率均有不同程度的提升,这说明了依赖金具缺陷特征的因果性比依赖特征分布的统计特性更能得到稳定且泛化性能好的模型,并且可以以较高的精度进行金具缺陷样本的分类且其缺陷误判率较低,这对于输电线路智能化巡检有着十分重要的应用价值。

3.2 金具缺陷数据集样本扩充实验结果及分析

本文主要针对已有金具缺陷数据集进行金具缺陷分类研究,故实验中将金具缺陷样本设为正样本,正常样本设为负样本。通过分析金具缺陷数据集特点,发现原始数据集中金具缺陷和正常状态下样本比例失衡,且金具缺陷样本较少,其中,防震锤变形样本22张、防震锤相交95张、屏蔽环锈蚀61张和均压环损坏168张。本文分别对3类金具进行缺陷和正常样本比例的定量分析,其中,防震锤样本比例失衡较为严重,即防震锤相交、变形和正常样本比例分别为1:11.2、1:48.36,屏蔽环锈蚀和正常样本数量均较少且样本比例为1:1.61,屏蔽环锈蚀和正常样本比例为1:5.27。为了提高模型分类性能,对防震锤正常图片进行优化,对4种金具缺陷样本采用本文的样本扩充方法进行扩充。此外,由于屏蔽环正常样本较少,同样采取本文扩充方法对其进行处理,扩充前后的数据集样本比例如图5所示。

表5为金具缺陷数据集样本扩充前后分别使用本文方法进行金具缺陷分类的实验结果。结果表明,4类金具缺陷数据集采用扩充方法后的准确率均有不同程度的提升。其中,防震锤由于目标形态差异较小和缺陷样本数量过少,其缺陷分

表 5 样本扩充前后本文方法实验结果
Table 5 Experimental results by proposed method before and after sample expansion

金具缺陷类别	扩充前准确率	扩充前召回率	扩充后准确率	扩充后召回率
防震锤相交	0.906 7	0.555 6	0.929 9	0.894 7
防震锤变形	0.901 5	0.333 4	0.911 8	0.666 7
屏蔽环锈蚀	0.927 5	0.833 4	0.956 7	0.902 7
均压环损坏	0.933 0	0.647 1	0.966 9	0.956 5

类准确率提升相对较小,但召回率提升较高。屏蔽环由于目标尺度较大、缺陷形态单一,相较于其他金具缺陷,扩充后屏蔽环锈蚀分类召回率提升较少。均压环缺陷样本存在类内差异较大、拍摄视角多等特点,且均压环由于单条主筋松动或者脱落导致其与中间绝缘子串存在微小角度变化而非90°,直接使用原始数据集致使均压环损坏样本难以识别,采用本文扩充方法可以有效提高均压环损坏分类的召回率。因此,使用样本扩充可以有助于基于本文方法构建的分类器实现对金具缺陷样本的精准分类。

3.3 不同分类器下公共数据集实验结果及分析

在公共数据集上对本文方法进行了评估,并从Cifar-10数据集中选取猫、狗2类图片作为实验所用数据集,其中,标签为猫样本设为正样本,标签为狗样本设为负样本,分类器以识别猫类样本为主。表6为使用不同分类器对所构建数据集进行分类的实验结果,其中,方案1、方案2和方案3与3.1节方案一致。

表6中的第2列为分类器的准确率,即分类器识别出猫狗正确样本占总样本数量的比例,第3列为分类器对正样本(猫类)的召回率,第4列为分类器的F1分数。实验结果表明,由于样本质量不佳且清晰度较低,导致分类器对猫、狗样本的分类准确率均不是很高,但相较于其他2种方法,使用本文方法进行猫、狗数据集分类的准确率和召回率均有所提升,这说明使用本文方法构建的分类模型具有较强的泛化能力,对其他数据集同样适用。

表 6 不同分类器的实验结果
Table 6 Experimental results of different classifiers

方案	准确率	召回率	F1 分数
方案 1	0.552 5	0.515 0	0.571 9
方案 2	0.702 5	0.725 0	0.700 5
方案 3	0.762 5	0.711 1	0.780 4

4 结 论

1) 针对输电线路单类金具正常和缺陷样本

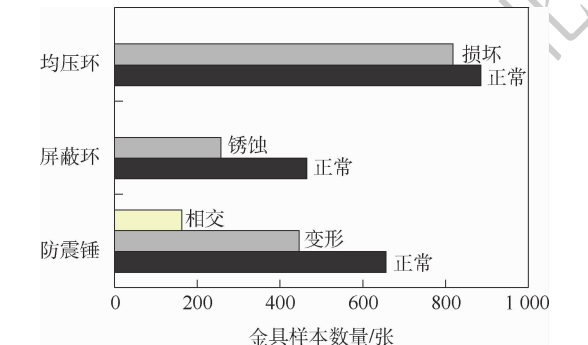


图 5 扩充后金具正常及其缺陷样本数量分布
Fig.5 Distribution of the number of fitting's normal and defect samples after expansion

比例失衡,采用水平翻转和角度旋转方法优化扩充缺陷样本,为后续金具缺陷分类工作奠定基础。实验结果表明,扩充后的金具缺陷分类准确率和召回率均有不同程度的提升。

2) 针对基于传统特征和深度特征的金具缺陷分类准确率较低等问题,采用 VGG16 进行特征提取并进行特征处理,通过因果逻辑回归模型识别因果特征有效提高了金具缺陷分类的准确率和召回率。

综上所述,本文方法展现出了在金具缺陷样本数量比例失衡严重、目标类内差异大和类间差异小情况下优良的分类性能。由于本文方法目前只能针对单类金具缺陷进行分类,具有一定的局限性,之后可以考虑进一步改进,同时实现多类金具缺陷分类,提升多类缺陷分类的准确率。

参考文献 (References)

- [1] 赵强. 输电线路金具理论与应用[M]. 北京: 中国电力出版社, 2013: 2-12.
ZHAO Q. Theory and application of transmission line fittings [M]. Beijing: China Electric Power Press, 2013: 2-12 (in Chinese).
- [2] DENG C, WANG S, HUANG Z, et al. Unmanned aerial vehicles for power line inspection: A cooperative way in platforms and communications[J]. Journal of Communications, 2014, 9(9): 687-692.
- [3] TONG W, YUAN J, LI B. Application of image processing in patrol inspection of overhead transmission line by helicopter [J]. Power System Technology, 2010, 34(12): 204-208.
- [4] 金哲, 尹洪, 吴启进. 典型 500 kV 输电线路地线金具腐蚀及磨损事件机理分析[J]. 电工技术, 2017(8): 66-69.
JIN Z, YIN H, WU Q J. Analysis on the corrosion and wear event mechanism of the ground wire of typical 500 kV transmission line[J]. Electric Engineering, 2017(8): 66-69 (in Chinese).
- [5] 陆旭, 罗汉武, 李文震, 等. 电力金具图像故障状态评估[J]. 红外技术, 2020, 42(7): 632-636.
LU X, LUO H W, LI W Z, et al. Evaluation of image failure state of power fittings[J]. Infrared Technology, 2020, 42(7): 632-636 (in Chinese).
- [6] 吴坤祥, 朱迪锋, 许杨勇, 等. ± 800 kV 特高压输电线路耐张线夹未压区鼓胀缺陷分析[J]. 浙江电力, 2017, 36(7): 11-13.
WU K X, ZHU D F, XU Y Y, et al. Analysis of bulging defects in uncompressed zone of tension clamp in ± 800 kV UHV transmission line[J]. Zhejiang Electric Power, 2017, 36(7): 11-13 (in Chinese).
- [7] 金立军, 胡娟, 闫书佳. 基于图像的高压输电线间隔棒故障诊断方法[J]. 高电压技术, 2013, 39(5): 1040-1045.
JIN L J, HU J, YAN S J. Image-based fault diagnosis method for spacers of high-voltage transmission lines[J]. High Voltage Engineering, 2013, 39(5): 1040-1045 (in Chinese).
- [8] 胡彩石, 吴功平, 曹珩, 等. 高压输电线路巡线机器人障碍物视觉检测识别研究[J]. 传感技术学报, 2008, 21(12): 2092-2096.
HU C S, WU G P, CAO H, et al. Research on visual inspection and recognition of obstacles for high-voltage transmission line patrol robot[J]. Chinese Journal of Sensors and Actuators, 2008, 21(12): 2092-2096 (in Chinese).
- [9] 宋伟, 左丹, 邓邦飞, 等. 高压输电线路防震锤锈蚀缺陷检测[J]. 仪器仪表学报, 2016, 37(S1): 113-117.
SONG W, ZUO D, DENG B F, et al. Corrosion defect detection of earthquake hammer for high voltage transmission line[J]. Chinese Journal of Scientific Instrument, 2016, 37(S1): 113-117 (in Chinese).
- [10] 付晶, 邵瑰玮, 吴亮, 等. 利用层次模型进行训练学习的线路设备缺陷检测方法[J]. 高电压技术, 2017, 43(1): 266-275.
FU J, SHAO G W, WU L, et al. Line equipment defect detection method using hierarchical model for training and learning[J]. High Voltage Engineering, 2017, 43(1): 266-275 (in Chinese).
- [11] 李辉, 钟平, 戴玉静, 等. 基于深度学习的输电线路锈蚀检测方法的研究[J]. 电子测量技术, 2018, 41(22): 54-59.
LI H, ZHONG P, DAI Y J, et al. Research on transmission line corrosion detection method based on deep learning[J]. Electronic Measurement Technology, 2018, 41(22): 54-59 (in Chinese).
- [12] 汤踊, 韩军, 魏文力, 等. 深度学习在输电线路中部件识别与缺陷检测的研究[J]. 电子测量技术, 2018, 41(6): 60-65.
TANG Y, HAN J, WEI W L, et al. Research on deep learning in component identification and defect detection in transmission lines[J]. Electronic Measurement Technology, 2018, 41(6): 60-65 (in Chinese).
- [13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 1-14.
- [14] MARCUS G. Deep learning: A critical appraisal[EB/OL]. (2018-01-02) [2020-08-01]. <http://export.arxiv.org/abs/1801.00631>.
- [15] GHOSH J, JUDEA P. Causality: Models, reasoning and inference[J]. International Statistical Review, 2011, 79(2): 289-290.
- [16] LOPEZPAZ D, NISHIHARA R, CHINTALA S, et al. Discovering causal signals in images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 58-66.
- [17] ZHONG J, SUN Y, YU Y, et al. Attribute-guided network for cross-modal zero-shot hashing[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 31(1): 321-330.
- [18] 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).

- [19] 杨博雄,杨雨琦.利用 PCA 进行深度学习图像特征提取后的降维研究[J].计算机系统应用,2019,28(1):279-283.
YANG B X,YANG Y Q. Research on dimensionality reduction after deep learning image feature extraction using PCA [J]. Computer Systems & Applications,2019,28(1):279-283 (in Chinese).
- [20] JOSE C. A fast on-line algorithm for PCA and its convergence characteristics [J]. IEEE Transactions on Neural Network, 2000,4(2):299-305.
- [21] 李航.统计学习方法[M].北京:清华大学出版社,2012:77-87.
LI H. Statistical learning method [M]. Beijing: Tsinghua University Press,2012:77-87 (in Chinese).
- [22] SHEN Z,CUI P,KUANG K,et al. Causally regularized learning with agnostic data selection bias [C] // ACM Multimedia. New York:ACM Press,2018:411-419.
- [23] 梁杰,陈嘉豪,张雪芹,等.基于独热编码和卷积神经网络的异常检测[J].清华大学学报(自然科学版),2019,59(7):523-529.
LIANG J,CHEN J H,ZHANG X Q,et al. One-hot encoding and convolutional neural network based anomaly detection [J]. Journal of Tsinghua University (Science and Technology), 2019,59(7):523-529 (in Chinese).
- [24] KUANG K,CUI P,LI B,et al. Estimating treatment effect in the wild via differentiated confounder balancing [C] // ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York:ACM Press,2017:265-274.
- [25] NEAL P,STEPHEN B. Proximal algorithms [J]. Foundations and Trends in Optimization,2014,1(3):127-239.

作者简介:

赵振兵 男,博士,副教授,硕士生导师。主要研究方向:电力视觉检测。

张薇 女,硕士研究生。主要研究方向:目标检测与图像处理。

戚银城 男,博士,教授,硕士生导师。主要研究方向:电力系统通信与信息处理。

Causal classification method of transmission lines fitting defect combined with deep features

ZHAO Zhenbing^{1,*}, ZHANG Wei¹, QI Yincheng¹, ZHAI Yongjie², ZHAO Wenqing²

(1. School of Electrical and Electronic Engineering, North China Electric Power University, Baoding 071003, China;

2. School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China)

Abstract: Aimed at the insufficient transmission lines fitting defect samples and diverse defect target shapes, a causal classification method combining deep network and logistic regression model is proposed to solve low defect classification accuracy when only using deep learning models. Firstly, rich and diverse datasets are obtained through the sample expansion method. Secondly, deep features are extracted based on the fine-tuned VGG16 model, and processed to construct an input feature set that conforms to causality learning. Finally, the causal relationship between fitting defect feature and label is learned through the global balance, and a causal logistic regression model is constructed to complete the classification of the fitting defects. Four types of fitting defect images collected by UAV are used respectively in the experiments to prove the effectiveness of the proposed method. The number of training and testing samples used is about 5 times higher than the original dataset. The experimental results show the proposed method can realize the accurate classification of the fitting defects, the classification accuracy of the shockproof hammer intersection and deformation reach 0.929 9 and 0.911 8 respectively, and the classification accuracy of the shielding ring corrosion and the grading ring damage reach 0.956 7 and 0.966 9 respectively.

Keywords: transmission line fitting defect; causality learning; deep features; logistic regression model; VGG

Received: 2020-08-24; Accepted: 2020-08-28; Published online: 2020-09-11 16:40

URL: kns.cnki.net/kcms/detail/11.2625.V.20200911.1032.002.html

Foundation items: National Natural Science Foundation of China (61871182,61773160); Beijing Natural Science Foundation (4192055); Natural Science Foundation of Hebei Province (F2020502009); the Fundamental Research Funds for the Central Universities (2018MS095,2020YJ006); Open Project Program of the National Laboratory of Pattern Recognition (201900051)

* Corresponding author. E-mail: zhaozhenbing@ncepu.edu.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0451

基于语义相关的多模态社交情感分析



胡慧君^{1,2,*}, 冯梦媛^{1,2}, 曹梦丽^{1,2}, 刘茂福^{1,2}

(1. 武汉科技大学 计算机科学与技术学院, 武汉 430065;

2. 智能信息处理与实时工业系统湖北省重点实验室, 武汉 430081)

摘 要: 社交平台允许用户采用多种信息模态发表意见与观点,多模态语义信息融合能够更有效地预测用户所表达的情感倾向。因此,多模态情感分析近年来受到了广泛关注。然而,多模态情感分析中视觉与文本存在的语义无关问题,导致情感分析效果不佳。针对这一问题,提出了基于语义相关的多模态社交情感分析(MSSA-SC)方法。采用图文语义相关性分类模型,对图文社交信息进行语义相关性识别,若图文语义相关,则对图文社交信息使用图文语义对齐多模态模型进行图文特征融合的情感分析;若图文语义无关,则仅对文本模态进行情感分析。在真实社交媒体数据集上进行了实验,由实验结果可知,所提方法能够有效降低图文语义无关情况对多模态社交媒体情感分析的影响。与此同时,所提方法的 Accuracy 和 Macro-F1 指标分别为 75.23% 和 70.18%,均高于基准模型。

关 键 词: 多模态; 社交媒体; 情感分析; 语义相关; 图文特征融合

中图分类号: TP37

文献标志码: A

文章编号: 1001-5965(2021)03-0469-09

随着社交网络的迅速发展,人们逐渐倾向于使用新浪微博、Twitter 等社交平台发表自己的观点与见解。如能对这些社交信息中蕴含的情感加以收集和分析,无疑能够对网络舆情进行分析^[1],了解大众对于热点事件的看法,有助于预防各种突发事件和危机。

以往的情感分析研究大多局限于单一的文本或视觉模态,然而,对于社交媒体进行情感分析来说,往往单一模态所表达的情感是不完整且具有局限性。因此,越来越多的研究者在对于社交媒体进行情感分析研究时,开始着眼于多模态^[2-4]的方法。与此同时,现有的多模态社交媒体情感分析研究仍存在以下问题:①不同模态特征融合的方法相对简单,造成多模态情感特征无法有效地融合;②社交媒体中用户发布的图文社交信息

可能存在图文语义无关的情况,现有方法未对图文语义相关性进行筛选,若对图文语义无关的图文社交信息直接进行图文特征融合的情感分析,则会对情感分类的结果造成影响。

因此,对于解决目前多模态社交媒体情感分析研究存在的问题,本文主要有 3 个方面的贡献:

1) 将图像纳入 RoBERTa^[5] 模型框架中,构建出语义对齐多模态 RoBERTa (Semantic Alignment Multimodal RoBERTa, SAMRoBERTa) 模型,该模型设计有图文语义对齐编码层,可以使图文信息相匹配的图像区域获得更多的权重信息,再把语义对齐后的图像特征与文本特征输入多模态编码层进行融合计算,从而使得图文特征能够有效地融合。

2) 在图文语义相关性数据集^[6]上对 SAM-

收稿日期: 2020-08-24; 录用日期: 2020-09-11; 网络出版时间: 2020-10-23 13:45

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201023.1144.001.html

基金项目: 国家社科基金重大项目(11&ZD189); 全军共用信息系统装备预先研究项目(31502030502)

* 通信作者. E-mail: huhuijun@wust.edu.cn

引用格式: 胡慧君, 冯梦媛, 曹梦丽, 等. 基于语义相关的多模态社交情感分析[J]. 北京航空航天大学学报, 2021, 47(3): 469-477. HU H J, FENG M Y, CAO M L, et al. Multimodal social sentiment analysis based on semantic correlation[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 469-477 (in Chinese).

RoBERTa 模型进行微调与训练,使得该模型中神经网络的权重参数适应于进行图文语义相关性分类任务,且将经过上述微调与训练的 SAMRoBERTa 模型称为语义相关性多模态 RoBERTa (Semantic Relevance Multimodal RoBERTa, SRMRoBERTa) 模型。因此,可以通过 SRMRoBERTa 模型对图文社交信息进行图文语义相关性的识别。

3) 针对多模态社交媒体情感分析任务,本文提出基于语义相关的多模态社交情感分析 (Multimodal Social Sentiment Analysis based on Semantic Correlation, MSSA-SC) 方法,该方法先通过 SRMRoBERTa 模型识别图文语义是否具有相关性,根据其识别结果,再选择对图文社交信息进行图文特征融合的多模态情感分类或者仅文本模态的情感分类。

1 相关工作

基于机器学习的方法需要大规模的语料数据集来给计算机进行学习,需要大量的人工标注数据。Xiao 和 Liang^[7]将双向长短期记忆 (Bi-directional Long Short Term Memory, BiLSTM) 模型用于对中文文本的情感分析,以原始中文文本的词嵌入 (Word Embedding) 为输入。Zheng 等^[8]使用 BERT^[9]模型对微博短文本进行情感分析,可以有效地提取语义特征,在预训练过程中,模型架构可以有效整合句子的语义特征,通过词嵌入、句子嵌入及位置嵌入将它们转换为具有上下文信息的特征向量。

Machajdik 和 Hanbury^[10]的研究实现了提取与整合具有图像情感信息的低级特征,并将其用于图像情感分类中。Jindal 和 Singh^[11]使用卷积神经网络 (Convolutional Neural Network, CNN) 构建图像情感预测框架,在大规模数据上进行预训练,以用于对象识别与进行迁移学习,并在人工标注的 Flickr 图像数据集上进行了实验。

在日益多样的信息传递媒介和目前较为成熟的自然语言处理技术支持下,多模态情感分析拥有巨大的研究空间与潜在价值。You 等^[12]利用视觉与文本信息联合进行情感分析,先对 CNN 模型进行微调以进行图像情感分析,再训练段落向量模型进行文本特征提取,将文本和图像特征整合在一起进行情感预测。Zhu 等^[13]提出通过合并视觉和文本信息来进行情感分类,所提模型基于双向循环神经网络 (Bidirectional Recurrent Neural Network, BiRNN) 并整合跨模式注意机制与语

义嵌入。缪裕青等^[14]提出将 CNN 模型在大规模图片数据集上进行预训练,再对其进行迁移,以微调的方式训练图片情感分类模型 (Fine-tuned CNN, FCNN), 通过将训练好的词向量输入 BiLSTM 中训练文本情感分类模型 (Word-embedding Bidirectional LSTM, WBLSTM), 通过决策融合的方法得到最终的图文情感结果。Zhao 等^[15]提出了基于图像和文本之间信息一致性的情感分析方法,先提取文本特征、图像特征及图文相似度特征,将这几种特征连接在一起输入到机器学习模型 SVM (Support Vector Machine) 中判断图文内容是否关联,根据其判断结果选择是否进行图文特征融合的情感分类。蔡国永等^[16]对于社交媒体情感分类任务,提出了利用多模态深度多重判别性相关分析提升图文语义配准,另外,利用多模态注意力融合网络图文特征,以进一步优化社交媒体的情感分类效果。

2 方法描述

本文提出的 MSSA-SC 方法对图文社交信息进行图文语义相关性分类,若图文语义相关性分类结果为图文语义相关,则将图文社交信息输入到 SAMRoBERTa 模型中进行情感分类;若分类结果为图文语义无关,则仅将文本模态输入到 RoBERTa 模型中进行情感分类。MSSA-SC 方法分为数据预处理、图文语义相关性分类模块及情感分类模块 3 个部分,如图 1 所示。

2.1 数据预处理

数据预处理是本文方法中的一个关键步骤,特别是对于来自社交媒体平台上用户生成的数据,其中数据是原始与非结构化的。因此,数据预处理这一步骤显得更加有必要。

在准备阶段,需要对来自微博的图文数据进行预处理,文本信息非常杂乱,并且有许多特殊字符,要进行文档切分、去停用词等操作。微博文本还有许多样板化的字符串,如“@”后面跟随的是微博用户名、“收起全文”及“网页链接”等。因此,对这些噪声信息要进行过滤操作。另外,微博图像的预处理包括去噪、消除光照不均及增强等。

2.2 图文语义相关性分类模块

2.2.1 图文关联的文本语义单元提取

图文关联的文本语义单元提取是图文社交信息在 SAMRoBERTa 模型与 SRMRoBERTa 模型中进行图文语义对齐的前提,图文关联的文本语义单元提取方法如图 2 所示。

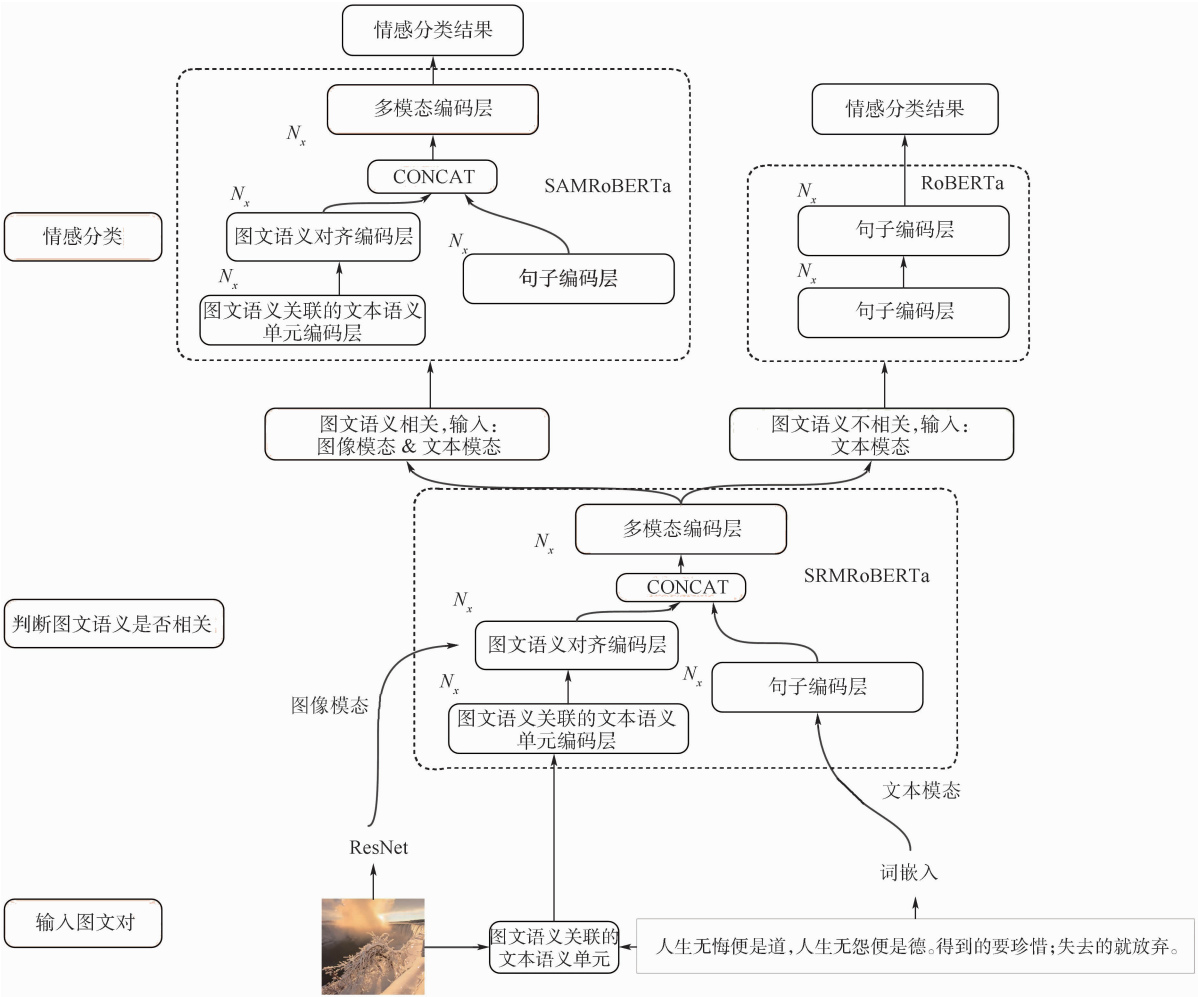


图 1 MSSA-SC 方法示意

Fig. 1 Schematic diagram of MSSA-SC method

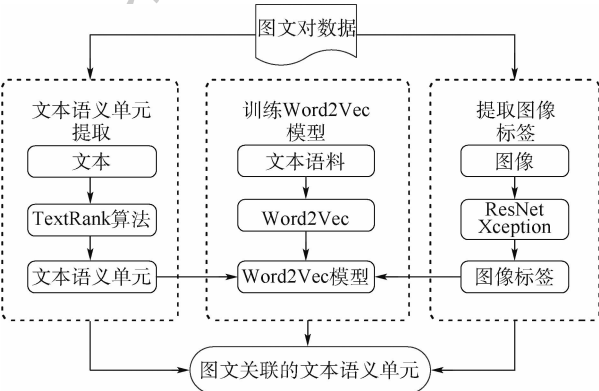


图 2 图文关联的文本语义单元提取方法架构

Fig. 2 Architecture of text semantic unit extraction method based on image-text correlation

首先,使用 TextRank^[17] 算法对文本提取文本语义单元;然后,对文本对应的图片使用 2 个经过预训练的基于 CNN 的图像分类模型 ResNet^[18] 和 Xception^[19] 分别提取图像中可信度最高的前 5 个语义标签,图 3 即为 ResNet 和 Xception 分别识别图像后提取的语义标签;最后,使用 Word2Vec^[20] 模型在大量微博文本语料上进行训练,利用训练



图 3 图像分类模型提取的语义标签

Fig. 3 Semantic tag extracted by image classification model

好的 Word2Vec 词向量模型分别对文本语义单元和图像语义标签中的每个词生成 300 维的词向量,通过向量与向量之间的余弦距离可以计算文本语义单元中每项和对应图片的语义标签词之间的相似度。式(1)表示计算词向量间的余弦相似度,根据语义相似度权值从大到小排序后取前

5 项作为图文关联的文本语义单元。

$$\text{similarity} = \cos \theta = \frac{\text{term}_i \cdot \text{term}_j}{\|\text{term}_i\| \cdot \|\text{term}_j\|} \quad (1)$$

式中: term_i 和 term_j 分别对应于文本语义单元与图像标签词的 300 维词向量。

2.2.2 SRMRoBERTa 模型

将图文社交信息里的图像纳入 RoBERTa 体系结构中,如图 4 所示。具体来说,对于相关的图像 I ,图像 I 通过模型 ResNet 获得最后一个卷积层的输出。ResNet 是在 ImageNet 图像数据集上训练的图像预训练模型,其本质上将原始图像分割成 $7 \times 7 = 49$ 个区域,每个区域用 2 048 维向量表示。具体公式为

$$\text{ResNet}(I) = \{r_s | r_s \in \mathbf{R}^{2048}, s = 1, 2, \dots, 49\} \quad (2)$$

式中: r_s 为图像被分割成 49 个区域后,其中一个区域上所提取的特征向量。

引入线性变换函数,使得视觉特征与文本特征具有相同的特征维度。

$$\mathbf{V}_I = \mathbf{W}_s \text{ResNet}(I) \quad (3)$$

式中: $\mathbf{W}_s \in \mathbf{R}^{2048}$ 为可学习的参数; $\mathbf{V}_I$ 为具有与文本特征相同维度的视觉特征。

为使图像区域与文本语义能够充分对齐,设计图文语义对齐(Image-Text Alignment,ITA)编码

层,该图文语义对齐编码层先将图文关联的文本语义单元 $\mathbf{T}_{\text{related}} = \{t_1, t_2, \dots\}$ 输入到编码器中得出图文关联的文本语义单元特征向量:

$$\mathbf{V}_{\text{related}} = \text{RoBERTa}(\mathbf{T}_{\text{related}}) \quad \mathbf{T}_{\text{related}} \in \mathbf{R}^{d \times L} \quad (4)$$

式中: d 为特征向量中的横向维度; L 为图文关联的文本语义单元的最大长度; $\mathbf{V}_{\text{related}}$ 为 RoBERTa 模型对图文关联的文本语义单元提取的特征向量。

将得到的 $\mathbf{V}_{\text{related}}$ 向量作为图文语义对齐编码层自注意力机制(Self-Attention)^[21]里的 queries 向量矩阵,将经过 ResNet 计算抽取的图像特征 $\mathbf{V}_I$ 作为图文语义对齐编码层自注意力机制里的 keys 与 values 向量矩阵。因此,在图文语义对齐编码层中,图文关联的文本语义单元特征引导图像特征进行向量计算,从而使得图像特征中与文本语义相关的图像区域分配更多的权重信息,具体公式为

$$\text{Attention}(\mathbf{V}_I, \mathbf{V}_{\text{related}}) = \text{Softmax}\left(\frac{[\mathbf{W}_Q \mathbf{V}_{\text{related}}]^T [\mathbf{W}_K \mathbf{V}_I]}{\sqrt{d_k}}\right) [\mathbf{W}_V \mathbf{V}_I]^T \quad (5)$$

式中: $\{\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V\} \in \mathbf{R}^{d_k \times d}$ 为可学习参数。

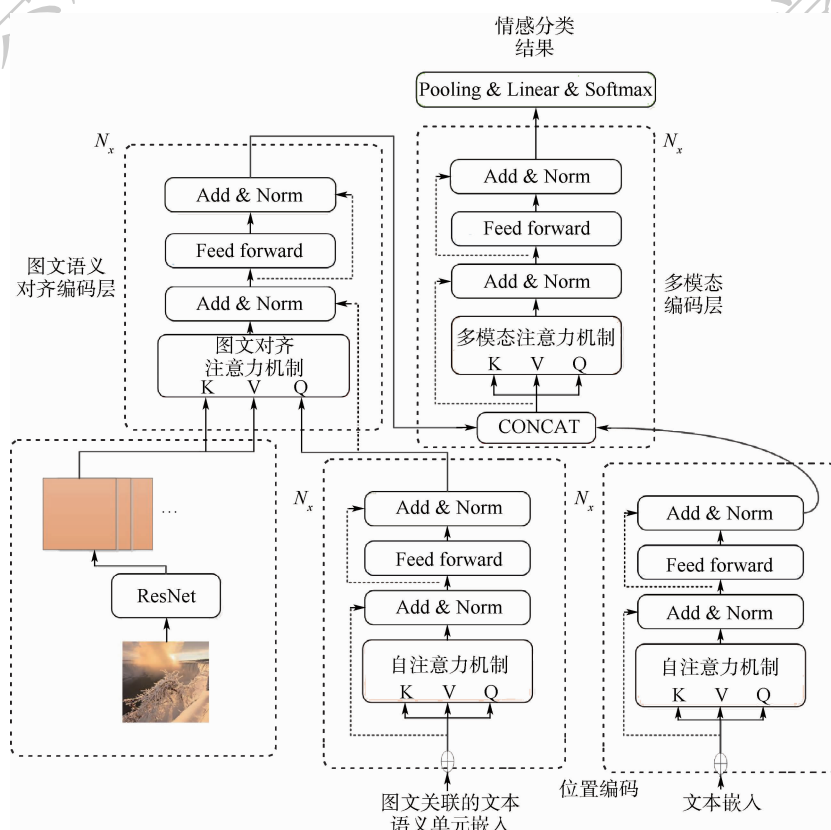


图 4 SAMRoBERTa 和 SRMRoBERTa 模型结构

Fig. 4 Structure of SAMRoBERTa and SRMRoBERTa model

图文语义对齐编码层与 Transformer 中的编码层类似,在图文语义对齐编码层自注意力机制之后采用一个全连接层与两层具有残差连接的层归一化,来获得图文对齐的视觉特征输出。

$$\mathbf{O}_M = \text{MLP}(\text{LN}(\mathbf{V}_{\text{related}} + M \cdot \text{Attention}(\mathbf{V}_I, \mathbf{V}_{\text{related}}))) \quad (6)$$

$$\text{ITA}(\mathbf{V}_I, \mathbf{V}_{\text{related}}) = \text{LN}(\mathbf{V}_{\text{related}} + \mathbf{O}_M) \quad (7)$$

式中: $\mathbf{O}_M$ 为图文语义对齐编码层中计算的中间产物,是 $\mathbf{V}_I$ 、 $\mathbf{V}_{\text{related}}$ 两个特征向量经过多头注意力机制后,将计算出的特征向量与 $\mathbf{V}_{\text{related}}$ 相加通过层归一化后再经过一层全连接层得到的特征向量; M 为多头自注意力机制中自注意机制的个数;MLP 代表全连接神经网络层;LN 代表层归一化(Layer Normalization)。

通过图文语义对齐编码层获得最终的视觉特征向量。

$$\mathbf{V}_F = \text{ITA}(\mathbf{V}_I, \mathbf{V}_{\text{related}}) \quad (8)$$

式中: $\mathbf{V}_F \in \mathbf{R}^{d \times M}$ 为视觉特征 $\mathbf{V}_I$ 和图文关联的文本语义单元特征 $\mathbf{V}_{\text{related}}$ 同时输入到图文语义对齐编码层中计算所得的视觉特征向量。

与此同时,将文本输入至 RoBERTa 模型中得到文本特征向量,即 $\mathbf{V}_T = \text{RoBERTa}(T)$,再把 $\mathbf{V}_F$ 与 $\mathbf{V}_T$ 相连接输入到多模态编码层(Multimodal Encoder, ME)中。

$$\text{ME}(\mathbf{V}_F, \mathbf{V}_T) = \text{RoBERTa}(\mathbf{V}_F, \mathbf{V}_T) \quad (9)$$

$$\mathbf{G} = \text{ME}(\mathbf{V}_F, \mathbf{V}_T) \quad (10)$$

将多模态编码层计算所输出的向量 $\mathbf{G}$ 输入到线性激活函数 Softmax 之中,以获得图文社交信息的情感分类结果。

$$P(y|\mathbf{G}) = \text{Softmax}(\mathbf{W}^T \mathbf{G}) \quad (11)$$

式中: $\mathbf{W} \in \mathbf{R}^{(2)d \times 3}$ 为可学习权重参数。

由于 Softmax 输出结果是一个概率分布,需要衡量预测标签与真实标签间的差异,训练过程中的损失函数设定为最小化交叉熵(Cross Entropy),其公式如下:

$$\text{loss} = \arg \min_w - \sum_{i=1}^n y \lg \hat{y} \quad (12)$$

式中: w 为神经网络所有训练参数; y 为预测标签; $\hat{y}$ 为真实标签。

2.3 情感分类模块

情感分类模块根据图文语义相关性分类模块对于图文社交信息的图文语义相关性的分类结果,选择将图文模态输入到 SAMRoBERTa 模型进行图文特征融合的情感分类还是仅将文本模态输入到 RoBERTa 模型中进行文本情感分类。

由于 SAMRoBERTa 模型结构与 SRMRoBERTa

Ta 模型相同,对于图文社交信息的输入与计算方法亦与 SRMRoBERTa 模型一样,两者仅仅是神经网络中的权重参数与最后分类结果输出的预测标签不同,所以这里不再赘述 SAMRoBERTa 模型。

RoBERTa 模型是基于 BERT 模型框架的鲁棒性预训练模型,其在模型神经网络架构方面没有改变 BERT,改变的只是预训练的方法和模型中的权重参数。将文本 T 输入到 RoBERTa 模型中,捕获的文本全局信息由多头自注意力机制完成。另外, T 会先经过嵌入层生成 T_w 作为模型的输入。

$$\text{Attention}(\mathbf{T}_w) = \text{Softmax}\left(\frac{[\mathbf{W}'_Q \mathbf{T}_w]^T [\mathbf{W}'_K \mathbf{T}_w]}{\sqrt{d_k}}\right) [\mathbf{W}'_V \mathbf{T}_w]^T \quad (13)$$

式中: $\{\mathbf{W}'_Q, \mathbf{W}'_K, \mathbf{W}'_V\} \in \mathbf{R}^{d_k \times d}$ 为可学习参数。

多头自注意力机制是把单个自注意力机制串联起来,具体公式为

$$M \cdot \text{Attention}(\mathbf{T}_w) = \mathbf{W}_m [\text{Attention}_1(\mathbf{T}_w), \dots, \text{Attention}_m(\mathbf{T}_w)]^T \quad (14)$$

式中: $\mathbf{W}_m \in \mathbf{R}^{(m \times d_k) \times d}$ 为可学习的权重参数。

将多头自注意力机制的输出结果输入到一个全连接层与两层具有残差连接的层归一化。

$$\mathbf{O}_T = \text{MLP}(\text{LN}(\mathbf{T}_w + M \cdot \text{Attention}(\mathbf{T}_w))) \quad (15)$$

$$\text{RoBERTa}(\mathbf{T}_w) = \text{LN}(\mathbf{T}_w + \mathbf{O}_T) \quad (16)$$

式中: $\mathbf{O}_T$ 为 RoBERTa 模型中编码层计算的中间产物,是文本 T 经过多头注意力机制后,将计算出的特征向量与 $\mathbf{V}_{\text{related}}$ 相加通过层归一化后再经过一层全连接层得到的特征向量。

综上所述,将模型的编码器全都进行堆叠,第 1 个分词(即[CLS])的最终隐藏状态被输入到 1 个线性变换函数中进行情感分类。

3 实验

3.1 实验设置

本文方法中提出的图文语义相关性分类模型 SRMRoBERTa 是 SAMRoBERTa 模型在 Liu 等^[6]提出的图文语义相关性数据集上进行微调与训练得到的。上述图文语义相关性数据集共有 9 817 个微博图文对,标签为 N 和 P,N 代表图像和文本语义不相关,P 代表图像和文本语义具有相关性,该数据集的具体数据分布如表 1 所示。

表 1 图文语义相关性数据集的数据分布

Table 1 Data distribution of image-text semantic correlation datasets

标签	微博图文对数
P	7 442
N	2 375

由于可以用于实验研究的公开中文图文情感数据集较少,从微博平台获取文本和对应的图片,微博平台允许一条微博可以配上多张图片,本文为便于研究只取文本对应的一张图片。获取的微博数据存在很多问题,如微博数据不完整、数据重复、图片及文本不合格等。在筛选不完整、数据重复和不合格的数据后,得到 20 130 对图文微博用于实验,再由 5 位标注者对其进行情感三分类标注工作。本文构建的微博图文情感分析数据集的数据分布如表 2 所示。

表 2 微博图文情感分析数据集的数据分布
Table 2 Data distribution of image-text microblog sentiment analysis datasets

标签	测试样例	训练样例
积极	981	4 660
中性	2 404	9 567
消极	615	1 903

评价指标选取准确率 (Accuracy) 和 Macro-F1 值判断模型识别效果,具体公式为

Accuracy = $\frac{\text{accNumber}}{N}$ (17)

式中:accNumber 为测试集中预测正确的数量;N 为测试集中图文对的总数量。

Precision = $\frac{TP}{TP + FP}$ (18)

Recall = $\frac{TP}{TP + FN}$ (19)

F1 = $\frac{2\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ (20)

式中:TP 为真正例;FP 为假正例;FN 为假负例。

通过式 (20) 得到三分类中每一类 F1 值,将所有 F1 值求平均,即得到 Macro-F1 值。

本文提出的 SAMRoBERTa 模型权重参数使用的是中文微调与训练好的 RoBERTa-zh-base 权重参数模型,本文模型与方法在 NVIDIA GeForce RTX2080Ti GPU 上运行,模型的实验参数设置如表 3 所示。

表 3 实验参数设置
Table 3 Experimental parameter setting

参数	数值
最大句子长度	256
批处理个数	12
学习率	2×10^{-5}
自注意力机制头数	6
预热学习率	0.1
全部样本训练次数	4.0

3.2 对比实验

ResNet^[18]:该模型中深度残差网络可以较好

地减小由于模型层数过深导致的模型梯度问题。

BiLSTM^[8]:先用单词嵌入将单词或短语从词汇映射到低维空间中的实数向量,再输入到 2 个并行运行的 LSTM 组成的 BiLSTM;一个在输入序列上,另一个在输入序列的反向上。此设置能够使隐藏层捕获到文本中过去和未来的语义信息。

BERT^[9]:让 BERT 模型在短微博文本数据集上进行微调,在预训练过程中,模型架构可以完全整合句子的语义特征和通过单词嵌入,句子嵌入和位置嵌入将它们转换为具有上下文信息的特征向量。

RoBERTa^[5]:其是通过改进 BERT 预训练的方法和改进模型里的优化函数得到的模型。因此,该模型是 BERT 模型的优化版,且可以用 BERT 模型直接加载 RoBERTa 的权重参数。

Multi-CNN^[22]:一种基于 CNN 的方法,2 个单独的 CNN 架构分别用于学习文本特征和视觉特征,再将图文特征整合到一起输入到另一个 CNN 的模型架构中,来进行图文特征融合的情感分类。

MBERT (Multimodal BERT)^[23]:将图像特征抽取部分融入到 BERT 的模型框架中,用预训练图像识别模型 ResNet 抽取图像的高级语义特征,文本则用 BERT 来提取语义特征,再将图像和文本特征整合输入到多模态编码层,使得 2 种模态特征充分对齐融合。

SAMRoBERTa:本文提出的基础模型,将图像模态融入到 RoBERTa 中的多模态模型,该模型有一个图文语义对齐编码层,可以使图文语义相匹配的图像区域获得更多的权重信息,将经过语义对齐后的图像特征和文本特征输入到多模态编码层进行融合计算。

MASS-SC:本文提出的基于语义相关的多模态社交情感分析方法,对图文社交信息进行图文语义相关性识别,若图文语义相关,则进行图文特征融合的多模态情感分类;若图文语义不相关,则仅对文本模态进行情感分类。

3.3 实验结果及分析

由表 4 可知,本文提出的 SAMRoBERTa 模型与 MSSA-SC 方法在 Accuracy 和 Macro-F1 指标上均高于基准模型,尤其是 MSSA-SC 方法在 Accuracy 指标上的数值达到 75.23%。

由此可知,本文提出的 SAMRoBERTa 中的图文语义对齐编码层可以有效提高图文对应的视觉区域特征权重,且多模态编码层亦可使得图文特征得到更充分的融合,因此,能够在多媒体社交媒体情感分析中得到比基准模型更好的分类效果。

表 4 微博图文情感分类实验结果

Table 4 Experimental results of image-text microblog sentiment classification

模态	方法	Accuracy/%	Macro-F1/%
图像	ResNet	58.91	45.27
	BiLSTM	70.92	63.48
文本	BERT	73.25	68.43
	RoBERTa	73.77	69.59
图文	Multi-CNN	68.31	62.46
	MBERT	74.52	68.70
	SAMRoBERTa	74.70	69.65
	MSSA-SC	75.23	70.18

MSSA-SC 方法在 Accuracy 和 Macro-F1 指标上均高于 SAMRoBERTa 模型, MSSA-SC 方法相较于 SAMRoBERTa 模型的区别在于: MSSA-SC 方法先对图文社交信息进行图文语义相关性的识别工作, 继而根据图文语义相关性的分类结果选择对图文社交信息进行图文模态融合情感分类亦或仅文本模态情感分类。然而, SAMRoBERTa 模型是直接对图文社交信息进行图文模态融合的情感分类。由此可知, MSSA-SC 方法中的图文语义相关性分类模块能有效筛选出在社交媒体中图文语义无关的情况。

3.4 实例分析

1) MBERT 模型与 SAMRoBERTa 模型对比实例

两者都是图文特征融合的多模态情感分析模型, 它们最大的不同之处是: 对于图像模态特征的处理方面, MBERT 的图像特征直接用 ResNet 提取出来, 而 SAMRoBERTa 模型中对于图像模态先用 ResNet 提取出图像特征, 再将图像特征和图文关联的语义单元一同输入到图文语义对齐编码层中, 使得图文语义相关的图像区域分配到更多的权重值。2 个模型的图像特征权重热力图如图 5 所示, 其中, 左边是实际的图文社交信息, 文本中标有下划线的词: 加油站、车与香瓜是图文关联的语义单元; 中间是 MBERT 模型的图像特征权重热力图; 右边是 SAMRoBERTa 模型的图像特征权重热力图。明显看出, SAMRoBERTa 模型中的图像特征经过图文语义对齐编码层中图文关联的语义单元特征的指导, 权重分布更加集中于加油站、车与香瓜的区域。

2) SAMRoBERTa 模型与 MSSA-SC 方法对比实例

表 5 显示了本文进行实验的图文情感分析数据集中的实例。该实例图文数据实际情感标签为消极, 进行图文语义相关性分类, 预测出来的结果是图文语义不相关, SAMRoBERTa 模型对其

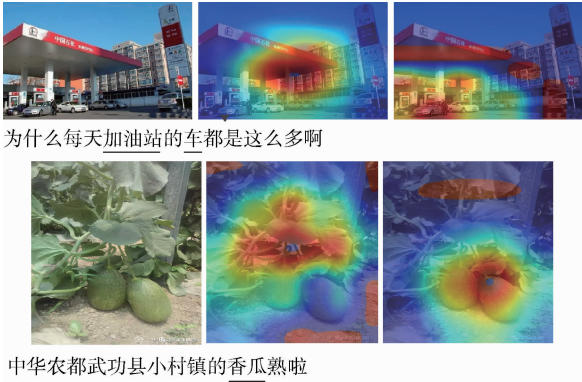


图 5 图文社交信息的图像特征权重热力图

Fig. 5 Heat map of image feature weight of image-text social information

表 5 图文情感分析数据集中的图文数据实例

Table 5 Image-text samples from image-text sentiment analysis dataset

图文数据实例	实际情感标签	SRMRoBERTa 图文语义相关性分类
 “迷惑绝了真是不知道又在搞些什么奇怪的疯狂操作”	消极	图文语义不相关

直接进行图文融合的情感分类结果为中性, 与实际情感标签不一致, 这是由于图片上的人物呈现出来的是一个“微笑”表情, 而文本表达出来的情感语义是消极的, 由于图文语义的互斥, 对图文信息进行融合导致情感分类的不准确, 而对这个图文数据使用本文的 MSSA-SC 方法得出的情感分类结果为消极, 与实际情感标签一致, 该方法先对图文数据实例进行语义相关性分类为图文语义无关, 再对其仅使用文本数据输入到 RoBERTa 中进行文本情感分类。

4 结 论

1) 本文针对多模态社交媒体上的情感分析任务提出 MSSA-SC 方法, 并在真实社交媒体数据集上进行了对比实验, MSSA-SC 方法的实验结果验证了其可行性与有效性。

2) 在真实社交媒体数据集上, 本文提出的 MSSA-SC 方法与 SAMRoBERTa 模型进行了消融实验, MSSA-SC 方法在 Accuracy 与 Macro-F1 指标上相较于 SAMRoBERTa 模型分别提升 0.53% 和 0.53%, 由此验证了图文社交信息上确实存在图

文语义无关的情况,且这种情况会对情感分类结果造成影响,而 MSSA-SC 方法中的 SRMRoBERTa 模型则可将图文语义无关的图文社交信息识别出来。

在未来对于社交媒体上情感分析的研究中:

1) 继续沿着图文语义相关性分类的方向进行研究,找到更加高效图文语义相关性分类方法。

2) 研究更好的特征融合方法以进一步提高多模态特征融合的性能。

3) 专门收集图片与文本语义不一致,但实际表达的内容是互补或者相反的图文社交信息,以此来训练与修改模型。

参考文献 (References)

- [1] 王英,龚花萍.基于情感维度的大数据网络舆情情感倾向性分析研究—以“南昌大学自主保洁”微博舆情事件为例[J].情报科学,2017,35(4):40-45.
WANG Y,GONG H P. Research on the sentiment tendency of big data network public opinion sentiment based on the sentiment dimension—Taking the “Nanchang University Independent Cleaning” Weibo public opinion event as an example[J]. Information Science,2017,35(4):40-45 (in Chinese).
- [2] SENG J K,ANG K L. Multimodal emotion and sentiment modeling from unstructured big data: Challenges, architecture & techniques[J]. IEEE Access,2019,7:90982-90998.
- [3] GHOSAL D,AKHTAR S,CHAUHAN D, et al. Contextual inter-modal attention for multi-modal sentiment analysis[C] // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL Press,2018: 3454-3466.
- [4] NIE L,WANG W,HONG R, et al. Multimodal dialog system: Generating responses via adaptive decoders[C] // Proceedings of the 27th ACM International Conference on Multimedia. New York: ACM Press,2019:1098-1106.
- [5] LIU Y,OTT M,GOYAL N, et al. RoBERTa: A robustly optimized BERT pretraining approach[EB/OL]. (2019-07-26) [2020-08-01]. <https://arxiv.org/abs/1907.11692>.
- [6] LIU M,ZHANG L,LIU Y, et al. Recognizing semantic correlation in image-text Weibo via feature space mapping[J]. Computer Vision and Image Understanding,2017,163:58-66.
- [7] XIAO Z,LIANG P. Chinese sentiment analysis using bidirectional LSTM with word embedding[C] // International Conference on Cloud Computing and Security. Berlin:Springer,2016: 601-610.
- [8] ZHENG J,CHEN X,DU Y, et al. Short text sentiment analysis of micro-blog based on BERT[C] // MUE 2019, FutureTech 2019: Advanced Multimedia and Ubiquitous Engineering. Berlin: Springer,2019:390-396.
- [9] DEVLIN J,CHANG M,LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[EB/OL]. (2018-10-11) [2020-08-01]. <https://arxiv.org/abs/1810.04805>.
- [10] MACHAJDIK J,HANBURY A. Affective image classification using features inspired by psychology and art theory[C] // Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press,2010: 83-92.
- [11] JINDAL S,SINGH S. Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning[C] // 2015 International Conference on Information Processing. Piscataway: IEEE Press,2015:447-451.
- [12] YOU Q,LUO J,JIN H, et al. Joint visual-textual sentiment analysis with deep neural networks[C] // Proceedings of the 23rd ACM International Conference on Multimedia. New York: ACM Press,2015: 1071-1074.
- [13] ZHU X,CAO B,XU S, et al. Joint visual-textual sentiment analysis based on cross-modality attention mechanism[C] // Proceedings of the International Conference on Multimedia Modeling. Berlin:Springer,2019:264-276.
- [14] 缪裕青,汪俊宏,刘同来,等.图文融合的微博情感分析方法[J].计算机工程与设计,2019,40(4):1099-1105.
MIAO Y Q,WANG J H,LIU T L, et al. Joint visual-textual approach for microblog sentiment analysis[J]. Computer Engineering and Design,2019,40(4):1099-1105 (in Chinese).
- [15] ZHAO Z Y,ZHU H Y,XUE Z H, et al. An image-text consistency driven multimodal sentiment analysis approach for social media[J]. Information Processing and Management,2019,56(6):102097.
- [16] 蔡国永,吕光瑞,徐智.基于层次化深度关联融合网络的社交媒体情感分类[J].计算机研究与发展,2019,56(6): 1312-1324.
CAI G Y,LV G R,XU Z. A hierarchical deep correlation fusion network for sentiment classification in social media[J]. Journal of Computer Research and Development,2019,56(6): 1312-1324 (in Chinese).
- [17] LI W,ZHAO J. TextRank algorithm by exploiting Wikipedia for short text keywords extraction[C] // 2016 3rd International Conference on Information Science and Control Engineering. Piscataway: IEEE Press,2016:683-686.
- [18] 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.
- [19] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2017:1251-1258.
- [20] MIKOLOV T,CHEN K,CORRADO G S, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-01-16) [2020-08-01]. <https://arxiv.org/abs/1301.3781>.
- [21] VASWANI A,SHAZEER N,PARMAR N, et al. Attention is all you need[C] // Proceedings of the 31st International Conference on Neural Information Processing Systems,2017: 5998-6008.
- [22] CAI G,XIA B. Convolutional neural networks for multimedia sentiment analysis[C] // NLPCC 2015: Natural Language Processing and Chinese Computing. Berlin: Springer,2015: 159-167.

[23] YU J, JIANG J. Adapting BERT for target-oriented multimodal sentiment classification[C] // Proceedings of the 28th International Joint Conference on Artificial Intelligence. San Francisco; Morgan Kaufmann, 2019:5408-5414.

冯梦媛 女, 硕士研究生。主要研究方向: 多模态情感分析。
曹梦丽 女, 硕士研究生。主要研究方向: 多模态情感分析。

作者简介:
胡慧君 女, 博士, 副教授, 硕士生导师。主要研究方向: 智能信息处理、图像分析与理解。
刘茂福 男, 博士, 教授, 博士生导师。主要研究方向: 自然语言处理、图像分析与理解。

Multimodal social sentiment analysis based on semantic correlation

HU Huijun^{1,2,*}, FENG Mengyuan^{1,2}, CAO Mengli^{1,2}, LIU Maofu^{1,2}

(1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;
2. Hubei Provincial Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan 430081, China)

Abstract: Social platforms allow users to express opinions in a variety of information modalities, and multi-modal semantic information fusion can more effectively predict the emotional tendencies expressed by users. Therefore, multimodal sentiment analysis has received extensive attention in recent years. However, in multi-modal sentiment analysis, there is a problem of unrelated semantics between vision and text, resulting in poor sentiment analysis. In order to solve this problem, this paper proposes the Multimodal Social Sentiment Analysis based on Semantic Correlation (MSSA-SC) method. The MSSA-SC firstly adopts the semantic relevance classification model of image and text to identify the semantic relevance of the image-text social media. If the image and text are semantically related, the image and text semantic alignment multimodal model is used for the image-text feature fusion for the image-text social media sentiment analysis. When the image and text semantics are irrelevant, only the sentiment analysis is performed on the text modality. The experimental results on real social media datasets show that the MSSA-SC method can effectively reduce the influence of unrelated image and text semantics on multimodal social sentiment analysis. Moreover, the Accuracy and Macro-F1 of the MSSA-SC method are 75.23% and 70.18%, respectively, and outperform those of the benchmark model.

Keywords: multimodal; social media; sentiment analysis; semantic correlation; image-text feature fusion

Received: 2020-08-24; Accepted: 2020-09-11; Published online: 2020-10-23 13:45
URL: kns.cnki.net/kcms/detail/11.2625.V.20201023.1144.001.html
Foundation items: Major Projects of National Social Science Foundation of China (11&ZD189); Pre-research Foundation of Whole Army Shared Information System Equipment (31502030502)
* Corresponding author. E-mail: huhuijun@wust.edu.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0457

基于深度多模态特征融合的短视频分类

张丽娟, 崔天舒, 井佩光*, 苏育挺

(天津大学 电气自动化与信息工程学院, 天津 300072)



摘

要: 目前, 短视频已经成为新媒体时代极具代表性的产物之一, 其天然的具有时短、强编辑等特点, 使得传统视频分类模型不再适合于短视频分类任务。针对综合短视频分类问题的特点, 提出了一种基于深度多模态特征融合的短视频分类算法。所提算法将视觉模态信息和音频模态信息输入到域分离网络中, 将整个特征空间划分为所有模态共享的公有域部分及由音频模态和视觉模态分别独有的私有域部分, 借助优化域分离网络, 最大程度地保留了不同模态特征间的差异性和相似性。在公开的短视频分类数据集上进行实验, 证明了所提算法可以有效减少特征融合时的冗余性, 并将分类的平均精度提高到 0.813。

关键词: 短视频; 多模态学习; 深度网络; 分类; 特征空间

中图分类号: TP181

文献标志码: A

文章编号: 1001-5965(2021)03-0478-08

伴随着智能手机和移动互联网的迅速普及, 短视频作为一种新型的用户生成内容, 已经广泛出现在各个社交平台上, 如抖音、Instagram 和 Vine。据抖音官方数据显示, 截至 2020 年 1 月 5 日, 其日活跃用户数已经突破 4 亿。与传统意义上的长视频相比较, 短视频的时间长度通常被限制在 2 min 以内, 这使得人们更容易获取和分享这些视频。除此之外, 短视频丰富的内容和多样的表现形式为受众提供了更优质的用户体验, 是短视频受欢迎的另一个原因。

已有的短视频内容分析相关研究主要包括场景估计^[1-3]、短视频的流行度预测^[4]及短视频的推荐^[5-6]等。例如, Zhang 等^[1]利用短视频的文本、音频和视觉模态特征解决多媒体场景分类的问题。Wei 等^[2]采用多模态特征融合的神经网络 (Neural Multimodal Cooperative Learning, NMCL) 解决短视频场景分类的问题。Nie 等^[3]致力于分析声音模态对于短视频场景类别预测的影响。Jing 等^[4]通过提出一种新颖的低秩多视角学习框

架解决短视频的流行度预测问题。Liu 等^[5]基于多模态特征融合的角度提出用户-视频联合注意力网络 (User-Video Co-attention Network, UVCA) 解决短视频的推荐任务。Shang 等^[6]提出了面向短视频大数据的推荐系统实现短视频的推荐。本文致力于解决短视频标签分类问题。

近年来, 随着计算机视觉领域理论的飞速发展和图像相关任务的理论日益成熟, 人们越来越多地去解决视频相关的问题而不是图像。从传统的人工处理获取视频特征到现在依赖于神经网络提取视频的特征, 从原来小范围的数据集到现在大范围的数据集, 视频分类领域的研究已经取得了巨大的进步。现有多模态特征融合的视频分类算法主要通过提取不同模态的特征并进行直接融合以解决分类问题, 然而, 该视频分类算法普遍针对的是传统长视频, 不能直接应用于短视频的分类, 主要原因有: ①与传统意义上的长视频相比, 短视频的时间长度通常只有 1~2 min, 所以需要在有限时间长度、有限内容的视频中提取出对分

收稿日期: 2020-08-24; 录用日期: 2020-08-28; 网络出版时间: 2020-09-11 20:52

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200911.0851.001.html

基金项目: 国家自然科学基金 (61802277); 中国博士后科学基金 (2019M651038)

* 通信作者。E-mail: pgjing@tju.edu.cn

引用格式: 张丽娟, 崔天舒, 井佩光, 等. 基于深度多模态特征融合的短视频分类[J]. 北京航空航天大学学报, 2021, 47(3): 478-485. ZHANG L J, CUI T S, JING P G, et al. Deep multimodal feature fusion for micro-video classification[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 478-485 (in Chinese).

类重要的特征。②短视频的来源广泛,表现形式多样,所以短视频相比较长视频,具有更高的信息复杂度和冗余度。③现有的特征融合方法多利用多模态特征之间的公共部分,而忽略了不同模态特征之间的私有部分。综上所述,现在的视频分类算法并不适合解决具有“时短”特性的短视频分类任务。

针对上述问题,本文提出了一种端到端的基于深度多模态特征融合的短视频分类算法,搭建基于音频模态的私有域、视觉模态的私有域及视觉模态的公有域组合而成的域分离网络,使用相似性损失函数探寻不同模态由公有域网络提取到的特征相似性,使用差异性损失函数探寻同一模态私有域网络和公有域网络提取到的特征差异性,并使用分类损失指导视频全局特征的分类。大量实验结果表明,本文算法可以很好地解决短视频的分类问题。

1 相关工作

本文主要从以下2个角度进行阐述:①深度特征学习,主要介绍利用深度卷积网络实现对短视频模态特征提取的相关工作;②多模态特征融合,简单介绍传统特征融合方式和现在普遍的特征融合方式。

1.1 深度特征学习

早期的视频特征提取方法采用2D卷积网络学习视频的每一帧特征,该方法的灵感来源于图像处理,然而其忽略了连续视频帧之间的时间关联性。为了保留时间相关性,现有很多方法通过聚合视频的帧特征作为一个整体的视频特征表示。Long等^[7]提出注意力簇网络(Attention Cluster Network, ACN),通过采用注意力单元将视频的局部特征聚合成视频的全局特征。Ma等^[8]通过设置每一时刻的特征和前一个时刻特征按权重进行加权实现特征融合,从而实现视频分类。近年来,为了充分利用视频的时空特征,3D卷积网络被提出用于学习视频的连续帧特征而非单一的视频帧特征,3D卷积网络的输入参数在保留了视频批量大小、视频通道、视频帧宽度和视频帧高度4个参数的基础上,增添了视频深度这一参数,用于记录每一个视频帧序列内的视频帧数量。Tran等^[9]提出C3D网络,利用3D卷积提取连续帧序列的时空域特征,并在视频分类准确率上取得了巨大的突破。近年来,以3D卷积为基础推出的一系列视频特征提取方法被广泛地应用在视频分

类、跟踪、分割等领域。例如,Carreira等^[10]提出I3D网络,通过在3D卷积网络基础上增加网络宽度的方式提高网络分类性能。Hara等^[11]则将原本应用在2D卷积网络的ResNet延伸到3D卷积网络,通过提出ResNet3D以解决视频分类的相关问题。Feichtenhofer等^[12]提出了SlowFast Network,整个网络通过构建快慢2个3D卷积网络去获取视频的全局特征。

然而,相比较传统的2D卷积网络,3D卷积网络需要更大的参数量及存储空间。为了解决该问题,Qiu等^[13]通过构建P3D网络,将3D卷积核用空间域的2D卷积及时间域1D卷积进行联合表示。Tran等^[14]运用R2+1D网络将3D卷积网络分解为独立的空间和时间模块。Xie等^[15]在S3D-G网络中采用(2D+1D)的卷积核代替I3D中的卷积核。

为了提高短视频分类的性能,本文采用3D卷积网络作为特征提取网络,同时还分别应用到公有域网络和私有域网络中,实验结果证明了该网络架构的普适性和有效性。

1.2 多模态特征融合

区别于传统的图像分类等任务,视频分类问题可利用的信息除了视觉信息外,还包括音频信息、光流信息及轨迹信息等多种模态信息。

传统的多模态特征融合策略大致分为前期融合和后期融合2种方式。前期融合通常将每个独立的模态特征拼接成一个全局特征,并且将这个全局特征放入分类器中进行分类^[16]。例如,D'Mello和Kory^[16]通过直接拼接音频模态特征和视觉模态特征获得视频的全局特征实现情感检测。后期融合则采用一些特定的数学方法去融合来自不同模态特征的判断得分,如平均、加权等操作。然而这些方法都忽视了不同模态特征在特征空间的关联性,而且在融合方式上也缺乏更有效的指导。

为了克服传统特征融合方法的缺陷,越来越多的人致力于寻找新的解决方案以提高分类的性能,代表性的方法包括联合学习、子空间学习、深度多模态特征学习等。联合学习是一种经典的用于多模态特征融合的半监督学习方法,该方法会对有标签的样本独立地训练每个模态的特征。针对于无标签的数据,整个学习过程会从每个独立的模态网络挑选标签置信度最高的无标签数据加入到训练网络。联合学习的优点在于:每次对于单一模态网络中数据的更新会参考其他模态的置信度。然而这种方法的实现依赖于训练集

中每个短视频不同的模态信息针对每一个类别的预测具有很高的置信度,这种要求对于短视频是很难实现的。

子空间学习是另一类用于多模态特征融合的常见学习方法^[17-19]。子空间学习的假设在于不同模态的特征会享有一个公共的子空间。例如,典型相关分析方法(Canonical Correlation Analysis, CCA)^[17,19]通过最大化不同模态的特征向量在潜在子空间上的关联性实现特征的融合及降维。Zhai 等^[17]在共享子空间中从多视角角度进行多视角度量学习。Franklin^[19]将典型相关分析方法应用在数据挖掘和预测等领域。除了典型相关分析方法,采用深度网络进行特征提取和融合也成为了子空间学习常用的方法之一。Feichtenhofer 等^[18]借助双流卷积网络(Convolutional Two-Stream Network, CTSN)探寻不同模态信息在特征空间的向量表示并进行特征融合。Wang 等^[20]则在此基础上使用时域分割网络(Temporal Segment

Network, TSN)提取不同视频片段的特征,并根据不同片段在不同模态特征空间的分类得分,采用分段函数叠加分类得分且融合不同模态的分类得分,最终实现对视频的分类。

受到子空间学习的启发,本文从特征表示空间的角度解决短视频分类问题,重点在于将特征空间划分为不同模态的私有域和所有模态的公有域,整个特征空间的数目取决于模态的数量。

2 深度多模态特征融合分类算法

本文算法框架如图 1 所示。图中: H_a^p 表示音频模态的私有域特征, H_v^p 表示视觉模态的私有域特征, H_a^c 表示音频模态的公有域特征, H_v^c 表示视觉模态的公有域特征。本节将对设计的短视频分类算法进行详细介绍,具体从特征提取网络、相似性损失函数、差异性损失函数及分类损失函数 4 个方面进行描述。

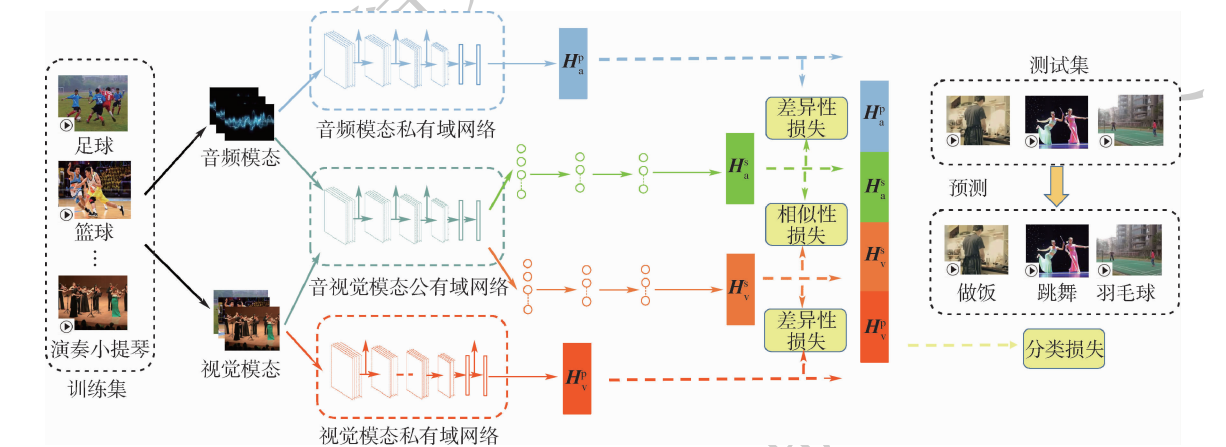


图 1 基于深度多模态特征融合的短视频分类网络

Fig. 1 Micro-video classification network based on deep multimodal feature fusion

2.1 特征提取网络

为了挖掘短视频视觉及音频模态的时域信息,本文利用 3D 卷积网络分别获取视频的视觉模态和音频模态各自的私有域特征及视觉模态和音频模态的公有域特征。针对视觉模态,按照 224×224 的尺寸从短视频中提取连续的 32 帧视频帧序列;针对音频模态,先通过从整个短视频中按照等间隔将短视频分为 32 个视频片段,再提取这 32 个视频片段的音频片段并将其转换成为频谱图表示这段音频的变化规律。需要注意的是,频谱图相比较视频帧而言只具有单通道。通过对比不同 3D 卷积网络的准确性和训练的复杂度,决定采用 I3D 网络作为特征提取网络,在该网络的基础上微调了网络的输出通道,增加了平均池化层,删除了最后用于分类的全连接层,最终输出

的特征向量维度为 512 维。

在网络整体结构设计上,本文将整个网络按照特征提取的角度划分为视觉模态信息的私有域网络、音视觉模态信息的公有域网络及音频模态信息的私有域网络 3 个模块。通过降低整体损失函数的数值来优化网络模型的参数,实现短视频的特征提取和分类。整个损失函数的实现包含 3 部分:①相似性损失 L_s ,用于探寻不同模态间(即视觉模态信息公有域特征和音频模态信息公有域特征)的相似性;②差异性损失 L_d ,用于衡量同一模态内的差异性,即视觉模态信息的私有域特征和公有域特征之间及音频模态信息的私有域特征和公有域特征之间;③分类损失 L_c ,用于将最后获得的融合后的全局特征进行分类。整个损失函数表示为

$$L = \alpha L_s + \beta L_D + \gamma L_C \quad (1)$$

式中: α 用于平衡相似性损失在整个损失中的权重; β 用于平衡差异性损失在整个损失中的权重; γ 用于平衡分类损失在整个损失中的权重。

2.2 相似性损失

通过构建相似性损失,可以获得视觉模态信息和音频模态信息公有域特征。本文借鉴了被广泛应用在人脸识别的孪生相似性损失去探寻不同模态公有域特征之间的相似性。

孪生相似性损失由 Chopra 等^[21]提出,主要应用于人脸识别领域并取得了良好的效果,在此基础上,越来越多的人致力于孪生网络结构的优化和使用,Zagoruyko 和 Komodakis^[22]优化了孪生网络并且将其应用在图像修复中。Bertinetto^[23]、Valmadre^[24]等将其使用延展到了目标跟踪,并且获得了理想的效果。受到孪生网络的启发,本文同样提出了基于孪生相似性损失的相似性损失,通过降低相似性损失 $L_{S_Siamese}$ 的值,实现对模态公有域网络的优化,具体公式为

$$L_{S_Siamese} = \frac{1}{2N} \sum_{n=1}^N (\|(\mathbf{H}_a^s)_n - (\mathbf{H}_v^s)_n\|_2)^2 \quad (2)$$

式中: n 为第 n 个短视频; N 为训练一次的样本数; $\mathbf{H}_a^s$ 和 $\mathbf{H}_v^s$ 分别为提取到的音频模态的公有域特征和视觉模态的公有域特征。通过减少每一层全连接层输出的视觉模态和音频模态在公有域特征的差异来探寻视觉模态公有域特征和音频模态公有域特征的相似性表示。实验证明, $L_{S_Siamese}$ 是有效的。

2.3 差异性损失

本文中,差异性损失被用于探寻单一模态下公有域特征和私有域特征差异性关系,具体将探讨视觉模态公有域特征和视觉模态私有域特征、音频模态公有域特征和音频模态私有域特征这两部分差异性的关系。

本文认为同一模态内私有域特征和公有域特征差异性主要集中在分布差异性和数值差异性两方面,接下来将以音频模态的私有域特征和音频模态的公有域特征为例,从分布差异性和数值差异性2个角度说明整个差异性损失。

针对分布上的差异,由于 KL (Kullback-Leibler) 散度被广泛地应用于评估模型输出的预测值分布和真值分布之间的差异,因此,在实验中将采用 KL 散度算法计算同一模态分布的差异性。在降低差异性损失 L_{D_KL} 的过程中,从分布差异的角度实现对模态私有域网络参数的优化,具体公式为

$$L_{D_KL} = \sum P(\mathbf{H}_a^p) \log_2 \frac{P(\mathbf{H}_a^p)}{Q(\mathbf{H}_a^s)} + \sum P(\mathbf{H}_v^p) \log_2 \frac{P(\mathbf{H}_v^p)}{Q(\mathbf{H}_v^s)} \quad (3)$$

式中: $P(\mathbf{H}_a^p)$ 为音频模态私有域特征的概率分布; $Q(\mathbf{H}_a^s)$ 为音频模态公有域特征的概率分布; $P(\mathbf{H}_v^p)$ 为视觉模态私有域特征的概率分布; $Q(\mathbf{H}_v^s)$ 为视觉模态公有域特征的概率分布。当 $P(\mathbf{H}_a^p)$ 和 $Q(\mathbf{H}_a^s)$ (或者 $P(\mathbf{H}_v^p)$ 和 $Q(\mathbf{H}_v^s)$) 的差异性越大时,损失函数的数值就越大,相同时损失函数的数值为0。然而,整个网络的目标是通过降低损失函数的数值去增大分布的差异性,因此对函数取负号作为实际使用的损失函数。

针对在数值上的差异,采用孪生网络差异性损失的方法去探寻差异性是一种可行的方式,在降低差异性损失 $L_{D_Siamese}$ 的过程中,从数值差异的角度实现对模态私有域网络参数的优化,具体公式为

$$L_{D_Siamese} = \frac{1}{2N} \sum_{n=1}^N (\max(d - \|(\mathbf{H}_a^p)_n - (\mathbf{H}_a^s)_n\|_2, 0)^2 + \max(d - \|(\mathbf{H}_v^p)_n - (\mathbf{H}_v^s)_n\|_2, 0)^2) \quad (4)$$

式中: d 为可调节的常数,通过设置常数的数值,调整音频模态私有域特征和音频模态公有域特征在数值上的差异性。实验结果表明,当 d 数值为3时,学习到的私有域特征更有利于提高短视频分类的准确性。

整个差异性损失的公式为

$$L_D = L_{D_KL} + L_{D_Siamese} \quad (5)$$

2.4 分类损失

在相似性损失和差异性损失的约束下,域分离网络已经将提取到的特征划分为音频模态的私有域特征子空间、音频模态和视觉模态的公有域特征子空间、视觉模态的私有域特征子空间。整个特征空间由4部分特征组成,即音频模态的公有域特征 $\mathbf{H}_a^s$ 、音频模态的私有域特征 $\mathbf{H}_a^p$ 、视觉模态的公有域特征 $\mathbf{H}_v^s$ 和视觉模态的私有域特征 $\mathbf{H}_v^p$ 。而后采用拼接的方法获取全局特征 $\mathbf{H}$ 。

通过设置分类损失指导整个模型实现分类,本文选择交叉熵损失作为分类损失函数,在降低分类损失 L_C 的过程中,实现对模态私有域网络参数的优化,具体公式为

$$L_C = - \sum_{n=1}^N \mathbf{y}_n \cdot \log_2 \hat{\mathbf{y}}_n \quad (6)$$

式中: $\mathbf{y}_n$ 为第 n 个视频的分类; $\hat{\mathbf{y}}_n$ 为对应的预测标签。

3 数据准备及实验分析

3.1 数据准备

实验数据集选择 AI-challenger-mlsv-2018 数据集,整个数据集包含大约 200 000 个短视频,涉及类别有 63 类。由于研究目标针对的是短视频单标签的分类,此次实验对存在多标签的短视频数据约 11 323 个进行了排除工作,将处理后的短视频数据按照 10:3 的比例划分为训练集和测试集,每个短视频的视频长度不超过 12 s。

3.2 实验分析

本节将从实验结果的角度对提出的网络进行系统全面的分析,在整个实验分析中包含以下 4 种指标:AR(Average Recall)、AP(Average Precision)、Micro-F1、Macro-F1。

3.2.1 网络复杂度

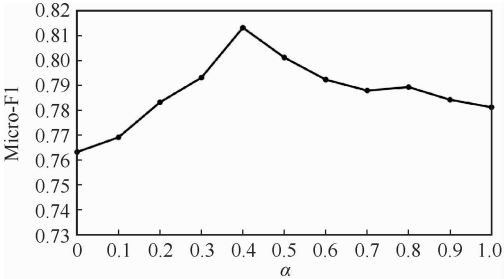
整个网络的实现由于采用的是 3 个结构相同的 3D 卷积神经网络分别探寻音频模态的私有域特征、视觉模态的私有域特征、音视觉模态的公有域特征 3 部分,而不是采用普遍的音频模态特征提取网络和视觉模态特征提取网络 2 部分,网络模型的参数量和输出的特征总数要高于现在普遍使用的特征融合网络。在时间复杂度上,因为同时要优化 3 个网络模型的参数及引入相似性损失、差异性损失计算损失,整个网络模型的时间复杂度要高于现阶段的视频分类网络。

对比表 1 中 5 种多模态特征融合分类算法(SlowFast Network^[12]、C3D(multimodal)、I3D(multimodal)、TSN^[20]、CTSN^[18])与本文算法在达到网络效果最好时所需要的迭代次数,发现本文算法需要的迭代次数最少,为 52 次,较为明显地减少了训练次数的时间成本。分析整个优化过

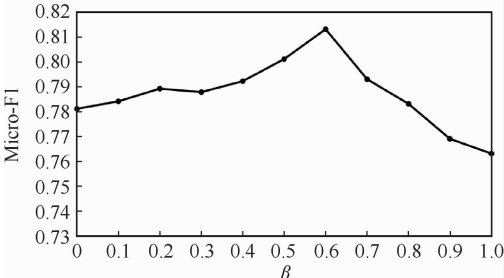
程,由于多模态算法通过计算差异性损失和相似性损失,加速了对网络私有域特征和公有域特征的分离,同时也加速了整个网络的收敛和参数的优化。

3.2.2 参数对比实验

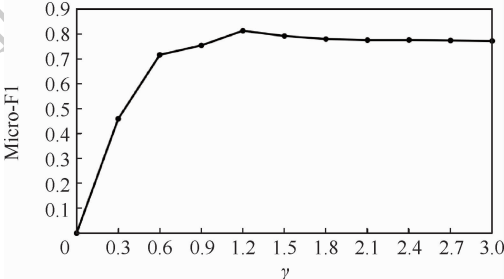
为了使整个网络的实验性能取得最好,探寻了相似性损失权重 α 、差异性损失权重 β 及分类损失权重 γ 对整个实验的影响。为了简化参数的学习过程,首先在固定分类损失权重 γ 的前提下,探寻相似性损失权重 α 和差异性损失权重 β 对整个实验结果的影响。参数 α 、 β 和 γ 的不同取值对网络性能的影响如图 2 所示。表明,当固定 γ 的值时,在调节参数 α 和参数 β 的过程中发现,当参数 α 的值为 0.4,参数 β 的值为 0.6 时,效果最好。然后固定参数 α 和参数 β 的数值,在调节参数 γ 的数值过程中,发现当参数 γ 的值为 1.2 时,提出的网络性能达到最好。



(a) 参数 α 对短视频分类任务的影响



(b) 参数 β 对短视频分类任务的影响



(c) 参数 γ 对短视频分类任务的影响

图 2 参数 α 、 β 、 γ 不同取值对短视频分类任务的影响
Fig. 2 Influence of different values of parameters α , β , γ on micro-video classification task

3.2.3 实验性能对比

为了证明本文所提模型的有效性,先后对比本文模型与 C3D^[9]、I3D^[10]、R2 + 1D^[14]、ResNet3D^[11]、GoogleNet^[25]、S3D-G^[15]、SlowFast Net-

表 1 不同网络短视频分类性能对比

Table 1 Performance comparison of micro-video classification in different networks

分类算法	AR	AP	Micro-F1	Macro-F1
C3D	0.555	0.568	0.569	0.560
R2 + 1D	0.622	0.643	0.651	0.631
单模态 ResNet3D	0.669	0.692	0.695	0.678
GoogleNet	0.679	0.709	0.719	0.691
I3D	0.706	0.727	0.733	0.715
S3D-G	0.741	0.750	0.755	0.744
多模态 C3D(multimodal)	0.612	0.623	0.632	0.617
I3D(multimodal)	0.751	0.763	0.772	0.755
TSN	0.763	0.772	0.781	0.770
CTSN	0.771	0.780	0.785	0.773
SlowFast Network	0.780	0.782	0.794	0.781
本文算法	0.782	0.795	0.813	0.789

work^[12]、C3D (multimodal)、I3D (multimodal)、TSN^[20]、CTSN^[18] 共 11 种视频分类算法在 AR、AP、Micro-F1、Macro-F1 这 4 项指标下的性能。其中,C3D^[9]、R2 + 1D^[14]、ResNet3D^[11]、S3D-G^[15]、GoogleNet^[25] 和 I3D^[10] 是常用的单模态特征视频分类网络,SlowFast Network^[12]、C3D (multimodal)、I3D(multimodal)、TSN^[20] 和 CTSN^[18] 是常用的多模态特征视频分类网络。通过观察表 1 的数据可以发现,本文算法在 AI-challenger-mlsv-2018 数据集中 AR、AP、Micro-F1、Macro-F1 取值分别为 0.782、0.795、0.813、0.789,这些数据反映了网络在短视频分类任务中的有效性。

3.2.4 消融对比实验

为了验证实验的有效性,表 2 分别比较了视觉模态特征、音频模态特征、前期融合特征、公有域特征、私有域特征及本文算法所提取的特征共 6 种特征。

表 2 中,视觉模态特征、音频模态特征是单独直接将对应的视觉模态信息、音频模态信息输入到 I3D 网络后得到的分类结果;前期融合是直接得到的视觉模态特征和音频模态特征融合后得到的全局特征;而公有域特征则是通过公有域网络,经过相似性损失约束得到的视觉模态和音频模态的公有域特征;私有域特征是将视觉模态信息和音频模态信息分别通过各自私有域网络,经过差异性损失约束得到的私有域特征。本文算法则是在得到的私有域特征和公有域特征上,对得到的私有域特征和公有域特征融合得到的全局特征,取得了更好的分类效果。

表 2 不同网络短视频特征分类性能对比

Table 2 Performance comparison of micro-video feature classification in different networks

方法	AR	AP	Micro-F1	Macro-F1
I3D(视觉模态)	0.720	0.718	0.733	0.720
I3D(音频模态)	0.328	0.319	0.341	0.328
前期融合	0.751	0.763	0.772	0.755
公有域网络	0.771	0.777	0.781	0.774
私有域网络	0.751	0.744	0.763	0.748
本文算法	0.782	0.795	0.813	0.789

4 结 论

1) 针对目前短视频分类任务中遇到的主要问题,本文提出了一种基于深度多模态特征融合的短视频分类算法,通过建立相似性损失和差异性损失,探寻短视频中不同模态之间的相似性和

同一模态的差异性,将视频的特征空间划分为视觉模态和音频模态的公有域,视觉模态的私有域和音频模态的私有域,获取到视频的不同模态的私有域特征和公有域特征,将其融合作为短视频的全局特征,用分类损失指导短视频的分类。

2) 在公开数据集上的大量实验表明,本文提出的算法成功地获取到了视频的不同模态信息在特征空间内的相似性表示和差异性表示,有效地降低了短视频多模态特征融合时的冗余性,提高了短视频分类的准确性,较好地解决了短视频的分类问题。

参考文献 (References)

[1] ZHANG J,NIE L,WANG X,et al. Shorter-is-better: Venue category estimation from micro-video[C] // Proceedings of ACM International Conference on Multimedia. New Nork: ACM Press, 2016:1415-1424.

[2] WEI Y,WANG X,GUAN W,et al. Neural multimodal cooperative learning toward micro-video understanding[J]. IEEE Transactions on Image Processing,2019,29:1-14.

[3] NIE L,WANG X,ZHANG J,et al. Enhancing micro-video understanding by harnessing external sounds[C] // Proceedings of ACM International Conference on Multimedia. New Nork: ACM Press,2017:1192-1200.

[4] JING P,SU Y,NIE L,et al. Low-rank multi-view embedding learning for micro-video popularity prediction[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 30 (8) : 1519-1532.

[5] LIU S,CHEN Z,LIU H,et al. User-video co-attention network for personalized micro-video recommendation[C] // Proceedings of World Wide Web Conference,2019: 3020-3026.

[6] SHANG S,SHI M,SHANG W,et al. A micro-video recommendation system based on big data[C] // Proceedings of International Conference on Computer and Information Science. Piscataway:IEEE Press,2016: 1-5.

[7] LONG X,GAN C,MELO G D,et al. Attention clusters: Purely attention based local feature integration for video classification [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7834-7843.

[8] MA C Y,CHEN M H,KIRA Z,et al. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition[J]. Signal Processing: Image Communication,2019,71: 76-87.

[9] TRAN D,BOURDEY L,FERGUS R,et al. Learning spatiotemporal features with 3D convolutional networks[C] // Proceedings of IEEE International Conference on Computer Vision. Piscataway: IEEE Press,2015:4489-4497.

[10] CARREIRA J,ZISSERMAN A,QUO V. Action recognition? A new model and the kinetics dataset[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2017: 6299-6308.

- [11] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? [C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6546-6555.
- [12] FEICHTENHOFER C, FAN H, MALIK J, et al. SlowFast networks for video recognition[C] // Proceedings of IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 6202-6211.
- [13] QIU Z, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C] // Proceedings of IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 5533-5541.
- [14] TRAN D, WANG H, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6450-6459.
- [15] XIE S, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning for video understanding[EB/OL]. (2017-12-13) [2020-08-01]. <https://arxiv.org/abs/1712.04851>.
- [16] D'ELLO S K, KORY J. A review and meta-analysis of multimodal affect detection systems[J]. ACM Computing Surveys, 2015, 47(3): 193-228.
- [17] ZHAI D, CHANG H, SHAN S, et al. Multiview metric learning with global consistency and local smoothness[J]. ACM Transactions on Intelligent Systems and Technology, 2012, 3(3): 1-22.
- [18] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 1933-1941.
- [19] FRANKLIN J. The elements of statistical learning: Data mining, inference and prediction[J]. The Mathematical Intelligencer, 2005, 27(2): 83-85.
- [20] WANG L, XIONG Y, WANG Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C] // Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016: 20-36.
- [21] CHOPRA S, HASSELL R, LECUN Y. Learning a similarity metric discriminatively, with application to face verification[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2005: 539-546.
- [22] ZAGORUYKO S, KOMODAKIS N. Learning to compare image patches via convolutional neural networks[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 4353-4361.
- [23] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C] // Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016: 850-865.
- [24] VALMADRE J, BERTINETTO L, HENRIQUES J, et al. End-to-end representation learning for correlation filter based tracking[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2805-2813.
- [25] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C] // Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1-9.

作者简介:

张丽娟 女, 硕士研究生。主要研究方向: 多媒体信息处理。

崔天舒 男, 硕士研究生。主要研究方向: 多媒体信息处理。

井佩光 男, 博士, 副教授, 硕士生导师。主要研究方向: 多媒体计算、机器学习、高阶数据分析。

苏育挺 男, 博士, 教授, 博士生导师。主要研究方向: 多媒体计算、机器学习。

Deep multimodal feature fusion for micro-video classification

ZHANG Lijuan, CUI Tianshu, JING Peiguang*, SU Yuting

(School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: Nowadays, micro-video has become one of the most representative products in the new media era. It has the characteristics of short time and strong editing, which makes the traditional video classification models no longer suitable for micro-video classification task. Based on the characteristics of the micro-video classification problem, the micro-video classification algorithm based on deep multimodal feature fusion is proposed. The proposed algorithm inputs the visual modal information and acoustic modal information into the domain separation network, and divides the entire feature space into a shared domain part shared by all modalities and the private domain part unique to the acoustic and visual modalities respectively. By optimizing the domain separation network, the differences and similarities among different modal features are preserved to the greatest extent. The experiments on the public micro-video classification dataset prove that the proposed algorithm can effectively reduce the redundancy of feature fusion and improve the average classification accuracy to 0.813.

Keywords: micro-video; multimodal learning; deep network; classification; feature space

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

DOI: 10.13700/j.bh.1001-5965.2020.0452

基于 FPGA 无人机影像快速低功耗 高精度三维重建



李杰*, 李一轩, 吴天生, 王昊榕, 梁敏

(山西财经大学 信息学院, 太原 030006)

摘 要: 现有无人机(UAV)影像三维重建方法在功耗、时效等方面无法满足移动终端对低功耗、高时效的需求。为此,在有限资源 FPGA 平台下,结合指令优化策略和软硬件协同优化方法,提出一种基于 FPGA 高吞吐量硬件优化架构的无人机航拍影像快速低功耗高精度三维重建方法。首先,构建多尺度深度图融合算法架构,增强传统 FPGA 相位相关算法对不可信区域的鲁棒性,如低纹理、河流等区域。其次,结合高并行指令优化策略,提出高性能软硬件协同优化方案,实现多尺度深度图融合算法架构在有限资源 FPGA 平台的高效运行。最后,将现有 CPU 方法、GPU 方法与 FPGA 方法进行综合实验比较,实验结果表明:FPGA 方法在重建时间消耗上与 GPU 方法接近,比 CPU 方法快近 20 倍,但功耗仅为 GPU 方法的 2.23%。

关 键 词: 低功耗; FPGA; 三维重建; 相位相关; 软硬件协同优化

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2021)03-0486-14

嵌入式计算机视觉系统是光学影像三维重建的重要应用领域之一,通过三维重建可获得相应场景的深度信息,其类似于人眼的基本机制,在诸如机器人导航^[1]、文物保护^[2]、航空摄影^[3-4]、太空探索^[5-7]等领域都有重要的潜在应用价值。无人机(Unmanned Aerial Vehicle, UAV)作为一种方便的地形观测设备,通常应用于航空摄影。假设 D 为无人机左右摄像机光心之间的距离, H 为机载摄像机相对地面的高度。一般而言,无人机连续拍摄间隔距离 D 值很小,其飞行高度 H 值很大,即航拍影像的基高比相对很低。另外,使用无人机进行地形拍摄时,很容易出现低纹理、动态纹理区域,尤其在多山地地区进行低空对地观测时,图像中常会包含高山、阴影、河流等区域,然而传统基于 FPGA 的三维重建方法通常不适用于这

一情形。

基于光学影像的三维重建系统通常包含大规模视差计算步骤,然而传统计算机因固有的指令周期延迟使得大规模视差计算的运行效率非常低。针对该类问题,研究人员开始尝试使用不同的硬件平台来提高三维重建方法的时效性,包括采用 FPGA^[5-6]、GPU^[8]、DSP^[9]、ASIC^[10] 等。然而,对于部分资源有限的硬件平台,如 FPGA、ASIC 等,不能像对待 CPU、GPU 或 DSP 的应用程序那样,直接通过简单地添加内存或增加磁盘空间的方法允许相同算法处理更大图像或更大视差范围的问题。其主要原因是:就基于 FPGA 的应用程序而言,可用逻辑资源数量有限,当逻辑单元耗尽时,唯一的解决方案是添加另一个设备或更改相应算法^[11]。这就导致许多本质上是高度迭

收稿日期: 2020-08-24; 录用日期: 2020-09-05; 网络出版时间: 2020-09-28 16:51

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20200928.1616.005.html

基金项目: 国家自然科学基金(61801279); 山西省应用基础研究计划(201801D221160); 山西省高等学校科技创新计划(STIP)(2019L0471); 山西省研究生教育创新计划(2020SY175, 2020SY176)

* 通信作者: E-mail: lijiescu@aliyun.com

引用格式: 李杰, 李一轩, 吴天生, 等. 基于 FPGA 无人机影像快速低功耗高精度三维重建[J]. 北京航空航天大学学报, 2021, 47(3): 486-499. LI J, LI Y X, WU T S, et al. Fast, low-power and high-precision 3D reconstruction of UAV images based on FPGA[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 486-499 (in Chinese).

代性的先进算法因FPGA资源受限难以有效实施。因此,在硬件编程复杂、资源受限的双重压力下,设计一套快速低功耗高精度三维重建方法是极具挑战意义的问题。

针对上述问题,本文基于有限资源FPGA平台,提出改进相位相关立体匹配算法来克服传统FPGA方法易受光照变化、遮挡、阴影及小旋转等因素影响,进而实现从包含低纹理区域、动态纹理区域及大视差范围等干扰因素的真实山地无人机影像中提取高可靠视差信息,大大提高基于FPGA的无人机航拍影像三维重建方法的鲁棒性、准确性及综合能耗等。与已有的研究成果相比,本文的主要贡献如下:

1) 基于多尺度深度图融合思想提出一种改进相位相关立体匹配算法,并利用FPGA平台构建低功耗快速实现硬件架构。

2) 基于有限资源FPGA平台,结合高并行指令优化策略和高性能软硬件协同优化方法,提出一种层级迭代、同层并行的高吞吐量硬件优化架构。

3) 本文方法充分利用FPGA芯片实现对外设影像和三维重建结果的自动处理和存储。此外,从时间、功耗、鲁棒性等方面对相关先进方法,如CPU^[3]、GPU^[8]等,进行充分的实验对比。

综上所述,本文方法不仅实现了基于无人机航拍影像的快速低功耗高精度三维重建,而且为各立体视觉应用领域提供快速低功耗高精度的三维重建解决方案。因此,本文极具理论意义和现实意义。

1 相关工作

1.1 图像匹配算法

图像匹配作为三维重建流程中的关键步骤之一,可以用于检测目标图像与参考图像之间各像素点的平移距离。在三维重建中,通常采用图像匹配算法实现立体图像对的密集视差图提取。目前,基于图像匹配的视差估计算法大致可分为3类:基于灰度的算法^[12-13]、基于特征的算法^[14-15]和基于频域的算法^[16-17]。

基于灰度的视差估计算法假设目标点与其对应点的图像强度在立体图像对中保持不变,通过找到具有最小灰度值统计距离对应像素块,可计算出这2个图像块之间的位移关系。但在实际应用中,图像容易受到噪声、照明变化和阴影等因素的干扰。因此,基于灰度的视差估计算法的结果在很大程度上取决于图像质量。

基于特征的立体匹配算法主要通过提取图像

中的特征描述(如边缘或直线段)来进行图像立体匹配。因此,对于纹理较少、信噪比(SNR)低且前景目标区域较大的图像(即特征描述不稳定且容易混淆的图像),匹配精度会大大降低。除此之外,为了追求更稳定的特征,特征提取过程通常具有很大的计算复杂度。

在基于频域的图像配准算法中,最典型的是基于相位相关的匹配算法^[18-20]。该算法主要利用傅里叶变换平移特性来估计2个图像之间的运动矢量。其中,平移特性是指空间域中的位移与频域中的线性相位变化之间的对应关系。基于相关函数的匹配算法不仅具有亚像素精度,还在图像对比度和图像强度方面不受全局线性照明变化的影响^[21]。因此,相位相关对噪声具有很高的鲁棒性,适用于噪声较大的场景。与传统的特征提取算法^[22]相比,该算法不依赖于特征提取,具有更快的更新速度。

1.2 基于FPGA的相位相关算法

Darabiha等^[23]在包含4个Xilinx Virtex2000E FPGA的Transmogriifier-3A(TM-3A)开发板上实现在 256×360 分辨率图像上以视频速率(30帧/s)生成8-bit亚像素级密集视差图。该系统使用局部加权相位相关算法,最多搜索20个像素范围的视差值。在Darabiha等^[23]的研究基础上,Masrani和Maclean^[11]使用更先进的硬件平台,包含4个Altera Stratix S80 FPGA的Transmogriifier-4开发板进一步实现以相同速率(30帧/s)处理更大视差范围的更大分辨率(480×640)图像。由于Transmogriifier-4开发板具有更加强大的硬件资源,Masrani和Maclean^[11]对局部加权相位相关算法进行了改进。采用2个可移动窗口取代了Darabiha等^[23]固定窗口方法进行相位相关计算。虽然多伦多大学团队的2项研究都以30帧/s速率实现三维重建,但很大程度上依赖于包含更大硬件资源的先进设备实现,并且不适用于大分辨率图像。除此之外,Masrani和Maclean^[11]提到,理论上最多可以实现128视差范围的搜索,但是大视差范围重建的时间消耗将重新估计。

据调研,截至目前为止,现有通过FPGA平台实现基于相位相关算法的光学影像三维重建研究相对较少,大多研究基本是基于空间域特征实现图像立体配准。Bhatia等^[6]在RTAX2000S FPGA上提出了一种利用相位相关算法精确计算月球着陆器水平速度的新方法,在测试实验中,该系统以2帧/s速率运行,但并没有给出所处理图像的分辨率值。Bhatia等^[24]在其另一项研究中详细描

述了该算法的优点,通过设计一种梯度滤波器,使改进相位相关算法适用于不同强度 and 对比度的图像序列,进而提高图像匹配精度。

Matsuo 等^[25]综述了基于多种硬件平台的相位相关算法及其实现,并就相关评价属性对其进行了详细实验比较。其中,硬件平台包括 GPU、ASIC 和 FPGA。最终结果表明,基于 GPU 的 POC 算法实现比基于 FPGA 实现要快近 1 倍的时间,如果仅考虑二维快速傅里叶变换(2D FFT)运算,则基于 GPU 实现比基于 FPGA 快 15.5 倍左右。虽然从速度上来看,基于 GPU 平台方法表现更好,但其存在需求性限制问题,如功耗较高。虽然 FPGA 不需在功耗方面考虑更多,但其硬件资源非常有限,这在无形之中给算法设计及程序实现带来了极高困难。此外,如果使用片外存储器(如 SD 卡)进行数据传输,更会进一步导致整个系统处理效率下降。因此,设计合理的 FPGA 硬件架构显得至关重要。

综上所述,目前基于 FPGA 的相位相关算法在三维重建方面的研究相对较少,依旧处于探索阶段,且大多研究集中在图像匹配领域,在应用层次方面鲜有涉及三维重建技术。基于此,本文在有限资源 FPGA 平台下,构建多尺度深度图融合算法架构,并结合指令优化策略和软硬件协同优化方法,提出一种基于 FPGA 高吞吐量硬件优化架构的无人机航拍影像快速低功耗高精度三维重建方法,满足了无人机航拍影像低功耗高时效三维重建的需求,克服了无人机航拍影像三维重建中对高山、阴影和河流等区域视差估计的不可信问题。

2 基于 FPGA 的无人机航拍影像三维重建方法

近年来,无人机已广泛用于工业检查、遥感、制图和测量等应用领域。但在低空影像三维重建方面,现有无人机航拍影像三维重建方法在功耗、时效等方面无法满足移动终端对低功耗、高时效的需求。因此,越来越多的研究人员将注意力聚焦于低功耗 FPGA 平台上。尽管在 FPGA 平台上已经成功部署了许多立体视觉算法,但许多本质上高度迭代或者依赖于不规律内存访问的先进算法,很难在有限资源 FPGA 平台上实现其高能效性。为了解决上述问题,本文基于 Xilinx ZCU104 FPGA 评估板,结合高并行指令优化策略和高性能软硬件协同优化方法,提出一种层级迭代、同层

并行的高吞吐量硬件优化架构。

Xilinx ZCU104 FPGA 评估板集成了嵌入式处理器(见图 1 中 PS 区域)与可编程逻辑(见图 1 中 PL 区域),既拥有 ARM 处理器灵活高效的数据运算和事务处理能力,又集成了 FPGA 的高速并行处理优势。因此,在系统开发时,通常可以利用软硬件协同优化方式将 PL 区域的硬件函数映射为 PS 区域一个或多个具有特定功能的外围设备。Xilinx ZCU104 FPGA 系统的体系结构如图 1 所示。一般来说,PS 区域主要负责整个系统的算法调度,以及执行具有复杂逻辑操作算法模块,而大量重复性计算通常交由 PL 区域执行,利用高并行指令硬件加速函数。

本文方法主要包含外围设备、嵌入式处理器(PS 区域)、可编程逻辑(PL 区域)、存储单元(RAM)4 部分。其中,外围设备包括搭载在无人机上的相机或 SD 卡。主要流程为(见图 2):FPGA 读取无人机相机捕获到的航拍图像序列,并对其执行预处理对齐操作,再对对齐立体图像对使用改进相位相关立体匹配算法进行视差估计,最终将多层视差图融合结果存储在 SD 卡中。

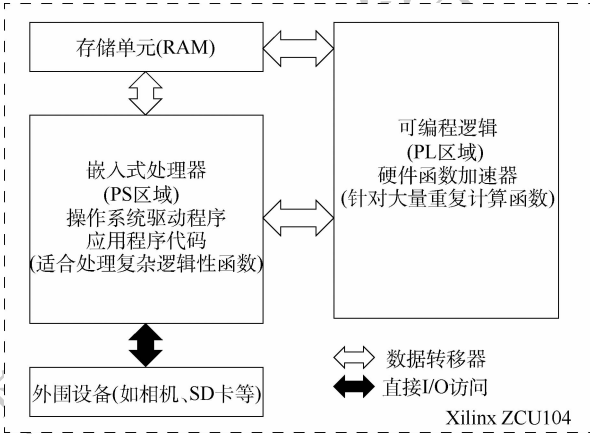


图 1 基于 Xilinx ZCU104 FPGA 系统的体系结构
Fig. 1 System structure based on Xilinx ZCU104 FPGA

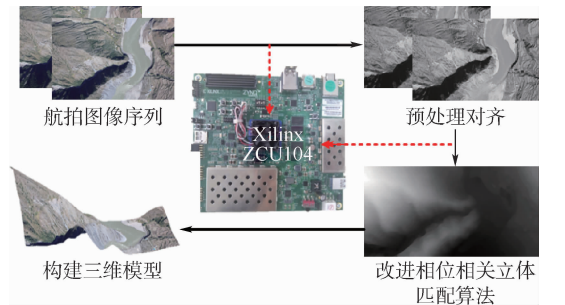


图 2 基于 FPGA 快速低功耗高精度三维重建方法流程
Fig. 2 Flowchart of fast, low-power and high-precision 3D reconstruction method based on FPGA

2.1 改进相位相关立体匹配算法

假设 $f_1(x, y)$ 、 $f_2(x, y)$ 分别表示无人机相机捕捉的 2 幅图像,满足以下关系:

$$f_1(x, y) = f_2(x - \Delta x, y - \Delta y) \quad (1)$$

即 $f_1(x, y)$ 平移 $(\Delta x, \Delta y)$ 向量后可以得到 $f_2(x, y)$,而改进相位相关立体匹配算法的目标是:求解具有亚像素级精度的平移向量 $(\Delta x, \Delta y)$ 。首先,对式(1)进行傅里叶变换得

$$F_1(u, v) = F_2(u, v) \exp[-2\pi j(u\Delta x + v\Delta y)] \quad (2)$$

式中: $F_1(u, v)$ 、 $F_2(u, v)$ 分别为图像 $f_1(x, y)$ 、 $f_2(x, y)$ 的频域信息。

进一步,用 2 幅图像的归一化互功率谱表示其相位相关性。

$$\text{CPS}(u, v) = \frac{F_1(u, v)}{F_2(u, v)} = \exp[-2\pi j(u\Delta x + v\Delta y)] \quad (3)$$

归一化互功率谱的傅里叶逆变换结果是空间域脉冲函数 $\delta(x - \Delta x, y - \Delta y)$,其对应的 Dirichlet 函数为

$$\text{ICPS}(u, v) = \frac{1}{AB} \cdot \frac{\sin(\pi(x - \Delta x))}{\sin\left(\frac{\pi}{A}(x - \Delta x)\right)} \cdot \frac{\sin(\pi(y - \Delta y))}{\sin\left(\frac{\pi}{B}(y - \Delta y)\right)} \quad (4)$$

式中: A 和 B 分别为图像块的宽和高。

式(4)可使用 $\sin$ 函数进行拟合,近似为

$$\text{ICPS}(u, v) \approx \frac{1}{AB} \cdot \frac{\sin(\pi(x - \Delta x))}{\frac{\pi}{A}(x - \Delta x)} \cdot \frac{\sin(\pi(y - \Delta y))}{\frac{\pi}{B}(y - \Delta y)} = \frac{\sin(\pi(x - \Delta x))}{\pi(x - \Delta x)} \cdot \frac{\sin(\pi(y - \Delta y))}{\pi(y - \Delta y)} \quad (5)$$

由于二维矩阵中各维度积分计算互不干涉,在下面的证明中只描述一维,这一结论也适用于其他维度。假设 (x_0, y_0) 为该维上最高点的坐标, (x_1, y_0) 为与最高点相邻的第二最高点,当满足 $x_1 = x_0 + 1$ 时,则

$$\begin{cases} (x_0, y_0) = \frac{\sin(\pi(x_0 - \Delta x))}{\pi(x_0 - \Delta x)} \cdot \frac{\sin(\pi(y_0 - \Delta y))}{\pi(y_0 - \Delta y)} \\ (x_1, y_0) = \frac{\sin(\pi(x_1 - \Delta x))}{\pi(x_1 - \Delta x)} \cdot \frac{\sin(\pi(y_0 - \Delta y))}{\pi(y_0 - \Delta y)} \\ \frac{(x_0, y_0)}{(x_1, y_0)} = \frac{\sin(\pi(x_0 - \Delta x))}{\pi(x_0 - \Delta x)} \cdot \frac{\sin(\pi(y_0 - \Delta y))}{\pi(y_0 - \Delta y)} / \left[\frac{\sin(\pi(x_1 - \Delta x))}{\pi(x_1 - \Delta x)} \cdot \frac{\sin(\pi(y_0 - \Delta y))}{\pi(y_0 - \Delta y)} \right] = \\ -\frac{x_1 - \Delta x}{x_0 - \Delta x} = -\left(1 + \frac{1}{x_0 - \Delta x}\right) \end{cases} \quad (6)$$

因此,在 x 方向上最大峰值点的亚像素平移为

$$\Delta x = x_0 + \frac{(x_1, y_0)}{(x_0, y_0) + (x_1, y_0)} \quad (7)$$

同理,当满足 $x_1 = x_0 - 1$ 时,此时最大峰值点的亚像素平移 $\Delta x = x_0 - \frac{(x_1, y_0)}{(x_0, y_0) - (x_1, y_0)}$ 。其他维度的亚像素平移求解方法与上述方法类似。

2.2 多尺度深度图融合算法

Liu 和 Hong^[26]指出,当使用基于相位相关算法进行三维重建时,其精度依赖于 PC 窗口的选择。如果窗口太小,不能覆盖足够大范围的像素变化,则无法准确地提取出较大范围的视差信息,在低纹理区域的重建结果也变得不可靠;如果窗口过大,则不仅难以识别局部细节信息,还会造成图像边界溢出。

为了解决上述问题,本文采用多尺度深度图融合思想进行视差优化,与 CPU 方法^[3]和 GPU 方法^[8]类似,利用上一层视差结果引导下一层视差图计算,充分结合不同尺寸窗口下 PC 计算优势,大大提高三维重建精度。具体的融合操作为:在下一层视差图计算的子像素块提取操作中,使用上一层视差结果作为该点子窗口提取的偏移量,实现在粗粒度视差结果基础上进行细粒度的视差计算,此时,粗粒度视差加上细粒度视差就是该层最终视差结果。

完整的多尺度深度图融合算法流程如图 3 所示。其中,图片 A 表示初始窗口下的视差图结果,图片 B 表示在初始窗口与减小窗口 2 种尺寸下的视差图融合结果,图片 C 表示 3 种不同尺寸窗口的视差图融合结果。从实验结果可以看出,当迭代次数为 1 时,视差图 A 中存在较为明显的边缘效应;而当迭代次数为 2 时,在视差图 A 引导下通过减小窗口尺寸后视差计算生成更为精细的纹理结构(见视差图 B),但仍存在细微的边缘效应;当迭代次数为 3 时,通过迭代更小窗口的视差结果使视差图 C 具有非常高的视差精度,几乎不存在肉眼可见的边缘效应。因此,使用多尺度深度图融合算法可以生成高精度的重建结果。

算法 1 多尺度深度图融合算法。

输入:无人机航拍影像对(左图 f 、右图 g)。

输出:视差图 M 。

初始化:深度图迭代 3 次,窗口尺寸分别为 W_1 、 W_2 、 W_3 ,每次窗口移动 S_1 、 S_2 、 S_3 次。

1. BEGIN

2. 利用 W_1 窗口,对无人机航拍图像对 f 、 g 分别取值,得到 W_1 大小的子图像块 f_1 、 g_1 。

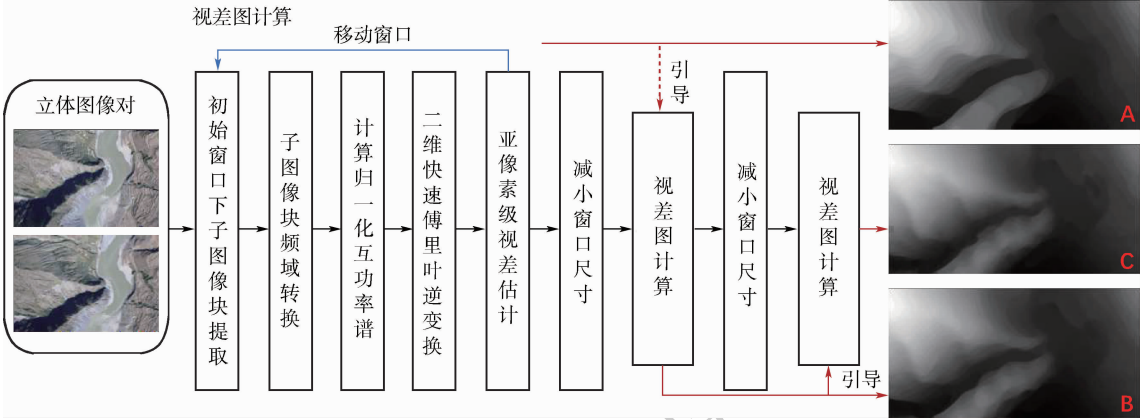


图3 多尺度深度图融合算法流程

Fig.3 Multi-scale depth map fusion algorithm flowchart

3. 分别对子图 f_1, g_1 进行二维快速傅里叶变换(2D FFT)得到 F_1, G_1 。
4. 利用式(3)求得归一化互功率谱 c_1 。
5. 对 c_1 进行二维快速傅里叶逆变换(2D IFFT)求得矩阵 C_1 。
6. 搜索 C_1 矩阵中最大值所在位置 (x_0, y_0) , 以及 x 方向的次大值 (x_1, y_0) , y 方向的次大值 (x_0, y_1) , 利用式(7)求得 x 方向的亚像素精度平移 x_m , 同理可求得 y 方向的亚像素精度平移 y_m , 则 W_1 窗口位置的最大视差为 (x_m, y_m) 。
7. 重复移动 W_1 窗口位置 S_1 次, 重复步骤2 ~ 步骤5, 得到 W_1 窗口下的视差图 M_1 。
8. 改变窗口尺寸为 W_2 , 将 W_1 视差结果作为 W_2 窗口下子图像块提取的偏移量, 对无人机航拍图像对 f, g 分别取值, 得到 W_2 大小的子图像块 f_2, g_2 。重复步骤3 ~ 步骤6, 得到视差图 M_2 。
9. 改变窗口尺寸为 W_3 , 重复步骤8, 得到多尺度深度图融合结果 M 。
10. END

2.3 基于FPGA平台的三维重建架构

在基于GPU平台实现无人机航拍影像三维重建时, Li 和 Liu^[8]使用单指令多数据流(Single Instruction Multiple Data, SIMD)体系架构和统一计算设备架构的快速傅里叶变换(CUDA Fast Fourier Transform, CUFFT)方法独立计算每个点的视差值, 但该方法依赖于大量的硬件资源支持(如2 GB 显卡内存), 是一种典型的利用大量硬件资源换取高效性能的实现方式。不同于GPU方法, 基于有限资源FPGA平台(如38 MB 内存 Xilinx ZCU104 FPGA 评估板), 研究人员将更多精力投入到软硬件协同优化方案的选择、高并行指令优化硬件加速模块等方面, 通过设计一种层级迭代、同层并行的高吞吐量硬件优化架构, 实现基于无人机航拍影像的快速低功耗高精度三维重建。

为将多尺度深度图融合算法在 Xilinx ZCU104 FPGA 平台上高效运行, 结合高并行指令优化策略和高性能软硬件协同优化方法, 提出一种层级迭代、同层并行的高吞吐量硬件优化架构。该方法实现的具体步骤如下:

- 步骤1 在CPU平台上开发出满足算法需求的可移植版本。
- 步骤2 将CPU版本移植到FPGA的嵌入式处理器(PS区域)。
- 步骤3 软硬件协同优化设计, 根据方法每个模块算法特性, 将适用于硬件加速的模块在可编程逻辑(PL区域)实现。
- 步骤4 对PL区域的硬件加速模块进行高并行指令优化。
- 步骤5 完成编译并在 Xilinx ZCU104 FPGA 评估板运行, 如果出现硬件资源不足无法编译或者执行效率达不到要求等情况, 则返回步骤3 调整软硬件协同优化方案, 直到选择出满足要求的硬件架构。

2.3.1 2D FFT 硬件实现

2D FFT 是基于FPGA 相位相关三维重建技术中最重要的一环之一。与基于CPU、GPU 平台直接调用2D FFT 函数不同, Xilinx ZCU104 FPGA 评估板仅支持一维快速傅里叶变换(1D FFT) IP core, 因此需要自主开发2D FFT 模块。在进行2D FFT 模块硬件开发时, 根据傅里叶变换可分离性质, 可以将其转化成2个1D FFT 实现。值得注意的是, 这里存在一个重要的数据依赖关系, 只有当图像块矩阵中所有元素都完成一维行FFT 之后, 才能进行一维列FFT 变换(见图4)。笔者认为, 这是相位相关算法难以在FPGA 平台上快速实现的主要原因之一。同理, 在傅里叶逆变换过程中, 2D IFFT 模块也是先转换成1D 行IFFT 变换后再进行1D 列IFFT 变换。此外, Xilinx 官方

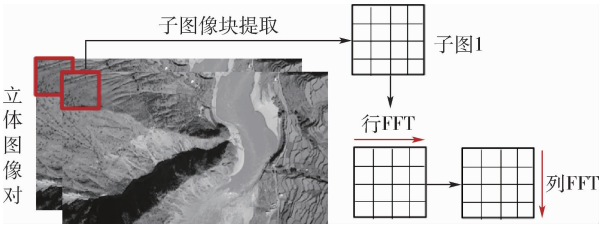


图4 2D FFT模块数据依赖关系示例

Fig. 4 Example of 2D FFT module data dependency

提供的1D FFT输入输出均为浮点复数形式,因此,还需将输入图像块矩阵转换成浮点复数形式。

2.3.2 软硬件协同优化

Xilinx公司的Xilinx ZCU104 FPGA评估板提供了一个灵活的原型设计平台,ZU7EV设备集成了四核ARM Cortex-A53处理系统和双核ARM Cortex-r5实时处理器,为应用开发者提供了前所未有的异构多处理能力。为了充分发挥异构平台的灵活性等优势,需要综合考虑每一种可能的软硬件协同优化方案,并对不同方案优缺点进行综合分析,才能选择出最适合的硬件架构。软硬件协同优化具体是指:开发者需要根据嵌入式处理器(PS区域)灵活高效的数据运算和事务处理能力,以及可编程逻辑(PL区域)高速并行处理的性能特点,为其分配相应功能,保证整个系统高效运行。

本文方法的目的是:在有限资源FPGA平台上实现多尺度深度图融合算法高效运行,但由于该算法的复杂性、不规律数据访问及迭代性,如果对每一个模块都按最高性能设计,则必将导致过多资源被占用。换句话说,如果一味追求最佳时效,则FPGA平台的有限硬件资源无法支持整个系统运行。因此,要想在有限资源FPGA平台上实现三维重建系统的所有功能,必须结合各种类型硬件资源使用情况,进行合理高效的软硬件协同优化。

在对无人机航拍影像序列进行子图像块提取时,一般会涉及到取窗口、边缘填充判断、不连续的内存读写及赋值运算等操作。此时,如果直接将无人机航拍影像序列从PS区域转移到PL区域内存中,再进行子图像块提取,则不需要每次都子图像块从PS区域转移到PL区域,进而降低了时间消耗,提高系统处理速度,但是这种方法势必会消耗PL区域大量的内存资源。而如果考虑在PS区域中完成子图像块提取操作,再将子图像块转移到PL区域进行下一步计算,则会最大程度节省PL区域内存资源占用,但是由于每次执行子图像块提取操作都需要从PS区域转移到PL区域,会在一定程度上增加时间成本,进而影响系统的整体性能。因此,基于有限资源FPGA平台的三维重建方法设计往往是在硬件资源与时效性之间的一种综合考量。一般而言,只有实现整个系统的基本功能之后,才会进一步考虑如何最大程度利用硬件资源的同时实现系统的最高性能。

本文方法的软硬件协同优化方案如图5所示。该方案具体流程为:FPGA将无人机相机捕获到的航拍图像序列临时存储在RAM中,再通过数据转移器交由PS区域与PL区域进行协同优化,最终将重建结果永久保存在SD卡中。此时,SD卡中的视差图可直接转换成需要的高程数字模型。其中,包含大量重复性计算功能的相位相关函数在可编程逻辑(PL区域)执行,而嵌入式处理器(PS区域)主要负责无人机航拍影像读取、取窗口、移动窗口、减小窗口、亚像素级视差估计及最终重建结果保存等复杂性逻辑操作。

2.3.3 高并行指令优化

本文方法并行架构设计时,主要利用 pipeline

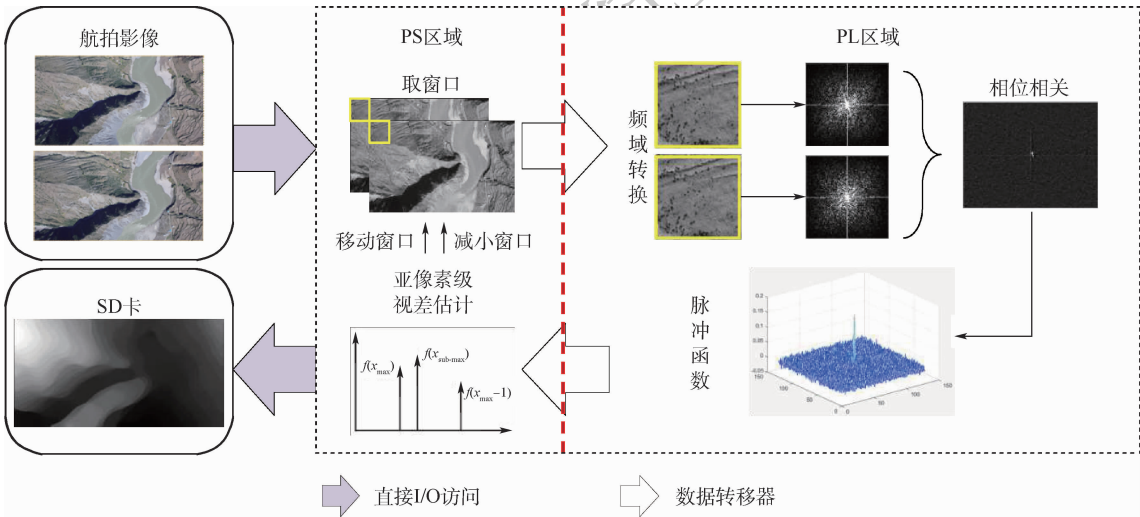


图5 软硬件协同优化流程

Fig. 5 Hardware-software co-design flowchart

(流水线)及 dataflow(数据流)高并行指令进行优化。在底层代码中,主要使用 pipeline 指令对 for 循环代码进行展开,pipeline 指令可以通过允许操作的并发执行来减少一个函数或循环的启动间隔,以此达到提高系统吞吐量的目的。dataflow 指令可以支持任务级管道处理,其允许函数和循环在各自的操作中进行重叠,这将增加 RTL 实现的并发性,进而可以增加设计的总体吞吐量。

为了最大程度降低正向、逆向二维快速傅里叶变换(FFT、IFFT)中数据依赖关系对整个系统性能造成的影响,在设计相位相关计算模块时,本文使用 dataflow 指令对频域转换、归一化互功率谱计算及 2D IFFT 三个函数进行并行加速,通过覆盖内存读写的时间消耗,最大程度上增加系统吞吐率。具体的操作为:将频域转换中一维列 FFT、互功率谱计算及 2D IFFT 中一维行 IFFT 三个操作合并成一个中转模块(见图 6),将中转模块、频域转换中一维行 FFT 模块及 2D IFFT 中一维列 IFFT 模块合并成一个硬件函数,并使用 dataflow 指令对其并行加速。

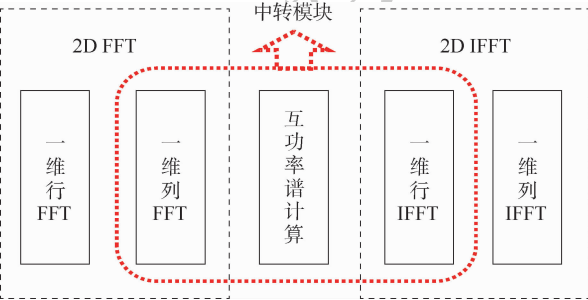


图 6 中转模块设计示意图

Fig. 6 Schematic diagram of transfer module design

3 实验评价

本文方法在 Xilinx ZCU104 FPGA 评估板上实现了完整的基于无人机航拍影像的快速低功耗高精度三维重建方法。实验图像采用 Liu^[26]、Li^[3,8] 等的无人机航拍影像数据,该组影像数据来源于中国西南山区拍摄的真实无人机图像序列。为了方便进行方法测试,从中选择了 2 张无人机航拍图像,并存储在 SD 卡中,用于测试的图像分辨率为 1 068 × 712。

Xilinx ZCU104 FPGA 评估板提供了一个灵活的原型设计平台,该平台具有高速 DDR4 内存接口、FMC 扩展端口、每秒千兆串行收发器、各种外围接口和可定制设计的 FPGA 结构。在 Xilinx ZCU FPGA 评估板上装有 Zynq UltraScale + XC-ZU7EV-2FFVC1156 MPSoC,其在同一器件中结合了功能强大的嵌入式处理系统(PS 区域)和可编程逻辑(PL 区域),共具有 504 000 个系统逻辑单元(LUT)、461 000 个 CLB 触发器(FF)、38 Mb 存储器(RAM)。在评估实验中,本文方法使用 Xilinx 官方提供的新一代 SDSoc™ 软件开发工具进行系统开发。其中,表 1 的数据来源于 Xilinx 官方软件 SDx 提供的 HLS 报告;表 2 及 FPGA 功耗数据来源于 Xilinx 官方软件 Vivado 提供的项目综合报告;CPU 方法^[3]、GPU 方法^[8]及本文方法分别基于 4 GB RAM 内存资源的 AMD Athlon II x2 240 型号 CPU (2.80 GHz)、英伟达 GTX760 型号 GPU (2 GB 显存)及 Xilinx ZCU104 FPGA (38Mb RAM 内存)评估板实现。

表 1 不同软硬件协同优化方案对比

Table 1 Comparison of different hardware-software co-design solutions

硬件资源 类型	方案 1					方案 2				
	行 FFT	列 FFT + CPS + 列 IFFT	行 IFFT	总计/%	子图像块提取	行 FFT	列 FFT + CPS + 列 IFFT	行 IFFT	视差估计	总计/%
BRAM_18K/个	12	18	6	34	0	12	18	6	0	49
FF/个	14 849	23 559	7 439	10	1 766	14 718	23 559	7 240	2 431	11
LUT/个	16 811	27 088	8 366	23	2 932	16 452	27 088	8 125	4 745	27
时间/s	8.1					6.4				

表 2 不同指令优化策略对比

Table 2 Comparison of different instruction optimization strategies

硬件资源 类型	方案 1		方案 2				方案 3			
	128 × 128 窗口	总计/%	128 × 128 窗口	64 × 64 窗口	32 × 32 窗口	总计/%	128 × 128 窗口	64 × 64 窗口	32 × 32 窗口	总计/%
BRAM/Mb	207.5	66.51	20	20	11	70.19	101	38	11	67.31
FF/个	66 047	14.33	38 858	34 804	31 608	39.03	25 955	22 477	19 319	20.41
LUT/个	47 847	20.77	27 963	24 889	23 630	57.73	21 104	18 371	17 098	31.38
时间/s	77.95		19.05	12.85	14.67	47.77	6.75	7.02	8.25	23.1

3.1 不同软硬件协同优化方案对比

在软硬件协同优化设计方面,有针对性地选择了 3 个具有代表性的软硬件协同设计方案,为了方便比较,仅考虑在初始窗口下的视差图计算,具体分区方案如图 7 所示。

软硬件协同优化方案 1 的具体细节包括:①将测试用的无人机航拍图像对读取到嵌入式处理器(PS 区域)中,并进行初始大小窗口的子图像块提取操作,再通过数据转移器 DMA 传入 PL 区域。②在 PL 区域通过 2.3.3 节描述的高并行相位相关计算模块,即利用 dataflow 指令对频域转换、归一化互功率谱计算及 2D IFFT 三个函数进行并行加速,求得空间域脉冲函数。③将脉冲函数转移到 PS 区域中,再使用 sin 函数对空间域脉冲函数进行亚像素级视差估计,这是因为峰值拟合时需要进行大量逻辑判断操作,更适用于嵌入式处理器实现。④在 PS 区域进行移动窗口提取子图像块操作,再重复步骤①~步骤③,直到获得初始窗口下的完整视差图。

软硬件协同优化方案 2 基于方案 1,在方案 2 中,将无人机航拍图像对转移到 PL 区域后,再进行子图像块提取、相位相关计算及亚像素级视差估计等操作。而软硬件协同优化方案 3 是在方案 2 基础上的一次大胆尝试。为了最大程度地提升系统性能,在方案 3 中,使用 dataflow 指令对子图像块提取操作进行并行优化加速。然而,在程序编译时,编译器提示因 RAM 资源不足导致无法编译。因此,在软硬件协同优化方案对比实验中,主要对方案 1 和方案 2 进行详细的实验对比论证。

为了对 2 种方案进行更深层次的比较研究,本文分别列出了其在 FPGA 平台实现时所占用的硬件资源情况,具体如表 1 所示。实验结果证明,

虽然方案 2 比方案 1 节省了约 20% 的时间,然而方案 2 仅初始大小窗口下的视差计算就占用了整个 FPGA 近一半的 BRAM_18K(18Kb Block RAM 块随机存储器)资源,最终导致该方案无法实现多尺度深度图融合算法的三维重建。

因此,考虑到 FPGA 硬件资源有限,本文方法最终选择方案 1 的软硬件协同优化方案。

3.2 不同指令优化策略对比

在指令优化策略方面,为了验证本文方法高并行指令优化策略的有效性,列出了 3 种不同指令优化策略进行比较,具体描述如下:

1) 方案 1 仅使用 pipeline 指令对相位相关计算模块中取窗口、频域转换、归一化互功率谱计算及 2D IFFT 等硬件加速函数分别优化。

2) 在方案 2 中,将频域转换中一维列 FFT、互功率谱计算及 2D IFFT 中一维行 IFFT 三个函数合并为一个中转模块,并与频域转换中一维行 FFT 模块及 2D IFFT 中一维列 IFFT 模块 3 个硬件函数分别使用 dataflow 指令优化并行加速。

3) 方案 3 是本文方法所采用的高并行指令优化策略,即将方案 2 提到的中转模块、频域转换中一维行 FFT 模块及 2D IFFT 中一维列 IFFT 模块合并成一个硬件函数后,使用 dataflow 指令对其并行加速。

3 种不同指令优化策略实验结果如表 2 所示。其中,包括方案 1 在 128×128 分辨率窗口下生成视差图所消耗的硬件资源情况及执行时间,方案 2 及方案 3 分别在 128×128、64×64、32×32 分辨率窗口下生成视差图及实现三层视差图融合三维重建方法所消耗的硬件资源情况、重建时间。表 2 结果显示,采用方案 1 指令优化策略,仅 128×128 分辨率窗口下的视差计算就消耗整个评估板

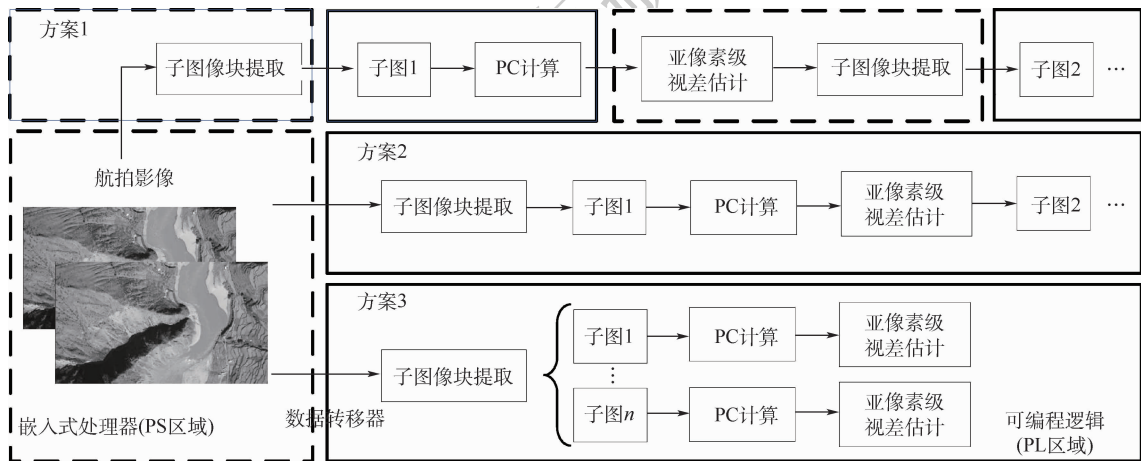


图 7 三种软硬件协同优化方案示意

Fig. 7 Schematic diagram of three hardware-software co-design solutions

66.51% 的 BRAM (块内存) 硬件资源,甚至无法完成第 2 层的视差图重建,笔者认为导致方案 1 性能差的主要原因是:未充分利用不同函数之间的数据依赖关系而进行指令优化。而方案 2 和方案 3 虽然都成功实现了三层视差图融合三维重建,但由于方案 2 未对不同函数模块之间的内存读写等待进行优化,导致硬件函数执行效率不高。

因此,本文方法最终采用方案 3 高并行指令优化策略,相比于其他方法,所消耗资源更少,重建速度更快,具有更高效性能。

3.3 与同领域先进方法的综合性比较

3.3.1 鲁棒性

将本文方法与其他 7 种先进方法(方法[27]、方法[28]、方法[29]、方法[30]、Agisoft stereoscan 商业软件方法 (<http://www.agisoft.com>)、CPU 方法^[3]及 GPU 方法^[8])进行比较研究。其中,初始参数设置如下:层次化方法[27-28]使用 32×32 初始窗口来估计视差;方法[29-30]使用 32×32 固定窗口来估计视差;对于 Agisoft stereoscan 商业软件方法,使用其默认初始参数;CPU 方法^[3]、GPU 方法^[8]和本文方法的初始窗口、初始采样间隔和初始运动矩阵的初始参数分别设置为 128×128 像素、16 像素和 0。此外,在 MATLAB 平台上测试了 CPU 方法^[3]、基于层次的方法[27, 29-30],并在 Visual Studio 环境中比较了方法[28]和 GPU 方法^[8]。因此,假设上述方法具有最合适的

初始条件和环境。值得注意的是,虽然本文中的视差 3D 网格模型不是真正的数字高程模型(Digital Elevation Model, DEM),但这不影响系统性能的分析 and 比较。具体三维重建结果如图 8 所示,其中矩形和椭圆分别代表重建结果中高山区域和河谷区域。实验结果表明,方法[28]、方法[29]及方法[30]的矩形区域和椭圆区域重建结果严重失真,即受高山区域和河谷区域影响很大,而方法[27]与 Agisoft stereoscan 商业软件的矩形区域成功恢复出高山地形特征,但是在椭圆区域却无法重建出河流区域的场景结构,只有 CPU 方法^[3]、GPU 方法^[8]与本文方法同时恢复出高山区域和河流区域的场景结构。

因此,本文方法与 CPU 方法^[3]、GPU 方法^[8]的重建结果相近,并显著优于现有方法[27-30]及 Agisoft stereoscan 商业软件的重建结果,表现出很强的鲁棒性。除此之外,为了充分证明本文方法的优越性,进一步提供了 2 组无人机图像的测试结果,分别如图 9 和图 10 所示。在图 9 中,矩形和椭圆分别代表重建结果中河流区域和高山区域。实验结果表明,方法[28]和方法[30]的矩形和椭圆区域重建结果受高山区域和河流区域影响严重,而 CPU 方法^[3]、GPU 方法^[8]及本文方法可最大程度地减少高山和河流的影响。另外,在图 10 中,椭圆代表重建结果中山脉地区。实验结果表明,方法[28]和方法[30]的椭圆区域重建结

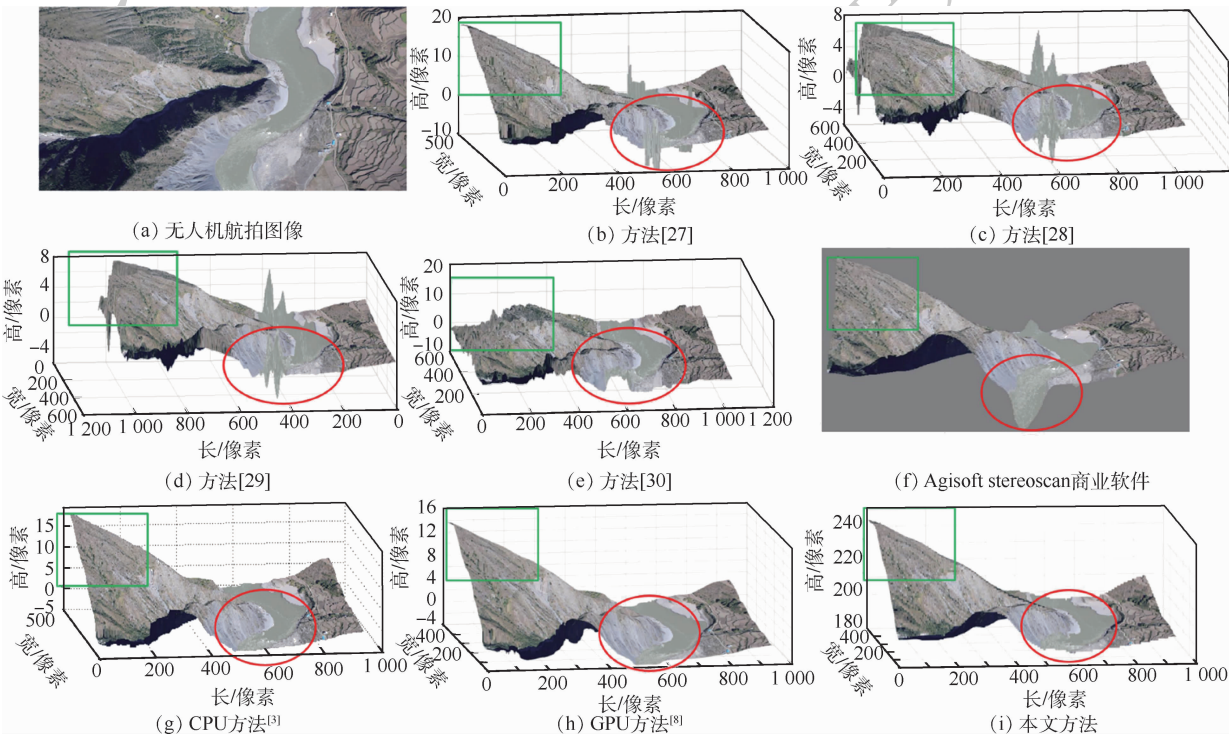


图 8 与先进方法重建结果比较

Fig. 8 Comparison with reconstruction results of state-of-the-art methods

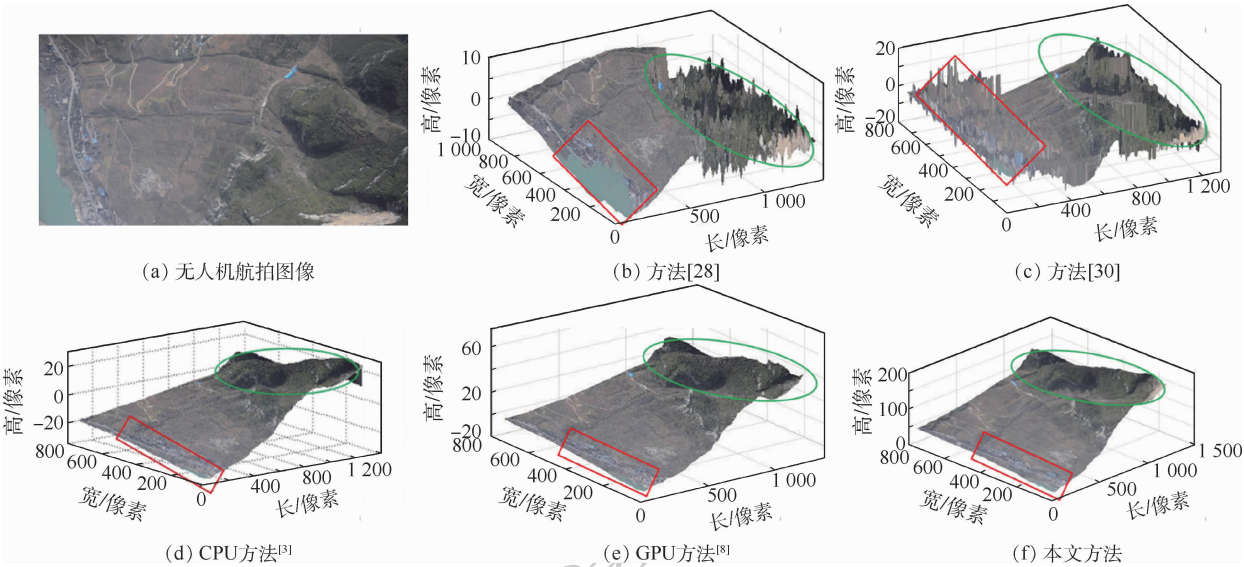


图9 高山、河流区域无人机图像的测试结果

Fig. 9 Test results of UAV image of mountain and river areas

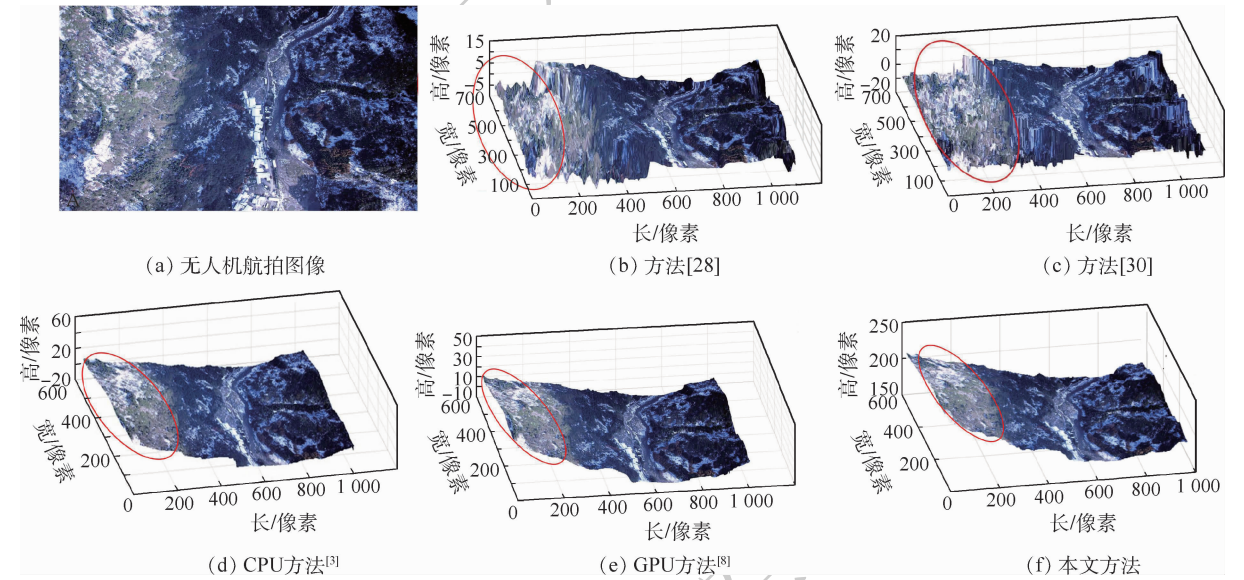


图10 山区无人机图像的测试结果

Fig. 10 Test results of UAV images of mountain area

果受山脉地区影响很大,几乎不能生成有效的重建结果。而CPU方法^[3]、GPU方法^[8]与本文方法可充分提取高山无人机图像的三维场景结构。因此,本文模型对高山、河流区域及山脉地区的无人机航拍影像具有很强的鲁棒性,并高度适用于多种复杂环境的无人机航拍图像。此外,本文方法在功耗及时效上具有更优异的表现。为了对本文方法的性能进行综合评估,将在3.3.2节从功耗及时效性方面,基于CPU方法^[3]及基于GPU方法^[8]进行进一步对比分析。

3.3.2 功耗和时效性

对本文方法与基于CPU^[3]、GPU^[8]平台的三维重建方法,在执行时间、功率损耗两方面进行详

细比较。实验结果如图11所示。实验数据显示,基于有限资源FPGA方法,在时间消耗上达到与

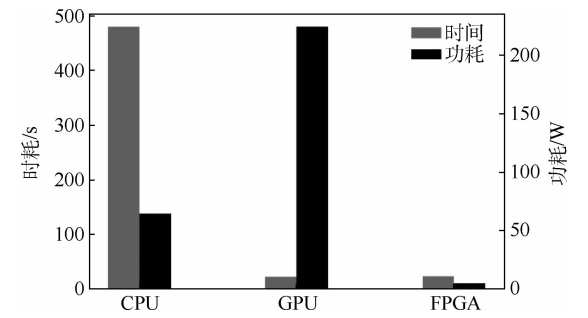


图11 基于CPU、GPU、FPGA方法的功耗和时效性比较

Fig. 11 Comparison on power consumption and time consumption based on CPU, GPU and FPGA methods

基于 GPU 方法相近的结果,并且与基于 CPU 平台方法相比,其时效性提高了近 20 倍;另外,在功耗方面,基于 FPGA 方法的功耗远低于 GPU 方法,仅为基于 GPU 方法的 2.12%,与基于 CPU 方法相比,基于 FPGA 方法是基于 CPU 方法的 7.42%。综上所述,本文方法成功使用有限资源 FPGA 芯片,实现了无人机航拍影像快速低功耗三维重建。为了进一步展示本文模型的高精度特性,将在 3.3.3 节从模型精度方面进行进一步分析。

3.3.3 精 度

为了充分展示本文模型的高精度特性,基于一组人工合成图像对 FPGA 方法和 CPU^[3]、GPU^[8]平台的三维重建方法,在模型精度方面进行详细比较。首先,设计了一张尺寸为 512 × 512

像素的脉冲噪声图像(噪声密度为 0.5)。然后,为了模拟精度为 1.5 的亚像素平移,将上述图像中央的 100 × 100 像素正方形区域在 x 轴上平移 1.5 个像素之后的结果作为目标图像。最后,分别使用本文方法与 CPU^[3]、GPU^[8]方法对上述立体图像对进行三维重建。值得注意的是,由于 FPGA 平台生成的视差结果以 BMP 位图图片格式存储,当利用 MATLAB 平台读取视差结果并进行三维可视化显示时,视差结果会自动转换为 0 ~ 255 范围的灰度值大小,需将其按比例缩放到相应尺寸显示。重建结果如图 12 所示。根据图 12 可视化结果和表 3 定量评估结果显示,本文方法成功基于有限资源 FPGA 芯片生成接近于 CPU 和 GPU 方法精度的三维重建结果。因此,本文方法具有亚像素级视差精度的三维重构能力。

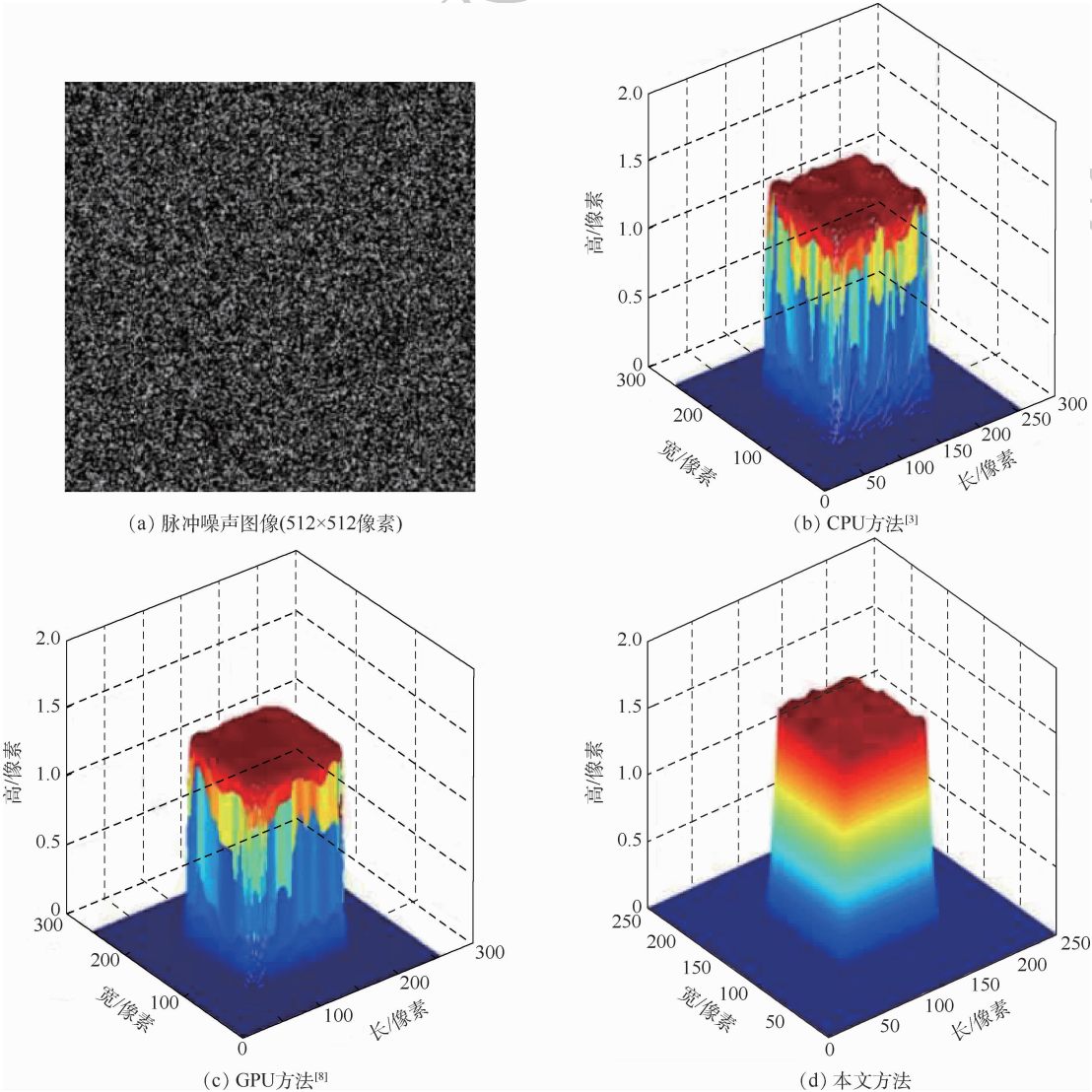


图 12 基于合成图像的三维重建结果可视化
Fig. 12 Visualization of 3D reconstruction results based on synthetic images

表 3 基于 CPU、GPU、FPGA 平台三维重建
方法的定量评估结果

Table 3 Quantitative evaluation of 3D reconstruction
results based on CPU, GPU and FPGA methods

方法	平均误差	均方根误差
CPU	1.5110	0.0149
GPU	1.5020	0.0123
FPGA	1.5195	0.0131

4 结 论

本文结合高并行指令优化策略,提出了一种基于 FPGA 软硬件协同优化方案的快速低功耗高精度三维重建方法。

1) 本文方法实现了对无人机航拍影像中复杂低纹理区域、动态纹理区域及大视差范围区域的低功耗快速高精度的视差信息求解。

2) 本文方法能够在单个 FPGA 上实现包括预处理和后处理阶段等完整三维重建方法,如无人机航拍图像序列读取、子图像提取、视差计算、亚像素级视差拟合及重建结果保存等。

3) 为了对本文方法进行全面评估,基于 CPU、GPU 及 FPGA 平台的方法进行了对比研究,实验结果表明,本文方法在执行速度方面可与 GPU 方法相媲美,更在功耗方面明显优于 GPU 方法。

在未来的工作中,将尝试在硬件架构(如多 FPGA 系统)或算法等方面,对三维重建方法的质量和重建速度进行进一步的优化,或考虑在立体视觉的其他应用领域展开探索。

参考文献 (References)

[1] 韩磊,徐梦溪,王鑫,等. 基于光场成像的多线索融合深度估计方法[J]. 计算机学报,2020,43(1):107-122.
HAN L,XU M X,WANG X,et al. Depth estimation from multiple cues based light-field cameras[J]. Chinese Journal of Computers,2020,43(1):107-122(in Chinese).

[2] 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]. 计算机学报,2020,43(1):41-63.
LI H S,WU Y J,ZHENG Y P,et al. A survey of 3D data analysis and understanding based on deep learning[J]. Chinese Journal of Computers,2020,43(1):41-63(in Chinese).

[3] LI J,LIU Y,DU S,et al. Hierarchical and adaptive phase correlation for precise disparity estimation of UAV images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2016, 54(12):7092-7104.

[4] TREVOR G,PATRICE D,SYLVAIN J,et al. A dedicated light-weight binocular stereo system for real-time depth-map generation[C]//Proceedings of the 2016 Conference on Design and Architectures for Signal and Image Processing (DASIP). Piscataway:IEEE Press,2016:215-221.

[5] LENTARIS G,MARAGOS K,SOUDRIS D,et al. Single- and multi-FPGA acceleration of dense stereo vision for planetary rovers[J]. ACM Transactions on Embedded Computing Systems,2019,18(2):1-27.

[6] BHATIA T,LAHA J,MAJI A,et al. Design of FPGA based horizontal velocity computation and image storage unit for lunar lander[C]//Proceedings of the 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON). Piscataway: IEEE Press, 2017: 521-526.

[7] LENTARIS G,STRATAKOS I,STAMOULIAS I,et al. High-performance vision-based navigation on SoC FPGA for spacecraft proximity operations[J]. IEEE Transactions on Circuits and Systems for Video Technology,2020,30(4):1188-1202.

[8] LI J,LIU Y. High precision and fast disparity estimation via parallel phase correlation hierarchical framework[J/OL]. Journal of Real-Time Image Processing, (2020-05-02) [2020-08-13]. <https://doi.org/10.1007/s11554-020-00972-1>.

[9] ZHENG J,ZHAI B,WANG Y,et al. Motion estimation using maximum sub-image and sub-pixel phase correlation on a DSP platform[J]. Multimedia Tools and Applications,2019,78(14):19019-19043.

[10] KUHN M,MOSER S,ISLER O,et al. Efficient ASIC implementation of a real-time depth mapping stereo vision system[C]//Proceedings of the 2003 46th Midwest Symposium on Circuits and Systems. Piscataway:IEEE Press,2003:1478-1481.

[11] MASRANI D K,MACLEAN W J. A real-time large disparity range stereo-system using FPGAs[C]//Proceedings of the 4th IEEE International Conference on Computer Vision Systems (ICVS'06). Piscataway:IEEE Press,2006:42-51.

[12] BATTIATO S,BRUNA A R,PUGLISI G,et al. A robust block-based image/video registration approach for mobile imaging devices[J]. IEEE Transactions on Multimedia,2010,12(7):622-635.

[13] MONTEIRO E,VIZZOTTO B B,DINIZ C U,et al. Applying CUDA architecture to accelerate full search block matching algorithm for high performance motion estimation in video encoding[C]//Proceedings of the 2011 23rd International Symposium on Computer Architecture and High Performance Computing. Piscataway:IEEE Press,2011:128-135.

[14] KIM S W,YIN S,YUN K,et al. Spatio-temporal weighting in local patches for direct estimation of camera motion in video stabilization[J]. Computer Vision and Image Understanding, 2014,118:71-83.

[15] WU X,ZHAO Q,BU W,et al. A SIFT-based contactless palmprint verification approach using iterative RANSAC and local palmprint descriptors[J]. Pattern Recognition,2014,47(10):3314-3326.

[16] CHEN J,XING M,SUN G,et al. A 2-D space-variant motion estimation and compensation method for ultrahigh-resolution airborne stepped-frequency SAR with long integration time[J]. IEEE Transactions on Geoscience and Remote Sensing,2017,55(11):6390-6401.

[17] ISMAIL Y,MCNEELY J,SHAABAN M,et al. A fast discrete transform architecture for frequency domain motion estimation

- [C] // Proceedings of the IEEE International Conference on Image Processing. Piscataway: IEEE Press, 2010: 1249-1252.
- [18] ALBA A, VIGUERAS-GOMEZ J F, ARCE-SANTANA E R, et al. Phase correlation with sub-pixel accuracy: A comparative study in 1D and 2D [J]. Computer Vision and Image Understanding, 2015, 137: 76-87.
- [19] DONG Y, LONG T, JIAO W, et al. A novel image registration method based on phase correlation using low-rank matrix factorization with mixture of gaussian [J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(1): 446-460.
- [20] KUMAR S, AZARTASH H, BISWAS M, et al. Real-time affine global motion estimation using phase correlation and its application for digital image stabilization [J]. IEEE Transactions on Image Processing, 2011, 20(12): 3406-3418.
- [21] TZIMIROPOULOS G, ARGYRIOU V, ZAFEIRIOU S, et al. Robust FFT-based scale-invariant image registration with image gradients [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(10): 1899-1906.
- [22] MATTHIES L, MAIMONE M, JOHNSON A E, et al. Computer vision on mars [J]. International Journal of Computer Vision, 2007, 75(1): 67-92.
- [23] DARABIHA A, ROSE J, MACLEAN J W, et al. Video-rate stereo depth measurement on programmable hardware [C] // Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2003: 203-210.
- [24] BHATIA T, MAJI A, LAHA J, et al. Practical approach for implementation of phase correlation based image registration in FPGA [C] // Proceedings of the 2018 International Conference on Current Trends Towards Converging Technologies (IC-CTCT). Piscataway: IEEE Press, 2018: 1-4.
- [25] MATSUO K, HAMADA T, MIYOSHI M, et al. Accelerating phase correlation functions using GPU and FPGA [C] // Proceedings of the 2009 NASA/ESA Conference on Adaptive Hardware and Systems. Piscataway: IEEE Press, 2009: 433-438.
- [26] LIU J G, HONG S Y. Robust phase correlation methods for sub-pixel feature matching [C] // Proceeding of the 1st Conference of Systems Engineering for Autonomous Systems. Piscataway: IEEE Press, 2012: 1-10.
- [27] TAKITA K, MUQUIT M A, AOKI T, et al. A sub-pixel correspondence search technique for computer vision applications [J]. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 2004, 87(8): 1913-1923.
- [28] CHEN T, LIU Y, LI J, et al. Fast narrow-baseline stereo matching using CUDA compatible GPUs [C] // Proceeding of the Advances in Image and Graphics Technologies (IGTA 2015). Piscataway: IEEE Press, 2015: 10-17.
- [29] MORGAN G L, LIU J, YAN H, et al. Precise subpixel disparity measurement from very narrow baseline stereo [J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(9): 3424-3433.
- [30] 刘怡光, 赵晨晖, 黄蓉刚, 等. 勿需图像矫正的高精度窄基线三维重建算法 [J]. 电子科技大学学报, 2014, 43(2): 262-267.
- LIU Y G, ZHAO C H, HUANG R G, et al. Rectification-free 3-dimensional reconstruction method based on phase correlation for narrow baseline image pairs [J]. Journal of University of Electronic Science and Technology of China, 2014, 43(2): 262-267 (in Chinese).

作者简介:

李杰 男, 博士, 副教授。主要研究方向: 机器学习、计算机视觉和图像处理。

李一轩 男, 硕士研究生。主要研究方向: 机器学习、计算机视觉和高能效计算。

吴天生 男, 硕士研究生。主要研究方向: 机器视觉、数字图像处理 and 三维重建。

王昊榕 女, 硕士研究生。主要研究方向: 机器视觉、数字图像处理 and 三维超分辨率重建。

梁敏 女, 博士, 副教授。主要研究方向: 大数据技术和应用。

Fast, low-power and high-precision 3D reconstruction of UAV images based on FPGA

LI Jie^{*}, LI Yixuan, WU Tiansheng, WANG Haorong, LIANG Min

(College of Information, Shanxi University of Finance and Economics, Taiyuan 030006, China)

Abstract: The existing 3D reconstruction methods based on Unmanned Aerial Vehicle (UAV) images cannot meet the mobile terminal's demand for low power consumption and high time efficiency. To tackle this issue, we propose a fast, low-power and high-precision 3D reconstruction method based on resource-constrained FPGA platforms, which combines instruction optimization strategy and hardware-software co-design method. First, we construct a multi-scale depth map fusion algorithm architecture to enhance the robustness of traditional FPGA phase correlation algorithms to untrustworthy areas, such as low-texture area and rivers. Secondly, based on the high parallel instruction optimization hardware acceleration function strategy, a high-performance hardware-software co-design scheme is proposed to realize the efficient operation of the multi-scale deep map fusion algorithm architecture on the FPGA platform with limited resources. Finally, we comprehensively compare the state-of-the-art CPU and GPU methods with our method. The experimental results show that our method is close to the GPU method in reconstruction time consumption, nearly 20 times faster than the CPU method, but the power consumption is only 2.23% of the GPU method.

Keywords: low-power; FPGA; 3D reconstruction; phase correlation; hardware-software co-design

Received: 2020-08-24; **Accepted:** 2020-09-05; **Published online:** 2020-09-28 16:51

URL: kns.cnki.net/kcms/detail/11.2625.V.20200928.1616.005.html

Foundation items: National Natural Science Foundation of China (61801279); The Applied Basic Research Programs of Shanxi Province (201801D221160); Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (STIP) (2019L0471); Shanxi Graduate Education Innovation Project (2020SY175, 2020SY176)

*** Corresponding author.** E-mail: lijiescu@aliyun.com

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

DOI: 10.13700/j.bh.1001-5965.2020.0475

融合句嵌入的 VAACGAN 多对多语音转换



李燕萍^{1,*}, 曹盼¹, 石杨¹, 张燕²

1. 南京邮电大学 通信与信息工程学院, 南京 210003; 2. 金陵科技学院 软件工程学院, 南京 211169)

摘

要: 针对非平行文本条件下语音转换质量不理想、说话人个性相似度不高的问题,提出一种融合句嵌入的变分自编码辅助分类器生成对抗网络(VAACGAN)语音转换方法,

在非平行文本条件下,有效实现了高质量的多对多语音转换。辅助分类器生成对抗网络的鉴别器中包含辅助解码器网络,能够在预测频谱特征真假的同时输出训练数据所属的说话人类别,使得生成对抗网络的训练更为稳定且加快其收敛速度。通过训练文本编码器获得句嵌入,将其作为一种语义内容约束融合到模型中,利用句嵌入包含的语义信息增强隐变量表征语音内容的能力,解决隐变量存在的过度正则化效应的问题,有效改善语音合成质量。实验结果表明:所提方法的转换语音平均 MCD 值较基准模型降低 6.67%,平均 MOS 值提升 8.33%,平均 ABX 值提升 11.56%,证明该方法在语音音质和说话人个性相似度方面均有显著提升,实现了高质量的语音转换。

关

键 词: 语音转换; 句嵌入; 文本编码器; 辅助分类器生成对抗网络(ACGAN); 变

分自编码器; 非平行文本; 多对多

中图分类号: TN912.3

文献标志码: A

文章编号: 1001-5965(2021)03-0500-09

语音转换是一种改变源说话人的声音,使之听起来像目标说话人声音的技术,但其中的语义信息保持不变^[1-2]。语音转换在构音障碍患者语音修复、影视配音、个性文语合成、增大语音数据和歌唱语音合成等方面有广阔的市场应用前景^[3]。

非平行文本条件下的语音转换研究取得了很大进展,主要经历了3个阶段:①借助自动语音识别(Automatic Speech Recognition, ASR)和文语转换(Text-To-Speech, TTS)等系统构建伪平行文本^[4-5],从而用平行转换方法完成转换;②采用说话人自适应^[6-7]和说话人归一化等方法对已有的平行语音转换模型实现更新,从而完成非平行文本转换;③当前的主流方法是通过解纠缠(Disen-

tangle)语义和说话人个性特征^[8-10],在转换过程中对语义信息和目标说话人个性信息进行重构。

基于C-VAE模型的语音转换方法^[8],其编码器解纠缠语音的语义和说话人的个性信息,解码器通过对语义和指定说话人身份标签进行重构完成语音的转换,不需要任何对齐操作,实现从平行文本向非平行文本的跨越,并且通过标签的引入实现了多对多的语音转换。由于C-VAE模型基于理想假设,认为观察到的数据服从高斯分布,导致解码器的输出语音过度平滑,转换后的语音质量不高。Hsu等^[9]提出的变分自编码生成对抗网络(Variational Autoencoding Wasserstein Generative Adversarial Network, VAWGAN)模型通过在C-VAE中引入Wasserstein生成对抗网络(Wasserstein

收稿日期: 2020-08-31; 录用日期: 2020-09-04; 网络出版时间: 2020-11-17 15:31

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201117.1417.002.html

基金项目: 国家自然科学基金(61401227, 61872199, 61872424); 金陵科技学院智能人机交互科技创新团队建设专项(218/010119200113)

* 通信作者: E-mail: liyp@njupt.edu.cn

引用格式: 李燕萍, 曹盼, 石杨, 等. 融合句嵌入的 VAACGAN 多对多语音转换[J]. 北京航空航天大学学报, 2021, 47(3): 500-508. LI Y P, CAO P, SHI Y, et al. Many-to-many voice conversion with sentence embedding based on VAACGAN[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 500-508 (in Chinese).

GAN, WGAN)^[11], 将 VAE 的解码器指定为 WGAN 的生成器来优化目标函数, 一定程度上提升了语音的质量。但是, 上述 2 种方法都基于优化相同的概率图模型, 而隐变量存在的过度正则化问题会直接导致转换语音模糊不清。此外, WGAN 在模型训练过程中存在性能不稳定及收敛速度较慢等问题, 导致转换后合成语音质量不高, 因此, 基于 VAWGAN 的语音转换系统的性能仍有待提升。

针对上述问题, Saito 等^[12]于 2018 年提出利用预训练的 ASR 模型得到音素后验概率 (Phonetic Posteriorgrams, PPG) 来增强隐变量表征语义内容的能力, 以进一步提高基于 C-VAE 模型的性能, 然而该方法直接依赖于经过大量语料和长时间训练得到的 ASR 系统的性能, 准确率也不稳定。鉴于此, 本文提出融合句嵌入 (Sentence Embedding, SE) 的辅助分类器生成对抗网络 (Auxiliary Classifier GAN, ACGAN) 语音转换方法。先构建基于变分自编码辅助分类器生成对抗网络 (Variational Autoencoding Auxiliary Classifier GAN, VAACGAN) 的语音转换模型, 通过引入 ACGAN 取代基准模型中的 WGAN, 利用其鉴别器同时鉴别输入频谱特征的真假及所属的说话人类别, 有效提升生成频谱的质量。在此基础上, 进一步提出通过预训练的文本编码器将说话人语句编码为句嵌入, 将其融合到 VAACGAN 模型中, 作为与说话人无关的语义表征直接参与模型训练, 能够显著增强隐变量表征语义信息的能力。

1 基于 VAWGAN 的语音转换基准方法

VAWGAN 模型由编码器、生成器和鉴别器 3 部分构成, 图 1 为基于 VAWGAN 模型的频谱转换原理图。给定源说话人和目标说话人的频谱特征帧分别为: $X_s = \{x_{s,n}, n = 1, 2, \dots, N_s\}$, $X_t = \{x_{t,n}, n = 1, 2, \dots, N_t\}$, N_s 为源说话人的帧总数, N_t 为目标说话人的帧总数, s 和 t 分别表示源和目标说话人身份的索引。在 VAWGAN 模型中, 通过编码器将说话人的频谱特征编码为隐变量 z_n , 认为是独立于说话人的语义信息。在此编码过程中, 由于可以不考虑输入频谱特征帧的说话人身份信息, 可将输入 $x_{s,n}$ 、 $x_{t,n}$ 统一表示为 x_n 。在可读性不受影响的前提下, 可忽略频谱特征帧索引。在上述过程中, x 表示输入频谱特征, y 表示说话人身份标签, z 表示语义表征隐变量。基于 VAWGAN 模型的频谱特征转换可以表示为

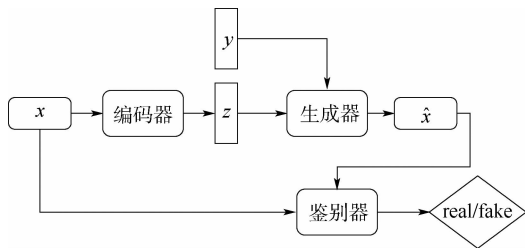


图 1 基于 VAWGAN 模型的频谱转换原理图

Fig. 1 Schematic diagram of spectrum conversion based on VAWGAN model

$$\hat{x} = \hat{f}(x, y) = f_{\theta}(z, y) = f_{\theta}(f_{\phi}(x), y) \quad (1)$$

式中: $\hat{f}(x, y)$ 表示通过输入的频谱特征 x 和目标说话人的标签 y 得到转换语音; $f_{\phi}(\cdot)$ 表示编码器, 说话人身份标签 y 由传统说话人编码向量 one-hot 表示, 以区分不同说话人的身份信息; 生成器 $f_{\theta}(\cdot)$ 通过结合隐变量 z 和说话人身份标签 y 构成联合特征 (z, y) , 以重构目标说话人的频谱特征帧 $\hat{x}$, 再将真实频谱特征 x 和生成器重构的频谱特征 $\hat{x}$ 一同输入鉴别器, 实现输入频谱真假 (real/fake) 的鉴别, 以生成接近目标说话人的频谱特征。

由上述可知, VAWGAN 模型通过在 C-VAE 模型中引入 WGAN, 将 C-VAE 的解码器指定为 WGAN 的生成器来优化目标函数。因此, 该模型完整的目标函数表示为

$$J_{\text{VAWGAN}} = L(x; \phi, \theta) + \alpha J_{\text{WGAN}} \quad (2)$$

式中: $L(x; \phi, \theta)$ 表示 C-VAE 模块的目标函数; α 为调节 WGAN 损失的系数; J_{WGAN} 表示 WGAN 模块的目标函数, 即

$$J_{\text{WGAN}} = E_{x \sim p_r} [D_{\psi}(x)] - E_{z \sim q_{\phi}(z|x)} [D_{\psi}(G_{\theta}(z, y))] \quad (3)$$

其中: $E_{x \sim p_r}[\cdot]$ 表示鉴别器鉴别输入为真实频谱特征分布的概率期望; $D_{\psi}(x)$ 表示鉴别器对输入真实频谱特征 x 的鉴别; $G_{\theta}(z, y)$ 为生成频谱特征; $E_{z \sim q_{\phi}(z|x)}[\cdot]$ 表示鉴别器鉴别输入为生成频谱特征分布的概率期望。

式 (2) 中, $L(x; \phi, \theta)$ 表达式为

$$L(x; \phi, \theta) = -D_{\text{KL}}(q_{\phi}(z|x) \| p_{\theta}(z)) + E_{q_{\phi}(z|x)} [\ln p_{\theta}(z|x, y)] \quad (4)$$

式中: $D_{\text{KL}}(\cdot \| \cdot)$ 表示 KL 散度 (Kullback-Leibler Divergence, KLD); $E_{q_{\phi}(z|x)}[\cdot]$ 表示解码过程中的重构损失; $p_{\theta}(z)$ 为隐变量 z 的先验分布, 服从标准多维高斯分布; $q_{\phi}(z|x)$ 表示编码过程; $p_{\theta}(z|x, y)$ 表示解码过程; ϕ 和 θ 分别表示编码器和生成器参数, 通常使用随机梯度下降法来更新 C-VAE 模型中的网络参数。

在式(4)中,KL 项先验的影响很大,假定观测数据严格服从标准高斯分布^[13],该假设将导致模型学习到的目标函数存在缺陷,获得的隐变量分布较为简单,难以完整代表语义信息的基础结构。

综上分析可知,VAWGAN 模型基于解纠缠语义和说话人个性信息的思想,完全解除了对平行训练数据的依赖。此外,通过在解码阶段指定说话人标签,能够实现多对多的语音转换,引入 WGAN 一定程度上提升了 C-VAE 模型的性能,但是仍然存在两大问题:①为了使训练更为稳定,基准模型中的 WGAN 采用权重剪切 (Weight Clipping) 操作,将权重剪切到一定范围内,但是仍然存在训练困难、收敛较慢等问题^[3];②由于 VAWGAN 与 C-VAE 模型都基于优化相同的概率图模型,在隐变量中经常观察到过度正则化效应,易导致语义内容趋于消失,转换后的语音模糊不清。基于此,本文将从改进生成对抗网络的性能及增加语义隐变量约束两方面对基准模型提出改进。

2 融合句嵌入的 VAACGAN 语音转换

本文提出融合句嵌入的变分自编码辅助分类器生成对抗网络 (VAACGAN-SE) 语音转换模型,创新点在于:①将 ACGAN 模型应用到语音转换领域中^[3,13],其鉴别器能够同时鉴别输入频谱包络特征的真假及其所属的说话人类别,以提升语音转换模型生成语音的清晰度和自然度;②从增加语义隐变量约束的角度,通过预训练一个文本编码器,将说话人语句编码为句嵌入,作为语义内容的有效辅助,再将其融入 VAACGAN-SE 模型的 C-VAE 模块中参与训练,以增强隐变量表征与说话人无关的语义内容的能力,使得源语音的语义内容在训练和转换阶段都能得到有效保持,有效克服隐变量的过度正则化效应问题,进一步改善模型的性能。

2.1 ACGAN 原理

ACGAN 是 Odena 等^[14]于 2017 年提出用于图像合成的一种方法,与传统 WGAN 不同之处在于:ACGAN 的鉴别器中包含能够输出训练数据的类别标签的辅助解码器网络^[3,13]。本文方法采用 ACGAN 取代基准模型中的 WGAN,通过联合训练编码器和生成器,有效区分不同说话人的频谱特征,生成的频谱包络特征更加接近目标说话人,从而提升语音质量,且训练过程更加稳定。

ACGAN 原理如图 2 所示。ACGAN 中,除了输入频谱特征外,每一个生成的频谱特征都具有

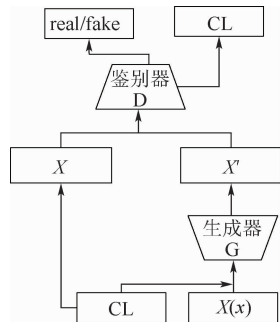


图 2 ACGAN 原理示意图

Fig.2 Schematic diagram of ACGAN

对应的类别标签 $cl \sim p_{cl}$, cl 利用传统说话人编码 one-hot 表示,在该方法中表示说话人身份标签 y 。生成器 G 使用频谱特征 x 和说话人类别标签 cl 来生成样本 $X' = G(x, cl)$,鉴别器 D 的输入为真实频谱特征 X 和生成样本 X' ,输出为所属的说话人类别 CL 及其预测的频谱特征的“真假” S ^[3]。

ACGAN 鉴别器的目标函数由 2 部分构成,如下:

$$L_s = E[\ln P(S = \text{real} | X)] + E[\ln P(S = \text{fake} | X')] \quad (5)$$

$$L_{cl} = E[\ln P(CL = cl | X)] + E[\ln P(CL = cl | X')] \quad (6)$$

式中: L_s 和 L_{cl} 分别为正确来源的对数似然和正确类的对数似然; $P(S = \text{real} | X)$ 表示鉴别器 D 判定频谱特征为真的概率; $P(CL = cl | X)$ 表示鉴别器 D 判定频谱特征属于说话人标签 cl 的概率^[3,13]。在模型的训练过程中,鉴别器 D 的训练目标是最大化 $L_s + L_{cl}$,同时训练生成器 G 是最大化 $L_s - L_{cl}$ 。

2.2 句嵌入的融合与训练

通过增强隐变量表征语义的能力,以进一步提升转换语音的质量是本文在上述研究基础上的深入探索。创新点在于:在 C-VAE 模块训练过程中融合了句嵌入,其代表与说话人无关的语义内容的隐变量,通过预训练文本编码器 (Text-Encoder) 来获得。因此,C-VAE 模块的目标函数由第 1 节中的式(4)更新为

$$L(x; \phi, \theta) = -D_{KL}(q_\phi(z|x) \| p_\theta(z)) + E_{q_\phi(z|x)}[\ln p_\theta(x|z, z_p, y)] \quad (7)$$

式中: z_p 表示句嵌入。解码器网络 $p_\theta(x|z, z_p, y)$ 中融入了句嵌入参与模型训练,增强了 C-VAE 模块提取与说话人无关的语义隐变量的能力,使得源语音中的语义信息得到充分的表征和重建。如图 3 所示,与图 3(a) 传统 C-VAE 模型不同,图 3(b) 为本文融合句嵌入的 C-VAE 模型原理示意图^[13],图中实线表示频谱特征参数 x 生成隐变量 z 和句嵌入 z_p ,虚线表示通过联合矢量 (z, z_p, y)

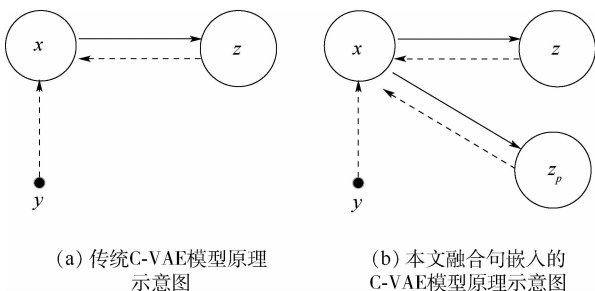


图 3 传统 C-VAE 模型与本文融合句嵌入的 C-VAE 模型对比

Fig. 3 Comparison of conventional C-VAE model and proposed model based on C-VAE with sentence embedding
来重构频谱特征参数 x 。

本文选取 VCC2018 语料库^[15]进行语音转换模型的训练和测试,该语料库是由国际行业内挑战赛提供的标准数据库,其中训练语料包含语句对应的转录,即句子的语义内容。因此,本文方法利用数据库所提供的转录,通过 Arora 等^[16]提出的由词嵌入获得训练语料的准确句嵌入算法。然而在转换阶段,由于任意的测试语料没有转录,需要提供给转换阶段一个预训练好的文本编码器,以生成测试语料对应的句嵌入。

2.2.1 训练语料的句嵌入提取

在训练阶段,通过 Arora 等^[16]提出的句嵌入提取算法,先将训练语料所对应转录中每个单词表示为词嵌入^[13],再将句嵌入表示成词嵌入的加权平均值:

$$z_u \leftarrow \frac{1}{|S|} \sum_{n \in S} \frac{a}{a + p(n)} v_n \quad (8)$$

式中: S 表示单词所处的句子,其绝对值表示句子中包含的单词数量; v_n 为单词 n 对应的词嵌入; z_u 为获得的句嵌入; a 为超参数; $p(n)$ 为单词 n 在整个语料库中出现的概率^[3]。将 z_u 形成矩阵的第一个奇异值记为 u ,通过 z_u 减去在向量 u 上的投影^[13],即可得到编码之后的句嵌入表示:

$$z_u \leftarrow z_u - uu^T z_u \quad (9)$$

2.2.2 文本编码器的训练及转换语料的句嵌入提取

基于 VAACGAN-SE 模型的频谱转换训练过程如图 4 所示,VAACGAN-SE 增加了能够生成句嵌入的文本编码器的预训练。

通过训练文本编码器来构建句嵌入和频谱包络特征之间的映射关系,如图 4 所示,将频谱包络特征输入文本编码器中可以获得相应语句的句嵌入。文本编码器具体训练过程如下^[13]:首先,将语料库提供的训练语句的转录编码为句嵌入 z_p ;然后,利用 {频谱包络特征 x ,句嵌入 z_p } 对训练文本编码器,经过训练输出预测得到句嵌入 $\hat{z}_p$, z_p 和 $\hat{z}_p$ 之间的误差通过 L_2 损失函数度量:

$$L^2(\hat{z}_p, z_p) = \sum_{m=0}^M (\hat{z}_p^{(m)} - z_p^{(m)})^2 \quad (10)$$

式中: M 为训练语句的数量。文本编码器与其他网络联合训练,迭代优化使文本编码器的损失函数尽可能小,直至实验设置的最大迭代次数。在转换阶段,将转换语料的频谱包络特征输入预训练好的文本编码器,就可以提取句嵌入。

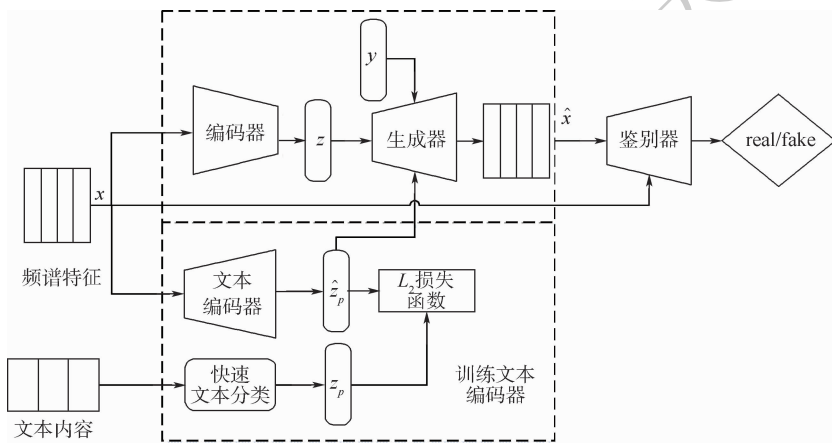


图 4 基于 VAACGAN-SE 模型的频谱转换的训练过程

Fig. 4 Training process of spectrum conversion based on VAACGAN-SE model

2.3 融合句嵌入的 VAACGAN 语音转换系统

2.3.1 训练阶段

采用 WORLD^[17]声码器,从说话人的训练语句中提取 3 部分特征,包括频谱包络特征 x 、基频 f_0 及非周期特征 (Aperiodicities, APs)^[18]。其中,频谱包络特征 x 输入 VAACGAN-SE 模型进行训

练,具体训练过程如图 4 所示,VAACGAN-SE 模型的输入是频谱包络特征 x ,充分训练后得到认为独立于说话人的语义信息 z ;同时将频谱包络特征 x 及由训练语料转录生成的准确句嵌入 z_p 作为文本编码器的输入,实现文本编码器的预训练。生成器将获得的隐变量 z 、句嵌入 z_p 和表示

说话人身份的 one-hot 向量相结合,重构特定说话人的频谱包络特征 $\hat{x}$ 。

VAACGAN-SE 模型的目标函数为

$$J_{\text{VAACGAN-SE}} = L(x; \phi, \theta) + \alpha J_{\text{ACGAN}} \quad (11)$$

式中: $L(x; \phi, \theta)$ 为 C-VAE 模块中融合了句嵌入 z_p 的目标函数,具体表达式如式(7)所示; α 为 AC-GAN 的损失系数; J_{ACGAN} 为 ACGAN 部分损失函数。

ACGAN 的生成器中结合了句嵌入 z_p :

$$\begin{aligned} J_{\text{ACGAN}} = & E_{x \sim p_{\theta}} [D_{\psi}(x)] - \\ & E_{z \sim q_{\phi}(z|x)} [D_{\psi}(G_{\theta}(z, z_p, y))] + \\ & E_{x \sim p_{\theta}} [\text{loss}(D_{\psi_{\text{cl}}}(x), y)] + \\ & E_{z \sim q_{\phi}(z|x)} [\text{loss}(D_{\psi_{\text{cl}}}(G_{\theta}(z, z_p, y)), y)] \quad (12) \end{aligned}$$

式中: $D_{\psi}(\cdot)$ 和 $D_{\psi_{\text{cl}}}(\cdot)$ 分别为鉴别器对频谱特征真假和所属类别的判定输出; $\text{loss}(\cdot)$ 函数表示鉴别器预测的频谱特征类别和真实类别之间的交叉熵损失。

频谱特征类别损失表示如下:

$$\begin{aligned} L_{\text{cl}} = & E_{x \sim p_{\theta}} [\text{loss}(D_{\psi_{\text{cl}}}(x), y)] + \\ & E_{z \sim q_{\phi}(z|x)} [\text{loss}(D_{\psi_{\text{cl}}}(G_{\theta}(z, z_p, y)), y)] \quad (13) \end{aligned}$$

在 ACGAN 中,生成器的损失函数为

$$\begin{aligned} L_G = & -E_{q_{\phi}(z|x)} [\ln p_{\theta}(x|z, z_p, y)] - \\ & \beta E_{z \sim q_{\phi}(z|x)} [D_{\psi}(G_{\theta}(z, z_p, y))] + L_{\text{cl}} \quad (14) \end{aligned}$$

式中: β 为鉴别器鉴别输入为生成频谱特征分布的概率期望系数;优化目标为 $\min_{\theta} \{L_G\}$ 。

在 ACGAN 中,鉴别器的损失函数为

$$\begin{aligned} L_D = & E_{x \sim p_{\theta}} [D_{\psi}(x)] - \\ & E_{z \sim q_{\phi}(z|x)} [D_{\psi}(G_{\theta}(z, z_p, y))] + L_{\text{cl}} \quad (15) \end{aligned}$$

优化目标为 $\max_{\psi} \{L_D\}$ 。

基频取对数后,在源说话人的对数基频 $\ln f_0$ 与目标说话人的对数基频 $\ln f'_0$ 之间建立映射关系^[19]:

$$\ln f'_0 = \mu' + \frac{\sigma'}{\sigma} (\ln f_0 - \mu) \quad (16)$$

式中: μ 、 μ' 和 σ 、 σ' 分别表示源说话人和目标说话人对数基频的均值和均方差。

2.3.2 转换阶段

采用 WORLD 声码器从测试语料中提取^[3,13]说话人语句的频谱包络特征 x 、基频及非周期性特征;将 x 输入训练好的 VAACGAN-SE 模型中,得到转换的频谱包络特征 $\hat{x}$;通过式(16)表示的映射关系^[19]实现源说话人和目标说话人基频的转换;非周期特征保持不变;联合 $\hat{x}$ 、 $\ln f'_0$ 和非周期特征通过 WORLD 声码器得到转换语音。

3 实验与分析

本文实验是在 Python 平台环境下实现,在 Intel(R) Xeon(R) CPU E5-2660 v4 @ 2.00 GHz、NVIDIA Tesla V100 (rev a1) 的 Linux 服务器上运行。采用标准通用数据库 VCC2018^[15] 中提供的非平行语料,共 4 名源说话人 (SF3、SF4、SM3 和 SM4) 和 4 名目标说话人 (TF1、TF2、TM1 和 TM2),共构成 16 个源-目标说话人转换对,即 16 个转换情形^[3]。每个说话人包含有 81 个训练语句,35 个测试语句。

在 VAACGAN-SE 模型中,文本编码器 (Text-Encoder) 与编码器 (Encoder) 的结构都是采用二维卷积神经网络,包括 5 个卷积层和 1 个全连接层^[3,13],5 个卷积层的过滤器大小为 7×1 ,步长为 3,过滤器深度设置为 16、32、64、128 和 300,激活函数采用 LReLU 函数^[18]。所述的鉴别器与 VAWGAN 网络不同,其中多了一层输出语音数据类别标签的网络层。编码器、生成器的网络结构与 VAWGAN 基准模型保持一致。图 5 为本文提出模型的网络结构示意图。图中: h 、 w 和 c 分别为高度、宽度和通道数, k 和 s 分别为卷积层的内

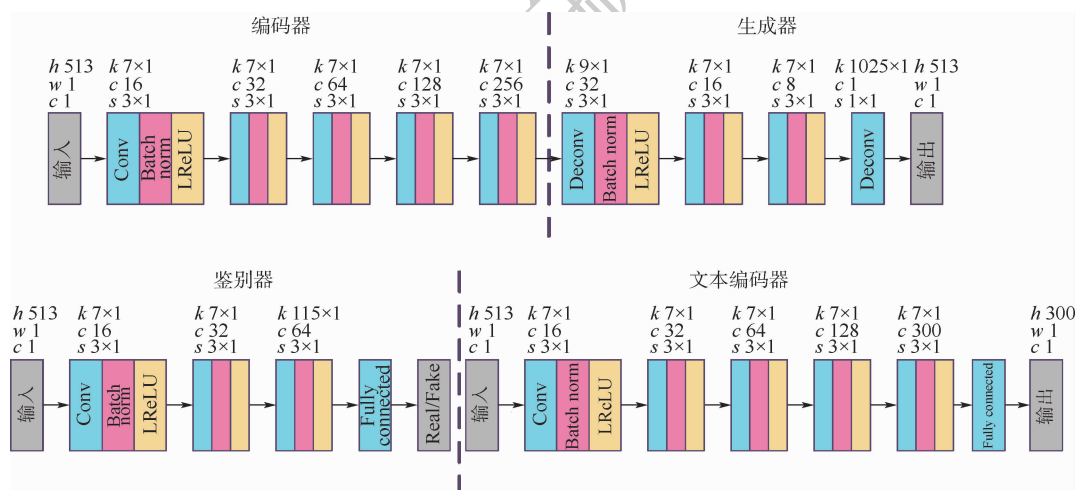


图 5 基于 VAACGAN-SE 模型的网络结构示意图

Fig. 5 Schematic diagram of network structure based on VAACGAN-SE model

核大小和步长, Conv、Deconv、Fully Connected 和 Batch norm 分别表示卷积、反卷积(转置卷积)、全连接层和批量归一化。频谱包络特征和非周期性特征为 513 维,隐变量 z 和句嵌入分别设置为 128 维和 300 维^[13],实验中训练批次大小为 16,学习率为 0.000 1,最大迭代次数设置 150 000。

采用通用的客观评价和主观评价对本文提出的 VAACGAN-SE 模型进行性能评估,并将其与基准模型 VAWGAN 进行对比。

3.1 客观评价

梅尔倒谱失真距离(Mel-Cepstral Distortion, MCD)作为业界公认的客观评价标准^[3],用来衡量转换语音与目标语音之间的频谱相似度^[1-2],MCD 的计算公式如下:

$$MCD = \frac{10}{\ln 10} \sqrt{2 \sum_{d=1}^D (c_d - \hat{c}_d)^2}$$
 (17)

式中: c_d 和 $\hat{c}_d$ 分别为目标语音和转换语音的第 d 维梅尔倒谱系数; D 为总维数^[3];MCD 单位为 dB。

图 6 为 4 种模型在 16 种转换情形下的 MCD 对比。分析可知,VAWGAN、VAACGAN、VAWGAN-SE 和 VAACGAN-SE 模型在 16 种转换情形

下的平均 MCD 值分别为 5.69、5.35、5.44、5.31 dB,本文提出的 VAACGAN、VAWGAN-SE 和 VAACGAN-SE 模型相比基准模型分别降低了 5.98%、4.39% 和 6.67%,表明 ACGAN 和句嵌入的引入,转换语音的频谱相似度更好,证实该方法能够有效地提高转换语音的合成音质。

如图 7 所示,在源-目标说话人对为 SF3-TM1 转换情形下,2 种模型得到的转换合成语音与源语音和目标语音的语谱图对比。分析可知,VAACGAN-SE

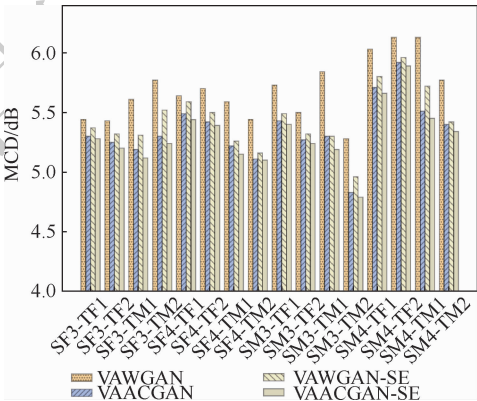
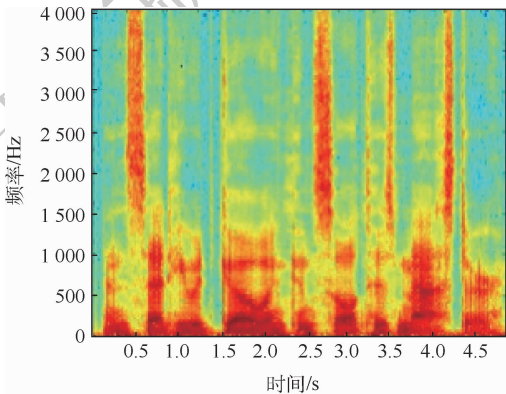
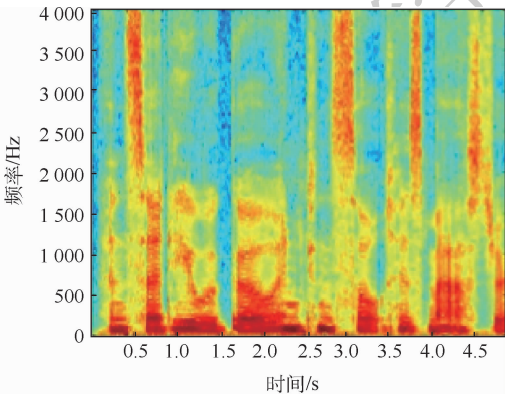


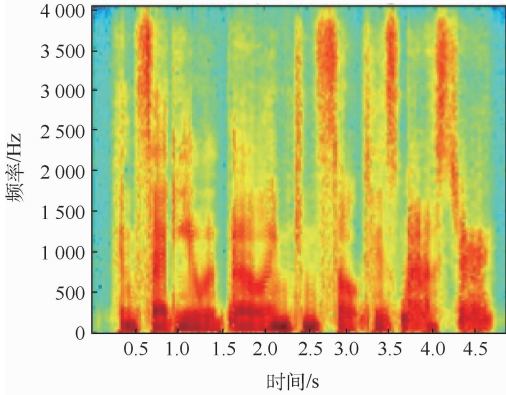
图 6 16 种转换情形下 4 种模型的转换语音 MCD 对比
Fig.6 Comparison of MCD of converted speech by four models in 16 kinds of conversion cases



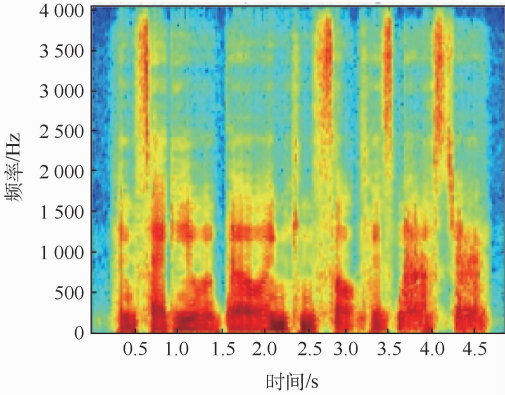
(a) 源说话人语音



(b) 目标说话人语音



(c) 基准模型 VAWGAN 得到的转换语音



(d) 本文提出模型 VAACGAN-SE 得到的转换语音

图 7 源-目标说话人对为 SF3-TM1 转换情形下基准模型 VAWGAN 与本文 VAACGAN-SE 模型转换语音的语谱图对比

Fig.7 Comparison of spectrogram between baseline VAWGAN and proposed VAACGAN-SE in voice conversion case of SF3-TM1

模型转换后合成语音的语谱图相比基准模型 VAWGAN,具有更完整的基音和谐波信息,以及更清晰的细节,在直观对比中,说明本文模型得到的转换语音与目标说话人的语音更为接近,在保持语义信息和说话人个性信息两方面更为完整。

3.2 主观评价

采用平均意见得分 (Mean Opinion Score, MOS) 和 ABX 值来分别评测转换语音的语音质量和个性相似度。主观评测人员为 20 名语音研究工作者,每个转换情形下有 35 句转换语音,从 4 种模型各自的 16 种转换情形下为每个人随机分配一句,并将不同方法对应的语句进行系统置乱。

在 MOS 测试中,评测人员对测听的转换语音进行质量打分,主要分为 5 个等级^[3]:1 分表示非常差,2 分表示较差,3 分表示可以接受,4 分表示良好,5 分表示非常好。在实验结果统计中,将 16 种转换情形根据不同性别划分为四大类:包括相同性别转换,即男到男、女到女 2 类;不同性别转换男到女、女到男 2 类。4 种模型转换语音的 MOS 值对比如图 8 所示。

从实验结果对比分析可得,VAWGAN、VAACGAN、VAWGAN-SE 和 VAACGAN-SE 的平均 MOS 值分别为 3.36、3.59、3.51 和 3.64,与基准模型 VAWGAN 比较,后 3 种模型的平均 MOS 值分别提高了 6.85%、4.46% 和 8.33%,表明本文提出的 ACGAN 和句嵌入的融合能够有效改善合成语音的音质,提高听觉质量。

在 ABX 测试中,X 代表转换语音,评测人员对测评语句进行评测,来判断转换语音 X 更加接近源说话人语音 A 还是目标说话人语音 B。本文将 16 种转换情形划分为相同性别转换和不同性别转换,测试结果分别如图 9 和图 10 所示。其中,

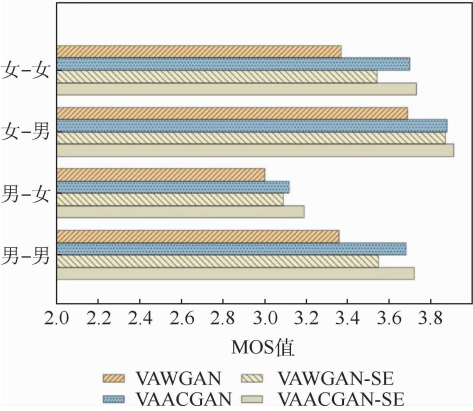


图 8 四类转换情形下 4 种模型转换语音的 MOS 值对比

Fig. 8 Comparison of MOS of voice conversion by four models under four conversion categories

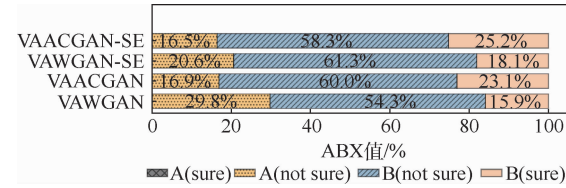


图 9 相同性别转换情形下 4 种模型转换语音的 ABX 值

Fig. 9 ABX of voice conversion by four models for intra-gender

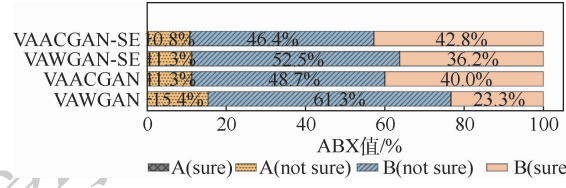


图 10 不同性别转换情形下 4 种模型转换语音的 ABX 值

Fig. 10 ABX of voice conversion by four models for inter-gender

A(sure) 用来表示确定判断转换语音为源说话人,A(not sure) 表示不完全确定转换语音是源说话人但很像^[3,13],B(sure) 表示确定判断转换语音为目标说话人,B(not sure) 表示不完全确定转换语音是目标说话人但很像。在结果分析中,将 B(not sure) 和 B(sure) 两项的比例之和作为评价转换语音个性相似性的指标。在所有测评中,没有评测人员确定判断转换语音为源说话人,实验结果对比图中 A(sure) 该项没有比例显示。

在相同性别转换情形对比中,基准模型 VAWGAN 和本文提出的 3 种模型 VAACGAN、VAWGAN-SE 和 VAACGAN-SE 的 ABX 值分别为 70.2%、83.1%、79.4% 和 83.5%,相比基准模型,本文提出的改进模型分别提升了 18.4%、13.10% 和 18.9%。异性别转换情形下,4 种模型的 ABX 值的比例分别为 84.6%、88.7%、88.7% 和 89.2%,相比基准模型,本文提出的改进模型分别提升了 4.8%、4.8% 和 5.4%。在不同性别转换情形下,4 种模型的说话人个性相似度均不同程度地优于相同性别转换情形,分析认为,这是因为异性别转换情形下,不同性别之间的转换更为明显,人耳对说话人倾向性的测听会更为突出。在相同性别和不同性别 2 种情形下,平均 ABX 值分别提升了 10.98%、8.59% 和 11.56%。

综上分析可得,相比基准模型,本文提出的 VAACGAN-SE 模型平均 MCD 值降低了 6.67%,平均 MOS 值提高了 8.33%,平均 ABX 值提高了 11.56%,充分验证本文提出的 ACGAN 和融合句嵌入的策略有效改善了转换语音的音质,同时对说话人个性相似度的提升也十分显著。

4 结 论

本文提出一种融合句嵌入的VAACGAN多对多语音转换方法,能够高质量地完成非平行文本条件下的多对多语音转换。

1) 利用ACGAN的鉴别器增加对样本分类的能力,实现了对频谱包络特征真假及其所属说话人类别的同时判定,使得生成的频谱包络特征与目标说话人的频谱特征更加接近。

2) 通过预训练文本编码器将说话人语音编码为句嵌入,利用其增加训练过程中的约束,增强隐变量表征语义内容的能力,有效克服了C-VAE模块中隐变量过度正则化效应导致语义内容趋于消失这一问题。

下一步的研究工作将考虑韵律特征尤其是序列到序列的语音转换。

参考文献 (References)

- [1] GODOY E, ROSEC O, CHONAVEL T. Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2012, 20(4): 1313-1323.
- [2] TODA T, CHEN L H, SAITO D, et al. The voice conversion challenge 2016 [C] // Interspeech, 2016: 1632-1636.
- [3] 李燕萍, 曹盼, 石杨, 等. 非平行文本下基于变分自编码器和辅助分类器生成对抗网络的语音转换[J]. 复旦学报(自然科学版), 2020, 59(3): 322-329.
LI Y P, CAO P, SHI Y, et al. Voice conversion based on variational autoencoder and auxiliary classifier generative adversarial network in non-parallel corpora [J]. Journal of Fudan University (Natural Science), 2020, 59(3): 322-329 (in Chinese).
- [4] DONG M, YANG C, LU Y, et al. Mapping frames with DNN-HMM recognizer for non-parallel voice conversion [C] // 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Piscataway: IEEE Press, 2015: 488-494.
- [5] ZHANG M, TAO J, TIAN J, et al. Text-independent voice conversion based on state mapped codebook [C] // 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2008: 4605-4608.
- [6] NAKASHIKA T, TAKIGUCHI T, MINAMI Y. Non-parallel training in voice conversion using an adaptive restricted Boltzmann machine [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(11): 2032-2045.
- [7] MOUCHTARIS A, VAN DER SPIEGEL J, MUELLER P. Non-parallel training for voice conversion based on a parameter adaptation approach [J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(3): 952-963.
- [8] HSU C C, HWANG H T, WU Y C, et al. Voice conversion from non-parallel corpora using variational auto-encoder [C] // 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA). Piscataway: IEEE Press, 2016: 1-6.
- [9] HSU C, HWANG H, WU Y, et al. Voice conversion from unaligned corpora using variational autoencoding Wasserstein generative adversarial networks [C] // Interspeech, 2017: 3364-3368.
- [10] ZHANG J, LING Z, DAI L R. Non-parallel sequence-to-sequence voice conversion with disentangled linguistic and speaker representations [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 28(1): 540-552.
- [11] ARJOVSKY M, CHINTALA S, BOTTU L. Wasserstein generative adversarial networks [C] // Proceedings of the 34th International Conference on Machine Learning (ICML). New York: ACM Press, 2017: 214-223.
- [12] SAITO Y, IJIMA Y, NISHIDA K, et al. Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors [C] // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 5274-5278.
- [13] 石杨. 非平行文本条件下基于文本编码器, VAE 和 ACGAN 的多对多语音转换研究 [D]. 南京: 南京邮电大学, 2019: 34-69.
SHI Y. Non-parallel voice conversion using ACGAN and variational autoencoders conditioned by sentence embedding [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2019: 34-69 (in Chinese).
- [14] ODENA A, OLAH C, SHLENS J. Conditional image synthesis with auxiliary classifier GANs [C] // Proceedings of the 34th International Conference on Machine Learning, 2017: 2642-2651.
- [15] LORENZO-TRUEBA J, YAMAGISHI J, TODA T, et al. The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods [C] // The Speaker and Language Recognition Workshop, 2018: 195-202.
- [16] ARORA S, LIANG Y Y, MA T Y. A simple but tough-to-beat baseline for sentence embeddings [C] // ICLR, 2017: 1-16.
- [17] MORISE M, YOKOMORI F, OZAWA K. WORLD: A vocoder-based high-quality speech synthesis system for real-time applications [J]. IEICE Transactions on Information and Systems, 2016, 99(7): 1877-1884.
- [18] MAAS A L, HANNUN A Y, NG A Y. Rectifier nonlinearities improve neural network acoustic models [J]. Computer Science, 2013, 30(1): 1152-1160.
- [19] 左宇涛. 非平行文本条件下基于 i-vector, VAE 和 GAN 的多对多语音转换算法研究 [D]. 南京: 南京邮电大学, 2019: 35-64.
ZUO Y T. Research on many-to-many voice conversion based on i-vector, variational auto-encoder and generative adversarial networks for non-parallel corpora [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2019: 35-64 (in Chinese).

作者简介:

李燕萍 女, 博士, 副教授。主要研究方向: 语音转换和说话人识别。

曹盼 女, 硕士。主要研究方向: 语音转换。

石杨 男, 硕士。主要研究方向: 语音转换。

张燕 女, 博士, 教授。主要研究方向: 模式识别和领域软件工程。

Many-to-many voice conversion with sentence embedding based on VAACGAN

LI Yanping^{1,*}, CAO Pan¹, SHI Yang¹, ZHANG Yan²

(1. College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

2. School of Software Engineering, Jinling Institute of Technology, Nanjing 211169, China)

Abstract: To solve the problems of poor speech quality and unsatisfactory speaker similarity for converted speech in existing non-parallel VC methods, this paper presents a novel voice conversion model based on Variational Autoencoding Auxiliary Classifier Generative Adversarial Network (VAACGAN) with sentence embedding, which achieves high-quality many-to-many voice conversion for non-parallel corpora. First, in the ACGAN, the discriminator contains an auxiliary decoder network that predicts the true or false of the spectral feature meanwhile the speaker category to which the training data belongs, thus achieving a more stable training process and a faster iterative convergence. Furthermore, sentence embedding is obtained by training the text encoder, which is introduced into the model as semantic content constraint, can enhance the ability of latent variables to characterize the speech content, effectively solve the over-regularization effect of latent variables and improve the quality of converted speech significantly. Experimental results show that the average value of MCD of the converted speech is decreased by 6.67%, MOS is increased by 8.33%, and ABX is increased by 11.56% compared with baseline method, which demonstrate that the proposed method significantly outperforms the baseline method in both speech quality and speaker similarity and achieves high-quality voice conversion.

Keywords: voice conversion; sentence embedding; text-encoder; Auxiliary Classifier Generative Adversarial Network (ACGAN); variational autoencoder; non-parallel corpora; many-to-many

Received: 2020-08-31; **Accepted:** 2020-09-04; **Published online:** 2020-11-17 15:31

URL: kns.cnki.net/kcms/detail/11.2625.V.20201117.1417.002.html

Foundation items: National Natural Science Foundation of China (61401227, 61872199, 61872424); Special Project of Intelligent Human-Computer Interaction Technology Innovation Team Building of Jinling Institute of Technology (218/010119200113)

* **Corresponding author.** E-mail: liyp@njupt.edu.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0444

一种时空特征聚合的水下珊瑚礁鱼检测方法



陈智能¹, 史存存², 李轩涯^{3,*}, 贾彩燕², 黄磊⁴

(1. 中国科学院自动化研究所 数字内容技术与服务研究中心, 北京 100190;

2. 北京交通大学 计算机与信息技术学院, 北京 100044;

3. 百度公司, 北京 100085; 4. 中国海洋大学 信息科学与工程学院, 青岛 266100)

摘

要: 水下监控视频中的珊瑚礁鱼检测面临着视频成像质量不高、水下环境复杂、

珊瑚礁鱼视觉多样性高等困难, 是一个极具挑战的视觉目标检测问题, 如何提取高辨识度的特征成为制约检测精度提升的关键。提出了一种时空特征聚合的水下珊瑚礁鱼检测方法, 通过设计视觉特征聚合和时序特征聚合 2 个模块, 融合多个维度的特征以实现这一目标。前者设计了自顶向下的切分和自底向上的归并方案, 可实现不同分辨率多层卷积特征图的有效聚合; 后者给出了一种帧差引导的相邻帧特征图融合方案, 可通过融合多帧特征图强化运动目标及其周边区域的特征表示。公开数据集上的实验表明: 基于以上 2 个模块设计的时空特征聚合网络可以实现对水下珊瑚礁鱼的有效检测, 相比于多个主流方法和模型取得了更高的检测精度。

关键词: 珊瑚礁鱼; 卷积神经网络; 时空联合特征; 目标检测; 特征融合

中图分类号: TP399

文献标志码: A

文章编号: 1001-5965(2021)03-0509-11

珊瑚礁鱼泛指生活在热带海洋珊瑚丛中的各种鱼, 它们种类繁多、形态各异、色彩斑斓, 是最富活力和观赏性的海洋生物群体之一。研究表明, 珊瑚礁鱼的种类、数量和活动痕迹是否丰富, 直接反映了珊瑚礁生态系统的健康状态和海洋生物多样性丰富程度^[1]。珊瑚礁鱼活动的显著变化, 则往往与温度剧变、水域污染和过度人类活动等事件紧密联系^[2]。通过监测分析珊瑚礁鱼的活动, 可以快速、准确、精细地掌握海洋生态系统的健康状况。此外, 珊瑚礁鱼的分析研究还将有助于帮助海洋生物学家研究不同海洋动物的行为及其之间的相互作用^[3]。目前, 随着全球大多数珊瑚礁生态系统呈退化趋势, 这一研究已经得到了广泛重视。

在海洋科学领域, 最初人们主要采取人工撒

网法^[4]和潜水调查法^[5]开展珊瑚礁鱼活动调查。人工撒网法先由人在珊瑚礁水域撒网捞鱼, 再经过海洋生物学家整理得到分析结果。潜水调查法由专业潜水员手持水下摄像机穿越珊瑚礁水域, 通过对所拍摄影像进行事后分析, 得到调查范围内的珊瑚和珊瑚礁鱼情况。虽然已被沿用很多年, 它们的缺点也很明显: 执行一次调查不仅消耗大量人力物力, 而且会对鱼类活动产生一定影响, 此外还难以获得大面积连续的监测数据。随着水下成像技术的发展, 在珊瑚礁水域特定位置安装水下摄像机, 采集珊瑚礁鱼活动影像并进行分析正在成为一种普遍接受的做法。相比于传统调查方法, 水下摄像既不影响珊瑚礁鱼行为, 同时也为后续分析提供了大量素材。目前, 全球多个国家和地区的珊瑚礁水域都部署了水下摄像系统并产

收稿日期: 2020-08-24; 录用日期: 2020-09-19; 网络出版时间: 2020-10-29 14:22

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201029.1355.001.html

基金项目: 国家自然科学基金(61772526, 61876016, 61872326); 百度开放研究基金

* 通信作者。E-mail: lixuanya@baidu.com

引用格式: 陈智能, 史存存, 李轩涯, 等. 一种时空特征聚合的水下珊瑚礁鱼检测方法[J]. 北京航空航天大学学报, 2021, 47(3): 509-519. CHEN Z N, SHI C C, LI X Y, et al. An underwater coral reef fish detection approach based on aggregation of spatio-temporal features[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 509-519 (in Chinese).

生了大量珊瑚礁鱼监控影像。对这些影像的分析催生了跨学科交叉研究需求:海洋生物学家手动分析每天产生的大量影像数据是不切实际的,迫切需要智能化的珊瑚礁鱼分析技术,能够从真实水下环境采集的影像中,自动得到珊瑚礁鱼的出现位置、种类、数量等信息。

在信息科学领域,视频大数据智能分析是一个广受关注的研究课题。特别是近年来,基于深度学习的视频分析与理解取得了显著突破,在大规模视频分类与检测^[6-10]、细粒度图像分类^[11-12]等任务上,深度学习相比于传统方法性能取得了大幅提升。但是,目前多以消费类视频图像及安防、交通等领域的监控视频为研究对象,对水下影像的分析相对较少。水下影像具有成像质量不高、水下环境复杂等分析难点,此外具体到珊瑚礁鱼检测上,还存在视觉多样性高、标注数据有限等挑战,这些困难决定了直接应用其他领域成熟的分析方法并不是最优方案,需要专门研究珊瑚礁鱼的有效检测方法。

目前,已经有一些针对珊瑚礁鱼检测分析的研究工作。早期研究多在受限情况下开展。例如,文献[13]提出了一种基于轮廓匹配的鱼识别方法,文献[14]提出了一种基于特征变换和支持向量机的罗非鱼自动分类方法,他们的实验都是在已捕捞、拍照时摆放较规则的鱼的图像上开展。针对水下自然环境中生活的鱼,文献[15-16]从不同角度提出了联合形状和纹理特征的鱼分类方法,结果表明,水下环境鱼检测分类的难度明显大于之前的受限环境。面对水下珊瑚礁鱼成像分辨率低的问题,Wei等^[17]提出了一种利用互联网高分辨率鱼图像进行数据增强的珊瑚礁鱼分类方法。围绕真实水下监控视频珊瑚礁鱼检测的 Sea-CLEF 系列国际竞赛^[18],来自韩国首尔大学^[19]、德国耶拿大学^[20]的团队采用了运动前景提取与基于深度学习分类相结合的解决方案,取得了较好成绩。近期相关工作更倾向于利用深度学习目标检测模型来解决珊瑚礁鱼检测问题。例如,文献[21]提出了一个相邻层特征融合的全卷积网络进行珊瑚礁鱼检测。为更好应对水下复杂环境,Zhuang等^[22]提出了先用 SSD 模型^[23]检测珊瑚礁鱼,再用 ResNet 网络^[24]对检测前景进行分类的方法。印度韦洛尔技术大学的研究团队^[25]评估了不同主干网 Faster R-CNN 模型^[26]在该竞赛上的检测性能。德国杜塞尔多夫大学的学者^[27]提出了一种基于 YOLO 模型^[28]的改进方法,取得了较好的检测效果。

上述研究虽然显著推动了珊瑚礁鱼自动检测技术的发展,但仍存在一些不足:将珊瑚礁鱼检测视为一个前景提取及分类的任务,或将其视为一个图像目标检测加上时序后处理的任务。前者虽然前景提取时能一定程度利用视频时序信息抑制水下复杂环境造成的负面影响,但其将检测过程切分成了前景提取和分类 2 个独立的子任务,二者无法相互促进和增强,检测性能受到限制。对于后者,由于检测时忽视了时序维度,且受目标大小有限、环境复杂等的影响,深度模型难以提取高质量检测特征,易造成误检和漏检,虽然时序后处理可消除一部分错误,但这也很难称之为视频时序信息的深入利用。

认识到特征辨识力不足是制约当前检测精度提升的关键因素,本文提出了一种时空特征聚合的水下珊瑚礁鱼检测方法。具体地,设计了一个新颖的卷积网络结构以提取更具辨识力的时空联合特征。该网络从 SSD 模型发展而来。同时,其包含一个多层视觉特征聚合模块,以提取更丰富的视觉特征,以及一个时序特征聚合模块,可结合运动目标生成时序强化的特征表示。通过以上 2 个模块实现对空间和时间 2 个维度特征的聚合,得到了可有效表征水下视觉目标的时空联合特征。公开数据集上的实验表明,本文方法可提升真实水下环境珊瑚礁鱼检测的精度。

本文的主要贡献如下:

- 1) 提出了一个多层视觉特征聚合的深度网络模块,设计了自顶向下的切分和自底向上的归并,可实现不同分辨率多层卷积特征图的聚合。
- 2) 提出了一个时序特征聚合的深度网络模块,可结合运动目标融合相邻帧的卷积特征图,从时序维度强化所提取特征。
- 3) 通过集成以上 2 个模块,提出了一个时空特征聚合的深度目标检测网络,可实现对视频目标特征的有效提取及检测。
- 4) 公开数据集的实验表明,本文方法可以有效检测真实水下环境中的珊瑚礁鱼,相比于传统方法和模型取得了更好的检测精度。

1 相关技术

本节对珊瑚礁鱼检测方法中涉及或相关的技术进行简要介绍,具体包括前景提取及分类、图像目标检测和视频目标检测 3 个方面。

1.1 前景提取及分类

前景提取及分类方法将珊瑚礁鱼检测视为一个前景目标提取及分类问题。由于当前水下摄像

采集的都是固定场景视频,借鉴安防、交通监控等领域的分析经验,利用多帧图像平均、高斯混合模型^[29]等方法可以对这类视频进行背景建模,进而可以通过背景差减和适当后处理,提取当前帧中的运动区域。将这些区域视为前景目标从图像中截取出来并归一化到特定大小,即可作为深度神经网络或其他机器学习模型的输入,构建相应分类模型实现珊瑚礁鱼检测。AlexNet^[30]、GoogleNet^[31]、ResNet^[24]都是现有文献中用到的分类网络。

前景提取的效果是这类方法能否取得好的结果的关键。由于珊瑚礁鱼在图像中通常只是一小部分,且受到影像分辨率低,以及水流、背景目标运动(如珊瑚摆动)等的影响,所提取的前景通常会有较多噪声。分类网络虽然可以滤除其中一部分,但因前景目标提取不完整、提取冗余等因素,不可避免会对分类精度造成一定影响。

1.2 图像目标检测

图像目标检测方法将珊瑚礁鱼检测视为一个基于单帧图像的目标检测问题。图像目标检测是随着深度学习技术发展性能得到显著提升的领域之一。根据检测原理的不同,现有检测方法主要分为两阶段方法和一阶段方法两大类。

两阶段方法一般包含2个网络:候选区域生成网络和检测网络。首先,使用候选区域生成网络在图像特征图上生成目标候选框;然后,使用检测网络对生成的目标候选框进行中心位置和长宽的回归,并进行分类。典型的两阶段方法包括Faster R-CNN^[26]、Cascade R-CNN^[32]等。

一阶段方法通过对目标位置、大小和长宽比进行密集的采样来检测目标。这类方法先在特征图的每个位置根据不同的尺寸和长宽比预定义固定数量的默认框,再对默认框的中心位置和长宽进行回归,并对其包含的物体进行分类判别。典型的一阶段方法有YOLO^[28]、SSD^[23]等。

除以上方法,近年来也有一些考虑目标定位损失^[33]和无需预定义默认框^[34]的方法被提出来并取得良好检测性能。此外,强化网络所提取特征的辨识力也是提升目标检测性能的重要方向。这方面代表性工作有特征金字塔网络FPN^[35]、默认框可适配学习的RefineDet^[36]等。

具体到珊瑚礁鱼检测方面,目前有文献用到Faster R-CNN^[26]、SSD^[23]和YOLO^[28]。因视频帧分辨率低、水下环境复杂、鱼体态呈多维变化等特点,所提取特征质量不可避免受到影响,制约了以上方法的精度。为此,文献[22,27]提出利用额外分类器来强化检测结果,这一做法的效果主要

体现在减少误判上,对漏判则作用不大。

1.3 视频目标检测

视频目标检测泛指同时利用图像静态特征和视频运动信息实现检测的各种方法。目前,主要有2类视频目标检测方法:第1类是图像目标检测及后处理。先在多帧图像上进行目标检测,再采取适当后处理,得到视频级检测结果。这类方法是图像目标检测方法的简单延伸,其存在难以充分利用检测结果的时序相关性等不足。第2类方法利用可同时接受视觉和时序信息作为输入的深度模型进行检测。目前,已经有一些相关的网络结构被提出来。例如,利用2个卷积神经网络分别处理视觉和运动信息的双流神经网络^[6],利用卷积神经网络提取单帧图像视觉特征,进而用长短时记忆神经网络建模相邻帧之间相关性的CNN-LSTM结构^[7],利用3D卷积提取时空联合特征的三维卷积神经网络^[9]。这些方法统筹考虑了时空域,因而可以提取到更为强大的特征,从而提升行为识别、视频分类等多个视频任务的分析精度。

由于珊瑚礁鱼通常只占视频帧的一小块区域,且其位置随着时间变化,构建有效的局部视觉和时序特征表示是视频目标检测的关键。这方面典型工作有光流引导特征聚合网络^[10],依据光流方向聚合相邻帧特征图以强化对当前帧目标的表示,可适当缓解目标因运动模糊、面积过小、罕见姿态等问题导致的检测困难。此外,也有学者研究了同时进行目标检测和跟踪的网络^[37],通过二者的互补提升检测性能。

2 本文方法

2.1 整体网络结构

本节介绍提出的时空特征聚合水下珊瑚礁鱼检测方法。图1给出了时空特征聚合神经网络的整体结构。可以看到,该网络接受当前视频帧及前后相邻帧作为输入。当前帧通过图中c1~cf组成的视觉特征聚合模块(Visual Feature Aggregation, VFA),生成一个融合了多层卷积特征图、信息更丰富的新视觉特征图。同时,对于其中参与目标预测的特征图(cf, c15~c19),通过图中实线框表示的时序特征聚合模块(Temporal Feature Aggregation, TFA)对当前帧及相邻帧特征图予以聚合,从而生成具备更强表示能力的时空聚合特征(Visual-Temporal Feature Aggregation, VTFA)。珊瑚礁鱼检测将在空间分辨率逐步降低的多个聚合特征图上进行。

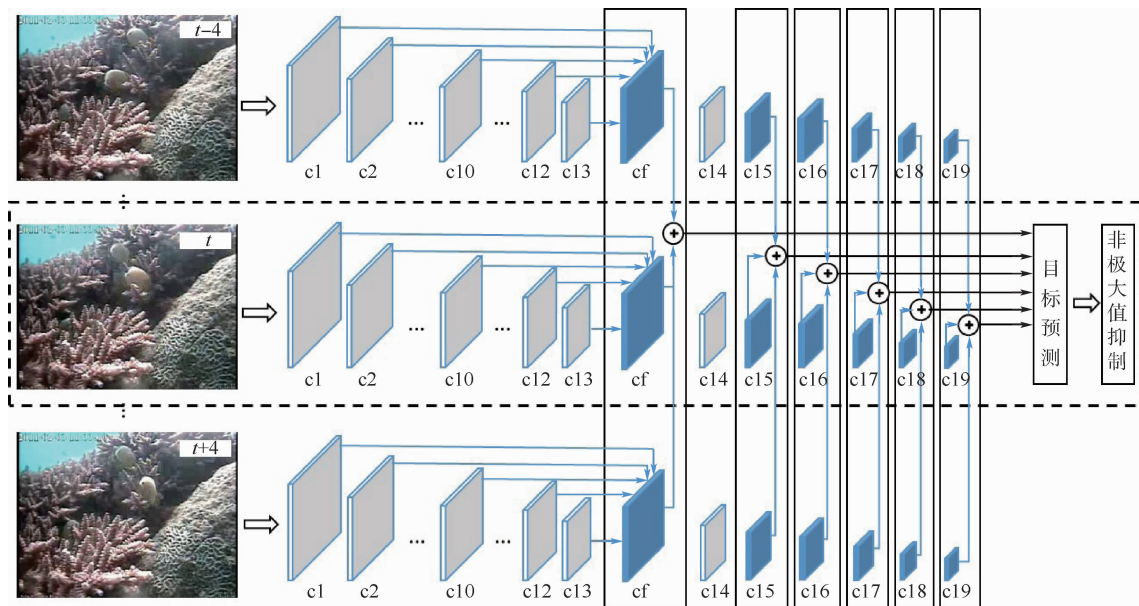


图1 本文时空特征聚合神经网络的整体结构

Fig. 1 Overall architecture of the proposed spatio-temporal features aggregation neural network

本文网络可视为一种从 SSD 模型^[23]发展而来的复合结构,在其单帧图像处理通道中(见图1中虚线框),类似于 SSD 利用卷积神经网络 VGG16^[38]作为特征提取主干网,其中基本卷积层 c1 ~ c13 与 VGG16 一致,最后 2 个全连接层和分类层被截断,予以替换的是 6 个空间分辨率逐步降低的新增卷积层(c14 ~ c19)。在 SSD 中,目标检测将在 c10、c15 ~ c19 这 6 个不同尺度的卷积特征图上进行。

损失函数方面,本文网络与 SSD 的形式相同,整个网络的损失函数定义为

$$L(\mathbf{x}, \mathbf{c}, \mathbf{l}, g) = \frac{1}{N} (L_{\text{conf}}(\mathbf{x}, \mathbf{c}) + aL_{\text{loc}}(\mathbf{x}, \mathbf{l}, g)) \quad (1)$$

式中: $\mathbf{x}$ 为记录预测框和标注框匹配情况的矩阵; $\mathbf{c}$ 为当前样本的预测置信度向量; $\mathbf{l}$ 和 g 分别为预测框和标注框的坐标信息; N 为当前帧预测框数量; L_{conf} 和 L_{loc} 分别为类别损失和定位损失; a 为一个用于平衡 2 类损失的参数,本文设置为 1。以上损失项的计算公式可参见文献[23]。

不同于 SSD,本文网络的目标预测是在经过单帧图像多层特征图聚合与/或相邻帧同层特征图聚合后生成的时空强化特征图上进行,这 2 个聚合模块正是本文的创新之处。

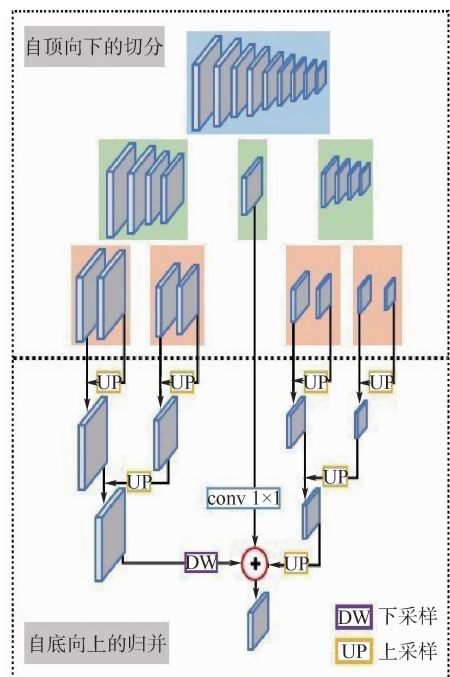
2.2 视觉特征聚合模块

源自 VGG16 的基本卷积层中,SSD 仅用 c10 进行预测,忽视了其他层信息。水下监控视频由于质量低、成像环境复杂且珊瑚礁鱼目标通常较小,基本卷积层特征图上的信息对于检测来说尤

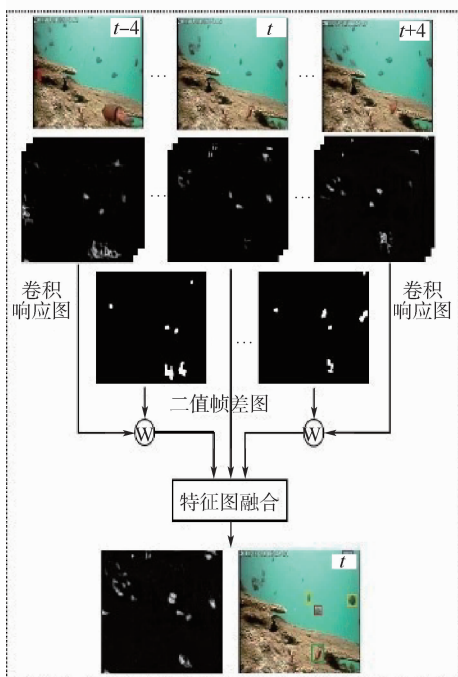
其重要。基于此,本文设计了一个视觉特征聚合模块对基本卷积层进行更有效的利用,以提高水下复杂环境中的珊瑚礁鱼检测性能。

视觉特征聚合模块由一个自顶向下的切分和一个自底向上的归并操作组成。切分过程迭代地将卷积层分成不同的组,形成了一个自顶向下的切分结构。在这个结构的最顶层,所有卷积层都在同一组。当卷积层数量 n 是偶数时,在下一层它们将从中间切分,分成 2 个各含有 $n/2$ 个层的组;当卷积层的数量 n 是奇数时,在下一层最中间的卷积层将被视为一个单独的组,其左右两边的卷积层被分为另外 2 个组,各含有 $(n-1)/2$ 个层。基于这一原则,卷积层可以不断被切分,直至每个组中卷积层的数量小于等于 2,此时切分过程停止。图 2(a)的上半部分给出了一个切分的例子。

基于以上切分结果,自底向上的归并从下往上不断合并每个组中的特征图,最终形成了一个聚合了所有卷积特征图的特征。具体地,卷积层归并时涉及对 2 个或 3 个分辨率和通道数可能不同的特征图融合。由于卷积神经网络特征图从浅层到深层满足分辨率不变或递减的规律,对于 2 个特征图融合的情况:若分辨率不同,将低分辨率特征图上采样到与高分辨率特征图具有相同大小,再进行融合;若分辨率相同则直接融合。对于 3 个特征图融合的情况,先保持中间层特征图的分辨率不变,若其浅层方向特征图分辨率大于中间层,则通过下采样将其降采样到与中间层特征图相同分辨率再融合;若其深层方向特征图分辨率小于中间层,则将其上采样到与中间层特征图



(a) 视觉特征聚合模块



(b) 时序特征聚合模块

图2 本文提出的视觉特征聚合模块和时序特征聚合模块

Fig.2 The proposed visual feature aggregation module and temporal feature aggregation module

一样大小再进行融合;若浅层和/或深层方向特征图分辨率与中间层的一致,则直接融合。通过迭代进行融合操作,最终所有特征图将会被融合成一个具有适中分辨率的聚合特征图,如图2(a)下半部分所示,该特征图将会替换c10进行预测。

对于融合过程中的特征图通道数可能不一致的情况,以(三层融合的)中间层或(两层融合中)浅层方向的特征图为准,融合前先通过 1×1 的

卷积将其他层特征图的通道数予以对齐。注意到,该特征聚合模块不仅适用于VGG16主干网,也可以推广到其他卷积主干网。相比于仅利用c10进行预测,以及文献[21]仅融合相邻预测层的方案,本文视觉特征聚合模块以一种合理且可扩展的方式聚合了多个基本卷积层的特征图,提供了更丰富的局部细节和上下文信息,有利于更好刻画视频帧中的珊瑚礁鱼目标。

2.3 时序特征聚合模块

珊瑚礁鱼在水下游动时体态呈多维变化,当其部分遮挡或以罕见体态出现时,不可避免会带来检测困难。融合相邻帧特征显然有利于缓解该问题。基于此,本文设计了时序特征聚合模块在相邻帧上融合运动目标相关的特征图,以生成更强大的特征表示。

时序特征聚合模块的示意图如图2(b)所示。对于输入到网络的当前视频帧及其前后相邻帧,利用图1所示的主干网结构提取每帧图像各个卷积层的特征图,这些特征图记录了目标在当前图像上的卷积响应值。基于此,在特征图上计算当前帧与每个相邻帧的帧差图,对帧差图进行灰度化和二值化并结合适当后处理,如图2(b)所示,记录了当前帧与其相邻帧之间运动信息的二值帧差图。

对于网络中参与目标预测的特征图,通过以下公式对其进行时序聚合:

$$f'_{i,n} = \sum_{I_j \in S_i} w_j \cdot B_{i,j} \otimes f_{j,n} \quad (2)$$

式中: $f_{j,n}$ 为视频帧 I_j 聚合前的第 n 层特征图; $f'_{i,n}$ 为视频帧 I_i 聚合后的第 n 层特征图; $B_{i,j}$ 为 I_i 与 I_j 的二值帧差图,当 $i=j$ 时, $B_{i,j}$ 为值全为1的矩阵;“ $\otimes$ ”为对应相乘操作; $S_i = (I_{i-k}, I_{i-k+1}, \dots, I_{i+1})$ 为 I_i 及其相邻帧集合, k 为邻域边界; w_j 为 I_j 对应的权重,即

$$w_j = \begin{cases} 2^{k+j-i}/\sigma & j \leq i \\ 2^{k+i-j}/\sigma & j > i \end{cases} \quad j \in [i-k, i+k] \quad (3)$$

其中: σ 为归一化因子,以确保所有权重之和为1。

$$\sigma = \sum_{i=1}^k 2^i + 2^k \quad (4)$$

式(2)以线性加权的方式,将相邻帧特征图中对应帧差运动区域的特征融合到当前帧同层特征图中。这一做法可以生成一个以当前帧运动目标为中心,适当囊括其周边区域,时序强化的特征。由于珊瑚礁鱼是视频中的运动主体,这一做法可有效缓解当前帧珊瑚礁鱼目标因运动模糊、罕见姿态等带来的特征表示困难。

以上时序融合在所有6个参与目标预测的卷

积特征图上都将进行,其中包括了一个通过视觉特征聚合模块生成的预测层。因此,网络可提取到时空联合的强化特征更好地进行目标检测。注意到,文献[10]也提出了一种光流引导的相邻帧特征图融合方法。本文与其有2点区别:①在融合区域确定上,文献[10]先计算两帧之间的光流,再依此将每帧光流前景对应的特征图位移后再与当前帧相应位置叠加融合。与之对应,本文采用了计算代价显著降低的帧差运算,融合区域也是相对更宽泛的帧差前景区域。采用这一做法主要是考虑到低质量视频中光流计算误差较大,容易导致位移估计不准确。此外,认为鱼周边区域的特征也有助于检测。②在相邻帧融合的权重上,文献[10]用余弦相似度动态计算当前帧与相邻帧的权重,而本文采用的是一个以当前帧为中心的高斯分布权重,直接赋予与当前帧更邻近的相邻帧更大权重。这一做法在降低计算量的同时,也一定程度避免了相似度计算对噪声敏感的影响。

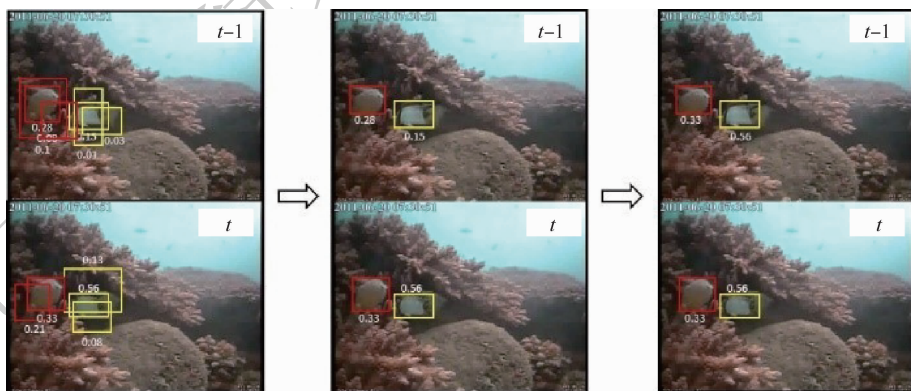


图3 非极大值抑制和本文提出的时序后处理

Fig. 3 Non-maximum suppression and the proposed temporal post-processing

本文网络实现时,由于当前帧预测需要利用前后相邻帧特征图,为避免重复提取图像特征,在确定邻域参数 k 及其采样间隔后,将申请一个公共缓存空间存储以当前帧为中心,邻域内所有采样图像参与预测的卷积特征图。这样,每帧检测时,只需计算当前帧的时空联合聚合特征图以开展以当前帧为中心的目标检测。对一个视频帧序列,则只需按时序对所有采样帧重复以上过程,相应调整缓存空间内容,即可依次计算所有采样帧上的检测结果。

3 实验与结果

3.1 数据集

用SeaCLEF国际竞赛^[18]数据作为本文实验数据集。该数据集提供了5个不同场景和日期的93个水下监控视频,给出了其中出现的15种珊

瑚礁鱼,上述融合中,时序聚合的邻域 k 是一个重要参数。大的 k 值融合的相邻帧多,但网络结构更复杂,计算代价更高;小的 k 值则有时序信息融合不充分的隐患。此外,选定 k 值后,邻域中图像分析的采样间隔也是一个需要明确的细节。将在消融实验中论证不同做法的区别。

2.4 后处理

基于帧级检测结果及置信度,本文先利用非极大值抑制消除单帧图像上的冗余检测框,再提出了一个时序后处理以提升珊瑚礁鱼检测得分置信度过低,易造成漏检和误检的现象。具体地,在得到单帧检测结果后,先将相邻帧中满足检测类别相同且 $\text{IoU} > 0.5$ 的检测框标记成检测对,再将检测对中检测框得分统一为置信度高的得分。通过这种方式,一定程度利用了目标的时序互补性强化了检测得分,使检测结果更稳定。上述非极大值抑制和时序后处理如图3所示。

瑚礁鱼的逐帧标注,包括鱼的类别和矩形框形式的位置信息,共有21 396个标注样例。该数据集涵盖了图像分辨率低、水下环境复杂、鱼体态变化大等一系列真实水下监控视频包含的检测难点。

竞赛将数据集分成了训练集和测试集,分别包含20个和73个视频的13 882个和7 514个标注实例。训练集和测试集都涵盖了全部的5个场景。但是,不同鱼在数据集分布并不均匀,出现次数最多的网纹宅泥鱼在训练集和测试集上分别出现了3 165和5 046次,15种鱼中黑缘单鳍鱼在测试集中仅出现了8次,甚至镜斑蝴蝶鱼和黑鳍粗唇鱼在测试集没有出现。因此参考文献[21],本文将上述3类鱼从检测任务中去除,构成了一个包含12种珊瑚礁鱼的目标检测任务。表1给出了这些鱼的名称,以及它们在训练集和测试集中的数量分布情况。

表 1 SeaCLEF 数据集中不同类别鱼的数量
Table 1 Numbers of different fish species on
SeaCLEF dataset

编号	珊瑚礁鱼名称	训练集样例数	测试集样例数
1	五带豆娘鱼	132	93
2	褐斑刺尾鲷	294	129
3	克氏双锯鱼	363	516
4	月斑蝴蝶鱼	1 217	1 896
5	川纹蝴蝶鱼	335	1 317
6	短身光鳃雀鲷	275	24
7	宅泥鱼	894	1 985
8	网纹宅泥鱼	3 165	5 046
9	康德锯鳞鱼	242	118
10	黄新雀鲷	85	1 593
11	迪克氏固齿鲷	737	700
12	宝石高鳍刺尾鱼	72	187

3.2 实验设置及评价指标

本文采用一个两步训练过程来训练网络。第 1 步基于单帧图像训练一个仅包含视觉特征聚合模块的目标检测网络:先读取 ImageNet 数据集的预训练参数,再采用批量随机梯度下降方法进行训练,批的大小为 32 张图像。设置网络总共迭代训练 120 000 次。先将学习率设置为 0.000 1 进行 1 000 次迭代的热身训练,完成热身训练以后将学习率升至 0.001,迭代训练 40 000 次和 80 000 次之后,分别将学习率降低为 0.000 1 和 0.000 01,以使网络更好地收敛。梯度更新动量值为 0.9。第 2 步训练基于第 1 步得到的参数训练整个网络结构。由于特征图相加操作的可导性,整个网络是端到端可训练的。第 2 步同样采用批量随机梯度下降的精调训练,批的大小为 1 张图像,迭代轮数设置为 60 000 次,其中前 40 000 次与后 20 000 次的学习率分别为 0.000 1 和 0.000 01。采用了随机剪裁和调整图像对比度的数据增强方式。图像在输入网络之前先将大小调整为 300 × 400。全部训练在一个 GTX Titan X GPU 上完成,基于 TensorFlow 平台完成整个模型训练需要约 34 h。

推理阶段,本文网络接受当前帧及其前后多个相邻帧作为输入,输出当前帧的检测结果。检测结果经过 2.4 节的后处理,可得到视频级检测结果。评价指标上,本文用目标检测领域广泛使用的平均精度均值 mAP,其定义为

$$\text{mAP} = \frac{\sum_{i=1}^n \text{AP}_i}{n}$$

(5)

式中:AP_{*i*} 为第 *i* 个目标类别通过改变阈值得到的不同召回率下的平均精度;*n* 为目标类别个数。

3.3 消融实验

通过消融实验来验证视觉特征聚合模块和时

序特征聚合模块中特征图的具体融合方式,以及时序融合时当前帧的邻域及采样间隔。

融合方式方面,验证视觉特征聚合模块时,将网络结构设置为仅输入当前帧 *I_t* 的情况,此时网络仅包含图 1 中虚线框的部分。对比了对应相加、取最大值和取平均值 3 种特征图融合策略。表 2 给出了相应的 mAP 值。可以看到,对应相加取得了更好的性能。

验证不同融合方式对时序特征聚合模块的影响。先将网络固定为输入 {*I_{t-4}*, *I_t*, *I_{t+4}*} 3 帧图像的情况。为简化起见,去除了网络中的视觉特征聚合模块。表 2 给出了上述 3 种情况下的 mAP 值。结果表明,取最大值进行融合更有利于进行珊瑚礁鱼检测。

分析采样邻域及间隔对结果的影响。受限于计算资源,仅考虑了输入不超过 5 个视频帧的情况。结合不同采样间隔,将其分成了如表 3 和表 4 所示的 11 种情况,其中 2 表示考虑 {*I_{t-2}*, *I_t*, *I_{t+2}*} 3 帧图像的情况,46 表示考虑 {*I_{t-6}*, *I_{t-4}*, *I_t*, *I_{t+4}*, *I_{t+6}*} 5 帧图像的情况,其余依此类推。实验中网络都未包括视觉特征聚合模块。

可以看到,输入 5 帧图像可以取得比 3 帧图像更好的检测结果。这一点是符合预期的,因为聚合更多相邻帧有利于提取更具辨识力的特征。此外,注意到与当前帧间隔为 6(3 帧情况),以及 4 和 8(5 帧情况)时,相比于其他间隔情况下取得了更好的结果。帧数间隔大一般关联着更大的运动区域,对应到本文网络则是更大范围的特征融合。当间隔相对适中时,可使得相邻特征图中目

表 2 不同融合方式及性能

融合方式	mAP	
	视觉特征聚合模块	时序特征聚合模块
对应相加	0.634 5	0.600 2
取最大值	0.632 8	0.601 2
取平均值	0.629 6	0.598 6

表 3 输入为 3 帧图像时不同参数下的网络性能
Table 3 Network performance under different
parameters when three-frame images are input

采样邻域及间隔	2	4	6	8
mAP	0.599	0.601	0.602	0.602

表 4 输入为 5 帧图像时不同参数下的网络性能
Table 4 Network performance under different
parameters when five-frame images are input

采样邻域及间隔	24	26	28	46	48	68
mAP	0.612	0.614	0.617	0.618	0.622	0.621

标及合适范围的周边上下文得到更强化的刻画,但若间隔过大,则容易融合到更多的噪声而起到负面作用。基于以上结果,本文网络结构最终确定为接受 $\{I_{i-8}, I_{i-4}, I_i, I_{i+4}, I_{i+8}\}$ 5 帧图像作为输入,并将对应相加和取最大值分别作为视觉特征聚合和时序特征聚合中多个特征图的融合方式。

3.4 实验结果及对比分析

为评估检测性能,将本文网络及其衍生结构和几种主流方法与模型进行了实验比较。

BS + GoogleNet^[20]:德国耶拿大学提出的基于前景提取及分类的珊瑚礁鱼检测方法。

Faster R-CNN^[26]、YOLOv3^[28]和 SSD^[23]:采用这 3 个主流目标检测模型进行珊瑚礁鱼检测。

FFDet^[21]:基于相邻卷积层特征融合的珊瑚礁鱼检测方法。

FGFA^[10]:光流引导的相邻帧特征图融合的检测方法。

Ours-VTFA、Ours-VFA 和 Ours-TFA:本文网络,以及本文网络分别去除时序特征聚合模块和视觉特征聚合模块后对应的珊瑚礁鱼检测方法。

表 5 给出了以上方法的图像级和视频级实验结果及时间消耗。可以看到,Ours-VTFA 方法相比于传统基于前景提取及分类、主流目标检测模型取得了8.8%~16.8%的相对性能提升,表明本文时空特征聚合网络能更好地检测水下珊瑚礁鱼。同时,该方法也取得了比仅考虑其中一种模态聚合的 Ours-VFA 和 Ours-TFA 更好的效果,说明从时间和空间 2 个维度强化特征提取的互补性和必要性。

视觉特征聚合方面,对比于没有特征融合的 SSD 和采用相邻层视觉特征融合的 FFDet, Ours-VFA 方法取得了更好的检测性能,说明基本卷积层聚合生成的特征图可以更好地描述珊瑚礁鱼的类别和位置信息,挖掘利用基本卷积层特征对低

质量水下视频中珊瑚礁鱼检测具有重要价值。此外,注意到 Ours-VFA 方法的检测速度显著快于 FGFA 等高精度方法,仅略逊于精度不如它的 SSD 和 FFDet。Ours-VFA 方法不失为一种速度和精度得到较好折中的检测方案。

时序特征聚合方面,Ours-TFA 方法与 SSD 的区别在于:前者在网络中聚合了相邻帧对应运动区域的特征图,即获得了 6.3% 的 mAP 相对提升,验证了时序维度的挖掘利用有助于提升检测性能。本文基于帧差的相邻帧聚合方法可以融合相邻帧目标周边的上下文区域,有助于提取更加有效的特征。该方法虽然检测性能低于 FGFA,但时间消耗减少了 2 倍以上,主要是帧差计算的代价显著低于光流计算。在 Ours-VFA 方法的基础上进一步融入时序特征聚合模块,可继续提升检测性能,再次说明了本文网络可以互补地聚合时间和空间维度的特征。注意到,Ours-VTFA 方法可以取得优于 FGFA 的实验结果,且检测时间也缩短了 2 倍以上,这也再一次凸显了聚合基本卷积层特征的重要性。此外,所有方法采用了本文后处理技术后,检测精度均有一定提升,表明网络内外挖掘的时序信息具有一定互补性,在不同层次利用它们可进一步提升检测精度。

图 4 给出了各种检测模型在不同珊瑚礁鱼类别上的检测结果。其中,每种鱼的 8 个检测结果从左到右分别是 Faster R-CNN、YOLOv3、SSD、FFDet、FGFA、Ours-VFA、Ours-TFA 和 Ours-VTFA 8 种方法取得的。可以看到,不同类型鱼的检测结果差异巨大。即使性能最好的方法,在褐斑刺尾鲷上的 AP 值也不超过 0.1。与之对应,无论是哪种方法,在宅泥鱼、克氏双锯鱼和月斑蝴蝶鱼上都取得了较高 AP 值。从两方面解释造成以上显著类间差异的原因:①珊瑚礁鱼样本数量在类别上分布不均。结合表 1 可看到,AP 值高的鱼类别样本较多,而 AP 低的鱼类别样本较少,训练数据是否丰富,一定程度上影响了所构建模型的检测精度。②不同类型鱼的视觉辨识难度各不相同,导致检测难度不一。水下低质量视频中鱼的检测主要依赖对鱼颜色和形态轮廓等的刻画,在训练数据相对充足的情况下,部分颜色和形态易于辨识,区分度较大的鱼可以被高精度的检测,宅泥鱼和月斑蝴蝶鱼均属于此类。但是,当训练数据较少,且鱼的区分性特征在低质量视频中不突出,易与其他目标混淆时,如褐斑刺尾鲷,检测精度则受到严重影响。以上实验表明,水下监控视频中珊瑚礁鱼的高精度检测仍然是一个极具挑战的技术难点。

表 5 不同方法的检测性能

模型	mAP		检测时间/s
	图像级	视频级	
BS + GoogleNet ^[20]	0.597	0.603	
Faster R-CNN ^[26]	0.571	0.581	0.153
YOLOv3 ^[28]	0.553	0.562	0.022
SSD ^[23]	0.576	0.586	0.050
FFDet ^[21]	0.614	0.628	0.065
FGFA ^[10]	0.643	0.647	0.384
Ours-VFA	0.624	0.635	0.067
Ours-TFA	0.619	0.622	0.113
Ours-VTFA	0.652	0.656	0.121

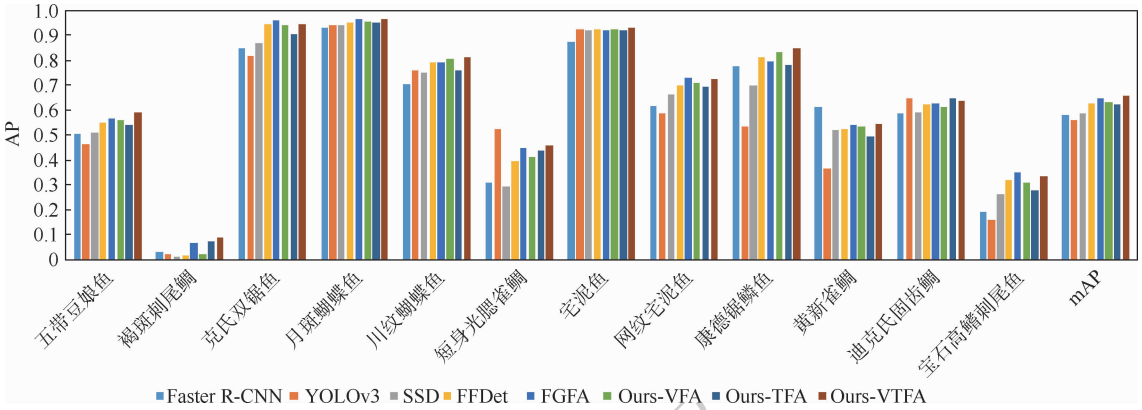


图 4 各种检测模型在不同珊瑚礁鱼类别上的检测结果
Fig.4 Detection results of different coral reef fish species by various detection models

4 结 束 语

认识到从充满挑战的水下监控视频中检测珊瑚礁鱼的重要价值,本文提出了一种时空特征聚合的水下珊瑚礁鱼检测方法。该方法从时间和空间 2 个维度出发,分别设计了一个视觉特征聚合模块以融合多层不同分辨率的卷积特征图,以及一个时序特征聚合模块以实现帧差引导的相邻帧卷积特征图融合。SeaCLEF 数据集上的实验基本验证了以上 2 个特征聚合模块的有效性。本文基于以上聚合模块提出的检测模型相比于多个典型方法和模型,可以取得更好的检测精度。

检测性能的提升主要归功于提取了更具表征力的时空联合特征。除了特征,数据和模型也是影响检测精度的关键因素。在下一步工作中,一方面将探索如何利用互联网上公开的珊瑚礁鱼视频图像,进一步提升低质量水下视频中珊瑚礁鱼的检测精度。另一方面,还将关注如何利用生成对抗网络等技术,生成更多高质量和多样化的珊瑚礁鱼训练样本,从而构建更加鲁棒有效的珊瑚礁鱼检测模型。此外,还计划在开源深度学习平台飞桨上复现该方法。

参考文献 (References)

[1] CINNER J E, HUCHERY C, MACNEIL M A, et al. Bright spots among the world's coral reefs[J]. Nature, 2016, 535 (7612): 416-419.

[2] 赵焕庭, 王丽荣, 袁家义. 南海诸岛珊瑚礁可持续发展[J]. 热带地理, 2016, 36(1): 55-65.

ZHAO H T, WANG L R, YUAN J Y. Sustainable development of the coral reefs in the South China Sea Islands[J]. Tropical Geography, 2016, 36(1): 55-65 (in Chinese).

[3] GRAHAM N A J, EVANS R D, RUSS G R. The effects of marine reserve protection on the trophic relationships of reef fishes on the Great Barrier Reef[J]. Environmental Conservation,

2003, 30(2): 200-208.

[4] HODGSON G. Reef check: The first step in community-based management[J]. Bulletin of Marine Science, 2001, 69 (2): 861-868.

[5] 李永振, 史赞荣, 艾红, 等. 南海珊瑚礁海域鱼类分类多样性大尺度分布格局[J]. 中国水产科学, 2011, 18(3): 619-628.

LI Y Z, SHI Y R, AI H, et al. Large scale distribution patterns of taxonomic diversity of fish in coral reef waters, South China Sea[J]. Journal of Fishery Sciences of China, 2011, 18 (3): 619-628 (in Chinese).

[6] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Advances in Neural Information Processing Systems, 2014: 568-576.

[7] DONAHUE J, HENDRICKS L A, ROHRBACH M, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 2625-2634.

[8] CHEN Z, AI S, JIA C. Structure-aware deep learning for product image classification[J]. ACM Transactions on Multimedia Computing, Communications, and Applications, 2019, 15 (1s): 1-20.

[9] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 4489-4497.

[10] ZHU X, WANG Y, DAI J, et al. Flow-guided feature aggregation for video object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 408-417.

[11] ZHAO B, WU X, FENG J, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245-1256.

[12] PU J, JIANG Y G, WANG J, et al. Which looks like which: Exploring inter-class relationships in fine-grained visual categorization[C]//European Conference on Computer Vision. Berlin: Springer, 2014: 425-440.

[13] LEE D J, SCHOENBERGER R B, SHIOZAWA D, et al. Contour matching for a fish recognition and migration-monitoring system[C]//Two- and Three-Dimensional Vision Systems for

- Inspection, Control, and Metrology II. International Society for Optics and Photonics, 2004, 5606: 37-48.
- [14] FOUAD M M M, ZAWBAA H M, EL-BENDARY N, et al. Automatic Nile tilapia fish classification approach using machine learning techniques[C] // 13th International Conference on Hybrid Intelligent Systems. Piscataway: IEEE Press, 2013: 173-178.
- [15] LARSEN R, OLAFSDOTTIR H, ERSBØLL B K. Shape and texture based classification of fish species[C] // SCIA 2009: Image Analysis. Berlin: Springer, 2009: 745-749.
- [16] SPAMPINATO C, GIORDANO D, DI SALVO R, et al. Automatic fish classification for underwater species behavior understanding [C] // Proceedings of the First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Image-ry Streams. New York: ACM Press, 2010: 45-50.
- [17] WEI G, WEI Z, HUANG L, et al. Robust underwater fish classification based on data augmentation by adding noises in random local regions[C] // Pacific Rim Conference on Multimedia. Berlin: Springer, 2018: 509-518.
- [18] JOLY A, GOËAU H, GLOTIN H, et al. LifeCLEF 2016: Multimedia life species identification challenges[C] // International Conference of the Cross-Language Evaluation Forum for European Languages. Berlin: Springer, 2016: 286-310.
- [19] CHOI S. Fish identification in underwater video with deep convolutional neural network: SNUMedinfo at LifeCLEF fish task 2015[C] // CLEF (Working Notes), 2015: 1-10.
- [20] JÄGER J, RODNER E, DENZLER J, et al. SeaCLEF 2016: Object proposal classification for fish detection in underwater videos[C] // CLEF (Working Notes), 2016: 481-489.
- [21] SHI C, JIA C, CHEN Z. FFDet: A fully convolutional network for coral reef fish detection by layer fusion[C] // 2018 IEEE International Conference on Visual Communications and Image Processing. Piscataway: IEEE Press, 2018: 1-4.
- [22] ZHUANG P Q, XING L J, LIU Y L, et al. Marine animal detection and recognition with advanced deep learning models[C] // Working Notes of the 8th International Conference of the CLEF Initiative, 2017: 1-9.
- [23] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C] // European Conference on Computer Vision. Berlin: Springer, 2016: 21-37.
- [24] 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.
- [25] JAISAKTHI S M, MIRUNALINI P, ARAVINDAN C. Coral reef annotation and localization using faster R-CNNs[C] // Working Notes of the 8th International Conference of the CLEF Initiative, 2019: 1-6.
- [26] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[C] // Advances in Neural Information Processing Systems, 2015: 91-99.
- [27] BOGOMASOV K, GRAWE P, CONRAD S. A two-staged approach for localization and classification of coral reef structures and compositions[C] // Working Notes of the 8th International Conference of the CLEF Initiative, 2019: 1-11.
- [28] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 779-788.
- [29] ZIVKOVIC Z. Improved adaptive Gaussian mixture model for background subtraction[C] // Proceedings of the 17th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE Press, 2004, 2: 28-31.
- [30] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C] // Advances in Neural Information Processing Systems, 2012: 1097-1105.
- [31] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 1-9.
- [32] CAI Z, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6154-6162.
- [33] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 658-666.
- [34] LAW H, DENG J. Cornernet: Detecting objects as paired keypoints[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 734-750.
- [35] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2117-2125.
- [36] ZHANG S, WEN L, BIAN X, et al. Single-shot refinement neural network for object detection[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4203-4212.
- [37] CORDTS M, OMRAN M, RAMOS S, et al. The cityscapes dataset for semantic urban scene understanding[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 3213-3223.
- [38] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2015-04-10) [2020-08-01]. <https://arxiv.org/abs/1409.1556>.

作者简介:

陈智能 男, 博士, 副研究员。主要研究方向: 多媒体内容分析与检索、医学影像分析。

史存存 女, 硕士, 工程师。主要研究方向: 计算机视觉、深度学习。

李轩涯 男, 博士。主要研究方向: 物联网计算、人工智能。

贾彩燕 女, 博士, 教授, 博士生导师。主要研究方向: 社会计算、文本聚类、网络社区发现。

黄磊 男,博士,副教授。主要研究方向:多媒体内容分析与检

索、计算机视觉。

An underwater coral reef fish detection approach based on aggregation of spatio-temporal features

CHEN Zhineng¹, SHI Cuncun², LI Xuanya^{3,*}, JIA Caiyan², HUANG Lei⁴

(1. Research Centre for Digital Content Technology and Services, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China;

2. School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China;

3. Baidu Inc., Beijing 100085, China; 4. College of Information Science and Engineering, Ocean University of China, Qingdao 266100, China)

Abstract: It is challenging to detect coral reef fish from underwater surveillance videos, due to issues like poor video imaging quality, complex underwater environment, high visual diversity of coral reef fish, etc. Extracting discriminative features to characterize the fishes has become a crucial issue that dominates the detection accuracy. This paper proposes an underwater coral reef fish detection method based on aggregation of spatio-temporal features. It is achieved by designing two modules for visual and temporal feature aggregation and fusing multi-dimensional features. The former designs a top-down partition and a bottom-up merging, which achieve effective aggregation of feature maps of different convolutional layers with varying resolutions. The latter devises a temporal feature fusion scheme based on the pixel difference between adjacent frames. It enhances the feature representation of moving objects and their surrounding area through the fusion of feature maps coming from adjacent frames. Experiments on a public dataset show that, by employing the spatio-temporal aggregation network built on top of the two proposed modules, we can effectively detect coral reef fishes in the challenging underwater environment. Higher detection accuracy are obtained compared with the existing methods and popular detection models.

Keywords: coral reef fish; convolutional neural network; spatio-temporal feature; object detection; feature fusion

Received: 2020-08-24; Accepted: 2020-09-19; Published online: 2020-10-29 14:22

URL: kns.cnki.net/kcms/detail/11.2625.V.20201029.1355.001.html

Foundation items: National Natural Science Foundation of China (61772526,61876016,61872326); Baidu Open Research Program

* **Corresponding author.** E-mail: lixuanya@baidu.com

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

DOI: 10.13700/j.bh.1001-5965.2020.0491

面向电力系统的多粒度隐患检测方法



徐晓华¹, 钱平², 王一达^{1,*}, 周昕悦¹, 徐汉麟¹, 徐李冰¹

(1. 国网浙江省电力有限公司杭州供电公司, 杭州 310020; 2. 国网浙江省电力有限公司, 杭州 310007)

摘 要: 由于电力系统的安全问题往往会造成严重的经济或社会影响, 隐患检测已成为电力系统不可或缺的重要环节。随着人工智能领域的发展, 基于深度学习的智能化电力系统隐患检测技术逐渐得到越来越多的关注。但目前的方法大多只是单一地考虑图像的全局特征或局部特征, 无法全面彻底表征图像, 进而难以捕捉电力领域尤其室外复杂背景下的隐患检测。为此, 基于深度学习技术, 提出了一种面向电力系统的多粒度隐患检测方法 MGNet。通过引入图像的多粒度信息, 构建全局和局部网络, 进行多粒度级检测; 并通过不同粒度级检测结果的协作式融合, 增强检测的全面性。在杆塔连接金具隐患和线路通道机械隐患 2 个数据集上进行了实验比较和分析, 对所提模型的检测性能进行评估。通过与现有最优隐患检测基准方法相比, 所提方法在 2 种不同数据集上的平均精度均值分别提升了 2.74% 和 2.77%, 验证了模型的有效性。

关 键 词: 隐患检测; 多粒度信息; 协作式融合; 深度学习; 电力系统

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2021)03-0520-11

电力资源是国民经济的基础和命脉, 与国计民生密切相关。然而, 由于受自然环境或人为因素等影响, 输电线路往往存在诸多隐患, 如输电杆塔中的销钉缺损和输电线路通道中的机械(如挖掘机)施工等, 这些隐患往往会造成巨大的经济损失和人身伤害。据国家能源局的相关数据统计, 2020 年第一季度全国发生多起电力事故, 其中电力设备事故直接经济损失高达 100 万元。因此, 隐患检测已成为电力系统亟待解决的重要问题。

传统的电力系统隐患检测主要采用人工检测的方式, 即通过检测人员肉眼观察电力场景的图片数据来检测电力系统的隐患问题。很明显, 这种人工检测的方式不仅需要耗费大量的人力和物力, 而且检测效率也十分低下; 同时, 检测的准确

率严重依赖于检测人员的业务水平。因而, 存在由于检测人员的视觉疲劳所导致的准确率降低问题。近年来, 随着计算机科学和人工智能领域的快速发展, 智能化电力系统隐患检测逐渐吸引了许多研究人员的注意^[1-4]。现有工作大多基于人为手工设计的特征和机器学习技术来进行隐患检测。由于人为手工设计的特征很大程度上依赖于人的先验知识, 使得这些方法难以准确捕捉图像丰富的语义信息^[5-6]。最近, 伴随神经网络技术的蓬勃兴起, 基于深度学习技术的计算机视觉算法在众多任务中展示了其强大的表示学习能力^[7-8]。因此, 本文旨在利用深度学习技术, 实现对电力系统隐患的精确检测。

然而, 基于深度学习技术的电力系统隐患检测任务是十分严峻的, 面临多项挑战。①电

收稿日期: 2020-09-02; 录用日期: 2020-09-04; 网络出版时间: 2020-11-05 09:28

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201104.0931.002.html

基金项目: 国网浙江省电力有限公司科技项目(5211HZ19014U)

* 通信作者. E-mail: wang_yida@zj.sgcc.com.cn

引用格式: 徐晓华, 钱平, 王一达, 等. 面向电力系统的多粒度隐患检测方法[J]. 北京航空航天大学学报, 2021, 47(3): 520-530.
XU X H, QIAN P, WANG Y D, et al. Multi-granularity hazard detection method for electrical power system[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 520-530 (in Chinese).

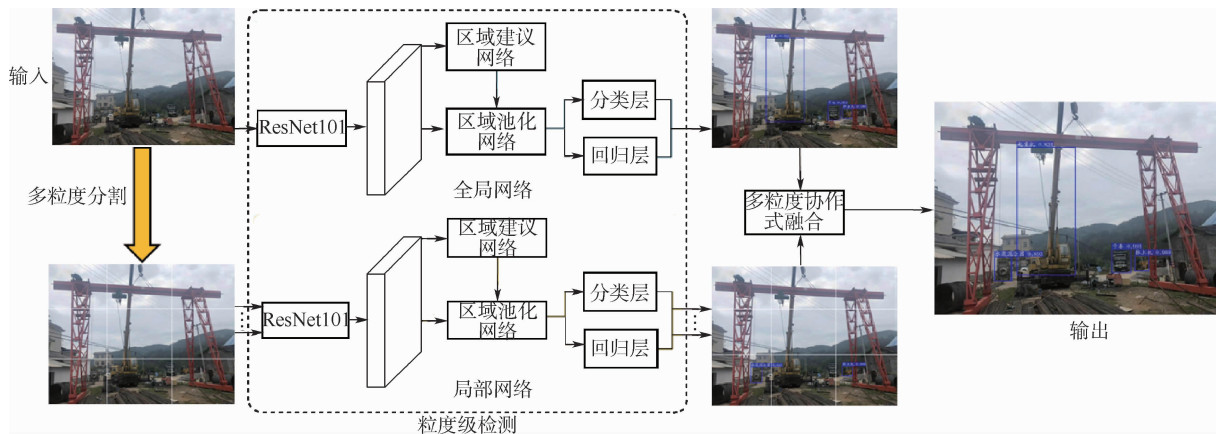


图1 多粒度隐患检测网络结构

Fig. 1 Structure of multi-granularity hazard detection network

为多粒度隐患检测网络对于给定图像 I 的隐患检测结果集合, 包含图像 I 中各隐患物体的类别信息和位置信息, θ_F 代表网络 $\mathcal{F}$ 待学习的相关参数。值得一提的是, 本文提出的多粒度隐患检测网络具有普适性, 可迁移到多种不同的检测任务中。

2.2 多粒度分割

为了全面表征图像, 本文对图像进行多粒度分割, 分别构建全局网络和局部网络, 进而学习图像的全局表示与局部表示。为使文章简洁, 暂时忽略隐患物体的下标索引。

2.2.1 全局网络

本文先从全局出发, 基于完整图像, 构建全局网络, 获得图像的全局表示。通过添加跳跃连接, 深度残差网络^[11] (ResNet) 可以捕捉到浅层网络的信息, 解决网络退化问题, 缓解梯度消失现象, 在计算机视觉领域多项任务中取得出色的效果^[12-13]。因此, 本文使用 ResNet 网络 $\mathcal{R}_g$ 来学习图像的全局表示。得到的图像全局特征图如下:

$$\tilde{\mathbf{V}} = \mathcal{R}_g(I | \theta_{\mathcal{R}_g}) \quad (1)$$

式中: $\theta_{\mathcal{R}_g}$ 为 ResNet 网络 $\mathcal{R}_g$ 的参数; $\tilde{\mathbf{V}} \in \mathbf{R}^{H_g \times W_g \times C_g}$ 为图像 I 的全局特征图表示, H_g 和 W_g 分别为图像全局特征图的长和宽尺寸, C_g 为特征图的通道个数。

2.2.2 局部网络

全局网络侧重图像的全局显著性特征, 而易忽略图像的局部细节特征, 不利于模型对于小型或远距离隐患的准确检测。因此, 本文在传统全局网络的基础上引入局部网络, 以提取更加具体的局部图像特征, 进而增强图像整体的隐患检测性能。

图像切块是一种不依赖辅助模型、简单且高效的方法^[14]。本文引入切块尺度因子 m , 对输入图像 I 采用 $m \times m$ 的网格划分, 得到图像网格划分集合 $\mathcal{J} = \{I'_1, I'_2, \dots, I'_{m \times m}\}$ 。将 $\mathcal{J}$ 中各网格图像 $I'_r, r \in \{1, 2, \dots, m \times m\}$ 依次送入 ResNet 网络 $\mathcal{R}_l$, 学习其特征图。具体公式如下:

$$\hat{\mathbf{V}}_r = \mathcal{R}_l(I'_r | \theta_{\mathcal{R}_l}) \quad (2)$$

式中: $\theta_{\mathcal{R}_l}$ 为 ResNet 网络 $\mathcal{R}_l$ 的参数; $\hat{\mathbf{V}}_r \in \mathbf{R}^{H'_r \times W'_r \times C'_r}$ 为图像 I 在第 r 个网格中局部特征图表示, H'_r 和 W'_r 分别为图像第 r 个局部特征图的长和宽尺寸, C'_r 为相应的通道个数。

2.3 粒度级检测

传统的基于区域推荐的目标检测方法, 如 R-CNN、Fast R-CNN 等框架, 大都采用选择性搜索算法^[15] 产生候选区域, 时间复杂度较高。针对此现象, Faster R-CNN 提出区域建议网络和区域池化网络, 实现了端到端的目标检测, 有效地平衡了目标检测任务的速度和精度^[16], 在众多检测任务中取得了杰出的表现^[17-18]。因此, 本文基于不同粒度的图像特征图, 即 $\hat{\mathbf{V}}$ 和 $\{\hat{\mathbf{V}}_1, \hat{\mathbf{V}}_2, \dots, \hat{\mathbf{V}}_{m \times m}\}$, 采用 Faster R-CNN 方法实现对图像中包含隐患物体的准确检测识别。具体公式如下:

$$\begin{cases} \mathcal{O}_g = \mathcal{T}_g(\tilde{\mathbf{V}} | \theta_{\mathcal{T}_g}) \\ \mathcal{O}_l = \mathcal{T}_l(\hat{\mathbf{V}}_1 | \theta_{\mathcal{T}_l}) \cup \mathcal{T}_l(\hat{\mathbf{V}}_2 | \theta_{\mathcal{T}_l}) \cup \dots \cup \mathcal{T}_l(\hat{\mathbf{V}}_{m \times m} | \theta_{\mathcal{T}_l}) \end{cases} \quad (3)$$

式中: $\mathcal{T}_g$ 为全局粒度级检测网络; $\theta_{\mathcal{T}_g}$ 为网络 $\mathcal{T}_g$ 的相关待学习参数; $\mathcal{O}_g = \{\tilde{o}_1, \tilde{o}_2, \dots, \tilde{o}_{N_g}\}$ 为图像 I 的全局隐患物体检测集合, 包含 N_g 个全局隐患物体, 每个隐患物体包含 2 部分信息, 即 $\tilde{o}_j = \{\tilde{\mathbf{y}}_j, \tilde{\mathbf{t}}_j\}$, $\tilde{\mathbf{y}}_j = (\tilde{y}_j^1, \tilde{y}_j^2, \dots, \tilde{y}_j^C) \in \mathbf{R}^C$ 为隐患物体 $\tilde{o}_j$ 的类

别向量, $\tilde{t}_j = (\tilde{x}_j, \tilde{z}_j, \tilde{w}_j, \tilde{e}_j)$ 为隐患物体 $\tilde{o}_j$ 的位置向量; $\mathcal{T}_1$ 为局部粒度级检测网络; $\theta_{\mathcal{T}_1}$ 为 $\mathcal{T}_1$ 网络的相关参数; $\mathcal{O}_1 = \{\hat{o}_1, \hat{o}_2, \dots, \hat{o}_{N_1}\}$ 为图像 I 的局部隐患物体检测集合, 包含 N_1 个局部隐患物体, 每个局部隐患物体表示为 $\hat{o}_k = \{\hat{y}_k, \hat{t}_k\}$, $k = \{1, 2, \dots, N_1\}$, $\hat{y}_k = (\hat{y}_k^1, \hat{y}_k^2, \dots, \hat{y}_k^C) \in \mathbf{R}^C$ 为隐患物体 $\hat{o}_k$ 的预测类别向量, $\hat{t}_k = (\hat{x}_k, \hat{z}_k, \hat{w}_k, \hat{e}_k)$ 为隐患物体 $\hat{o}_k$ 的预测位置向量。

在本文工作中, 不论是全局目标检测网络还是局部目标检测网络, 均由区域建议网络和区域池化网络 2 部分构成。

2.3.1 区域建议网络

区域建议网络本质上是卷积神经网络, 可以对输入图像对应生成高质量的区域建议框。通过与多粒度分割中的全局网络和局部网络共享全局和局部特征图, 缓解了区域选择产生的高耗时问题^[4]。基于式(1)和式(2)生成的全局和局部特征图, 即 $\tilde{\mathbf{V}}$ 和 $\{\hat{\mathbf{V}}_1, \hat{\mathbf{V}}_2, \dots, \hat{\mathbf{V}}_{m \times m}\}$, 区域建议网络使用 $n \times n$ 滑动窗口对特征图以固定步长进行滑动卷积, 同时引入锚框机制, 预先定义 a 个不同尺度和比例的锚框, 以滑动窗口中心点为基准, 匹配预先定义锚框作为初始候选框。对于初始候选框所对应的检测区域, 构建分类层和回归层。分类层对检测区域进行初步判断, 得到其为隐患物体和背景区域的概率。回归层旨在调整边框位置, 得到检测区域回归偏移量, 进而拟合预测物体的位置坐标。基于 a 个锚框, 分类层可以得到 $2a$ 个概率评分, 对应每个锚框为隐患物体和背景区域的概率。回归层得到 $4a$ 个锚框坐标编码。图 2 为基于全局特征图 $\tilde{\mathbf{V}}$ 的区域建议网络结构。为缓解检测区域冗余问题, 本文方法进一步采用非极大值抑制方法^[6], 基于检测区域与实际检测目标的

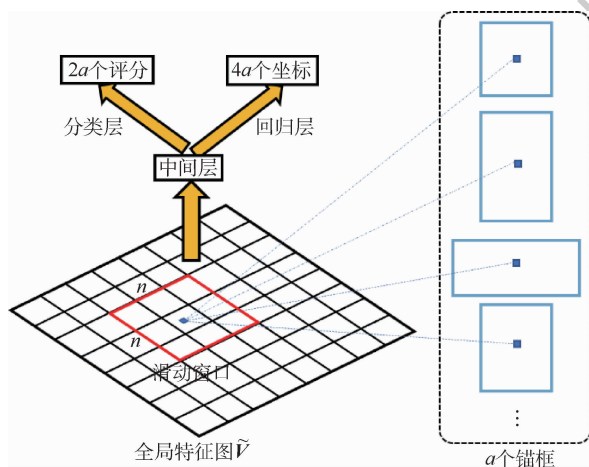


图2 区域建议网络结构

Fig.2 Structure of region proposal network

重叠面积及检测区域之间的重叠面积, 过滤多余的检测区域。按照分类层概率评分, 得到指定数目的候选区域, 从而实现对检测区域的初步定位。

2.3.2 区域池化网络

全局和局部特征经过区域建议网络后, 可以得到一系列目标物体的候选区域。由于候选区域的尺寸大小不一, 本文采用单尺度池化操作, 先将候选区域分割成相同大小尺寸的分块, 对每个分块进行最大池化操作, 从而将大小不一的候选区域转换为固定尺寸特征图, 再使用卷积神经网络, 对固定尺寸特征图进行编码, 得到隐患物体的具体分类及精确位置。

2.4 模型优化

多粒度隐患检测网络的代价函数分为全局代价函数和局部代价函数。以全局代价函数为例, 具体过程如下:

$$L(\{\tilde{y}_j^c\}, \{\tilde{t}_j^c\}) = \frac{1}{N_{\text{cls}}} \sum_{j=1}^{N_g} \sum_{c=1}^C L_{\text{cls}}(y_j^c, \tilde{y}_j^c) + \rho \frac{1}{N_{\text{loc}}} \sum_{j=1}^{N_g} \sum_{c=1}^C y_j^c L_{\text{loc}}(t_j, \tilde{t}_j) \quad (4)$$

式中: N_{cls} 为参与训练锚框的数目; N_{loc} 为参与训练锚框中正确类别锚框的数目; ρ 为分类代价和位置回归代价的调节平衡参数。

代价函数中的 2 项分别对应隐患物体的分类代价和位置回归代价, 其中分类代价函数为

$$L_{\text{cls}}(y_j^c, \tilde{y}_j^c) = -\log_2[y_j^c \tilde{y}_j^c + (1 - y_j^c)(1 - \tilde{y}_j^c)] \quad (5)$$

位置回归代价函数为

$$L_{\text{loc}}(t_j, \tilde{t}_j) = R(\tilde{t}_j - t_j) \quad (6)$$

式中: R 为鲁棒性代价函数。

具体方法如下:

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & |x| > 1 \\ |x| - 0.5 & \text{其他} \end{cases} \quad (7)$$

回归代价函数通过与 y_j^c 相乘, 只对正确类别的隐患物体进行位置回归。值得一提的是, 针对局部目标检测网络的优化, 需要根据每个网格区域的图像信息, 重新定义相应的隐患类别真实值向量和隐患区域位置信息。

2.5 多粒度协作式融合

为增强对于图像 I 的隐患检测效果, 本文设计多粒度协作式融合图像全局网络检测结果 $\mathcal{O}_g$ 和局部网络检测结果 $\mathcal{O}_l$ 。

以全局隐患物体 $\tilde{o}_j$ 和局部隐患物体 $\hat{o}_k$ 为例 (这里需要先将局部隐患物体 $\hat{o}_k$ 映射回全局图像), 具体融合操作如下:

$$\begin{cases} \tilde{o}_j & \tilde{o}_j \cap \mathcal{O}_1 = \emptyset \\ \hat{o}_k & \hat{o}_k \cap \mathcal{O}_g = \emptyset \\ \tilde{o}_j \cup \hat{o}_k & 0 < S_j^k < \lambda \\ \hat{o}_k & \lambda \leq S_j^k \end{cases} \quad (8)$$

式中: S_j^k 为 $\tilde{o}_j$ 和 $\hat{o}_k$ 面积的交并比, 即 $\tilde{o}_j$ 和 $\hat{o}_k$ 面积的交集和并集的比例; λ 为预设的阈值。

据式(8)可知, 若图像全局隐患物体 $\tilde{o}_j$ 与局部隐患物体集合 $\mathcal{O}_1$ 交集为空集, 则保留 $\tilde{o}_j$ 。相似地, 若图像局部隐患物体 $\hat{o}_k$ 与全局隐患物体集合 $\mathcal{O}_g$ 交集为空集, 则保留 $\hat{o}_k$ 。若与全局隐患物体 $\tilde{o}_j$ 有重叠, 且交并比为 S_j^k , 则根据预设阈值 λ , 进行保留或合并。具体地, 若 $\lambda > S_j^k$, 则将 2 个检测结果均保留; 若 $\lambda \leq S_j^k$, 则两者视为同一物体。

通过对图像全局检测和局部检测结果进行多粒度协作式融合, 多粒度隐患检测网络可以得到图像 I 的最终隐患物体检测集合 $\tilde{\mathcal{H}}_I$ 。多粒度隐患检测网络具体过程如算法 1 所示。

算法 1 多粒度隐患检测网络。

输入: 图像 I 。

输出: 图像 I 的最终隐患物体检测集合 $\tilde{\mathcal{H}}_I$ 。

1. 初始化网络参数 θ_{R_g} 、 θ_{R_1} 、 θ_{T_g} 、 θ_{T_1} 。
2. REPEAT:
3. 根据式(1)学习图像全局表示。
4. 根据式(3)对图像进行全局粒度级隐患检测, 得到全局网络检测结果集合 $\mathcal{O}_g$ 。
5. 根据全局代价函数优化全局网络参数。
6. UNTIL: 全局代价函数收敛
7. 对图像 I 进行多粒度分割, 得到图像网格划分集合 $\mathcal{J}$ 。
8. REPEAT:
9. FOR $r = 1, 2, \dots, m \times m$
10. 根据式(2)学习图像每个网格 I'_r 的表示。
11. 根据式(3)对图像进行局部粒度级隐患检测, 得到局部网络检测结果 $\mathcal{O}_1$ 。
12. 根据局部代价函数优化局部网络参数。
13. END FOR
14. UNTIL: 局部代价函数收敛
15. 对隐患物体集合 $\mathcal{O}_g$ 和 $\mathcal{O}_1$ 进行多粒度协作式融合, 得到图像 I 的最终隐患物体检测集合 $\tilde{\mathcal{H}}_I$ 。

3 实验结果与分析

3.1 数据集

为全面验证模型效果及鲁棒性, 本文引入 2 个真实世界数据集: ①杆塔连接金具隐患数据集; ②线路通道机械隐患数据集, 分别侧重于小尺寸和大尺寸隐患检测。其中, 杆塔连接金具隐患数据集主要包含 2 种隐患类别: 销钉缺损和销钉安装位置错误; 线路通道机械隐患数据集主要包含 7 种隐患类别: 卡车、推土机、起重机、挖掘机、泵车、水泥混合器、打桩机。本文雇佣 10 名专业技术人员, 使用智能化标注工具 LabelImg 对数据集中的隐患物体进行标注。2 个数据集最终规模分别如表 2 和表 3 所示。

表 2 杆塔连接金具隐患数据集规模
Table 2 Statistics of tower connection fitting hazard datasets

隐患类别	隐患物体数目
销钉缺损	6 232
销钉安装位置错误	3 272

表 3 线路通道机械隐患数据集规模
Table 3 Statistics of transmission line channel mechanical hazard datasets

隐患类别	隐患物体数目
卡车	7 073
推土机	7 579
起重机	15 796
挖掘机	18 568
泵车	4 808
水泥混合器	2 449
打桩机	1 336

3.2 实验设置

本文实验环境搭建在含有 4 块 TITAN Xp GPU 的服务器中。本文提出的多粒度隐患检测网络基于 Tensorflow 和 PaddlePaddle 框架实现, 使用 Momentum 优化器优化网络参数, 并采用目标检测任务中常用的评价指标——平均精度均值 mAP 来评价模型。其中, 学习速率被设置为 0.001, 动量参数被设置为 0.9。此外, 本文采用网格搜索策略来决定网络中超参数的最优值。具体地, 式(4)中分类代价和位置回归代价的调节平衡参数 ρ 取值为 10, 式(8)中阈值 λ 在杆塔连接金具隐患数据集和线路通道机械隐患数据集取值分别为 0.6 和 0.5。输入图像的尺寸预先经过缩放处理和均值中值化处理后为 $1\,400 \times 1\,800$ 。对数据集数据进行分割, 其中 80% 作为训练集, 10% 作为验证集, 10% 作为测试集。通过多次迭

代训练调优,得到最佳网络模型,代价函数收敛如图 3 所示。可知,代价函数开始时下降迅速,后期逐渐趋于稳定。这一现象验证了本文模型良好的收敛性。此外,给定一张图片,MGNet 网络对其进行隐患检测所花费的时间为 0.998 8 s。因此,本文模型检测效率较高,具有可行性。

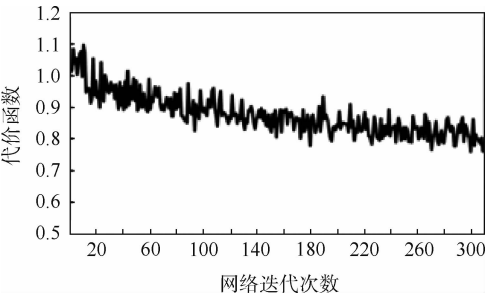


图 3 代价函数收敛图
Fig.3 Cost function convergence graph

3.3 结果分析

选择以下基准方法作为比较,来评估本文模型,方法描述如下:

1) SSD^[19]:不同于构建的多粒度隐患检测网络,该方法通过学习图像的特征金字塔,基于图像多个大小不同的特征图进行检测,解决图像数据中所含物体尺度多样性问题。

2) Faster R-CNN^[20]:该方法可以通过去掉本文模型的局部网络而得到。和本文模型相比,该方法只使用图像的全局特征表示。

3) YOLOv3^[21]:该方法将图像全局表示特征图划分为多个网格单元,基于待检测物体的中心坐标位置,只选择特定网格单元进行目标检测。

4) Cascade R-CNN^[22]:该方法针对区域建议网络中的回归层,提出级联多个具有不同阈值的回归层,并通过多次调整边框位置,以提高检测边框准确度。

5) YOLOv4^[23]:该方法基于 YOLOv3 进行改进,通过引入数据扩增策略,为数据集扩增不同尺寸的目标物体。

表 4 展示了本文模型 MGNet 和不同方法对于 mAP 的对比结果。通过观察表 4 对比结果,可以得到以下结论:①在杆塔连接金具隐患数据集和线路通道机械隐患数据集中,SSD 方法均取得了最差的效果,这一现象可能是因为:SSD 方法直接基于特征图进行隐患类别检测,舍弃了初步区分隐患物体和背景区域的步骤,致使检测过程噪声较大,无法准确捕捉隐患物体。②相比于其他基准方法,Faster R-CNN 在 2 个数据集中的性能表现相对变化最为明显。这一现象说明,仅基于

图像全局特征进行检测会使模型对于隐患物体尺寸较为敏感,从而具有不稳定性。③所有的基准方法在线路通道机械隐患数据集中的表现都要优于其在杆塔连接金具隐患数据集中的表现。这一现象可能是因为:线路通道机械隐患数据集中的隐患物体大多为卡车等大尺寸物体,相对于杆塔连接金具隐患数据集中的小尺寸销钉隐患易于区分。④本文提出的 MGNet 模型在 2 个领域数据集上都取得了比其他先进方法更好的检测效果,验证了模型 MGNet 的有效性及将图像的多粒度信息引入到隐患检测过程的必要性。

为了更深入地理解 MGNet,进一步研究了 MGNet 在不同隐患类别上的表现。根据表 4 中展示的结果,选择在杆塔连接金具隐患数据集和线路通道机械隐患数据集上最好的基准方法 YOLOv4 作为该实验的对比方法。在图 4 中,YOLOv4 对于销钉缺损类别的表现略高于本文模型。在图 5 中,MGNet 和 YOLOv4 在杆塔连接金具隐患数据集中的 2 个隐患类别检测效果差别较大,即对于销钉缺损的检测效果远远高于对销钉安装位置错误的检测。一个可能的解释是隐患物体形态缺损的特征比位置变动的特征更加明显,因而易于捕捉。

表 4 不同隐患检测方法在 2 个数据集上的平均精度均值
Table 4 Mean average precision of different hazard detection methods on two datasets

模型	mAP/%	
	杆塔连接金具 隐患数据集	线路通道机械 隐患数据集
SSD	19.99	60.93
Faster R-CNN	38.96	71.65
YOLOv3	40.59	61.32
Cascade R-CNN	41.21	68.50
YOLOv4	44.99	72.34
MGNet	47.73	75.11

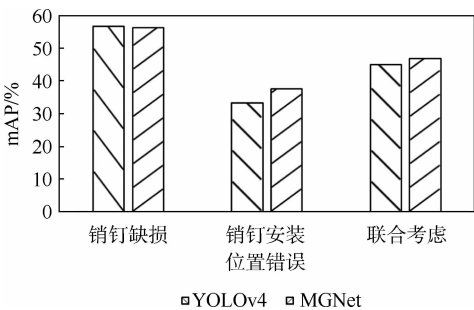


图 4 MGNet 和 YOLOv4 在杆塔连接金具隐患数据集中各隐患类别表现
Fig.4 Performance of MGNet and YOLOv4 on different categories of tower connection fitting hazard datasets

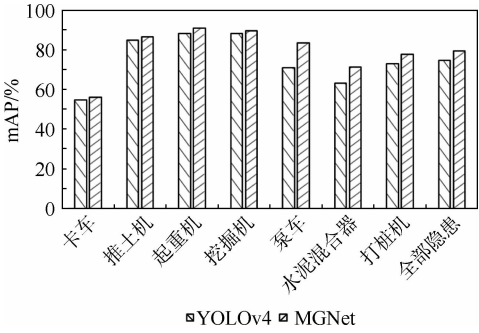


图5 MGNet 和 YOLOv4 在线路通道机械隐患数据集中各隐患类别表现

Fig. 5 Performance of MGNet and YOLOv4 on different categories of transmission line channel mechanical hazard datasets

3.4 消融实验

为了验证多粒度隐患检测网络中多粒度协作式融合的有效性,引入 2 个衍生模型:MGNet_NoGlobal 和 MGNet_NoLocal。在这 2 个衍生模型中均去掉了多粒度的协作式融合操作,而分别只考虑了局部网络和全局网络。

表 5 给出了本文模型和这 2 个衍生模型基于 2 个数据集的实验结果对比。可以看出,本文模型 MGNet 在 2 个数据集中的表现均高于这 2 个衍生模型。这一实验结果说明,通过全局网络和局部网络的协作式融合,可增强图像中所含物体检测效果,有利于更加精确的隐患检测。从某种意义上说,这是因为:全局网络可以捕获图像的整体特征,而局部网络则侧重于图像局部区域细节特征。两者互相促进,均不可或缺。除此之外还观察到,在杆塔连接金具隐患数据集中, MGNet_NoGlobal 比 MGNet_NoLocal 的效果要好。这说明在面向杆塔连接金具的隐患检测中,局部网络比全局网络的贡献大。这可能是因为:相对于全局网络侧重于全局特征的提取,局部网络更加注重图像中的局部区域,从而使小尺寸隐患物体的检测结果更加精准。相反地,在线路通道机械隐患数据集中, MGNet_NoLocal 比 MGNet_NoGlobal 的效果要好。这说明在线路通道机械隐患数据集中,全局网络对于隐患检测的贡献大于局部网络。

表 5 多粒度隐患检测网络消融实验

Table 5 Ablation experiments of multi-granularity hazard detection network

模型	mAP/%	
	杆塔连接金具 隐患数据集	线路通道机械 隐患数据集
MGNet	47.73	75.11
MGNet_NoGlobal	46.87	41.24
MGNet_NoLocal	38.96	71.65

背后的原因可能是:对于大尺寸隐患物体,只引入局部网络,可能会破坏隐患物体的完整性,从而影响局部网络的检测结果。

为了更深入地理解全局网络和局部网络的性能表现,进一步研究了 MGNet 及 2 个消融衍生模型 (MGNet_NoGlobal 和 MGNet_NoLocal) 在不同隐患类别上的表现。如图 6 所示,本文模型 MGNet 对于线路通道机械隐患数据集所有类别检测的精度均高于 2 个消融衍生模型,验证了本文模型 MGNet 的有效性。在图 7 中, MGNet_NoGlobal 模型对于销钉安装位置错误类别的检测表现略高于模型 MGNet,这可能是因为:局部网络注重图像中的局部区域,对位置变换更加敏感。同时发现,对于线路通道机械隐患数据集中的各隐患类别 (见图 6), MGNet_NoLocal 的表现均优于 MGNet_NoGlobal,这一现象进一步验证了只引入局部网络会损坏大尺寸隐患物体完整性,从而造成隐患检测准确性下降的结论。

为了更加直观地理解本文所提的多粒度协作式融合,给出了 MGNet 及其 2 个消融衍生网络 (MGNet_NoGlobal 和 MGNet_NoLocal) 在线路通

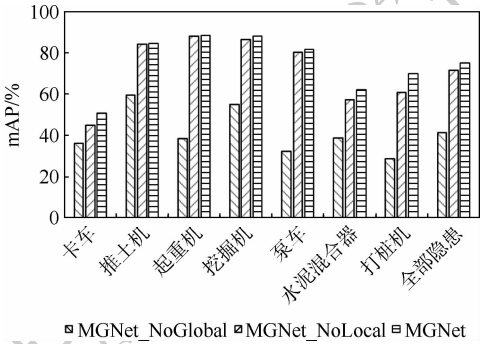


图 6 MGNet 及其消融衍生模型在线路通道机械隐患数据集中各隐患类别表现

Fig. 6 Performance of MGNet and its ablation derived models on different categories of transmission line channel mechanical hazard datasets

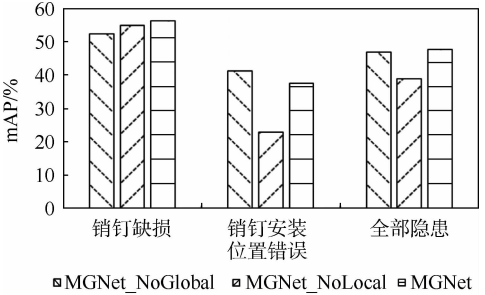


图 7 MGNet 及其消融衍生模型在杆塔连接金具隐患数据集中各隐患类别表现

Fig. 7 Performance of MGNet and its ablation derived models on different categories of tower connection fitting hazard datasets

道机械隐患检测方面的测试结果,如图 8 和图 9 所示(图 9 为图 8 中检测结果的局部放大图)。其中,局部网络的切块尺度因子 $m = 3$ 。基于图 8 和图 9 可以看出,全局网络检测可以捕捉到图片中较为显著的起重机、卡车和推土机隐患,局部网络可以检测到水泥混合器和推土机隐患。而多粒度隐患检测网络通过协作式融合,可以全面检测到这 4 种不同的隐患。由此可见,全局网络侧重于检测图像中比较显著的特征,可以捕捉到起重机等大型隐患物体,但是忽略了水泥混合器这类

不显著隐患物体。局部网络可以充分捕捉图像中水泥混合器等不显著隐患物体,但是与此同时分割损害了起重机这类大型隐患物体的特征,致使无法充分捕捉这类大型隐患物体。多粒度隐患检测网络检测比较全面,可以综合考虑图像整体与局部特征,全面对图像中的物体进行检测。从图 8 和图 9 所示的不同网络检测结果可以直观看出,相对于消融衍生网络 MGNet_NoGlobal 和 MGNet_NoLocal, MGNet 在隐患检测方面具有一定的优势。



图 8 消融实验检测结果对照图 ($m = 3$)
Fig. 8 Comparison of detection results from ablation experiment ($m = 3$)

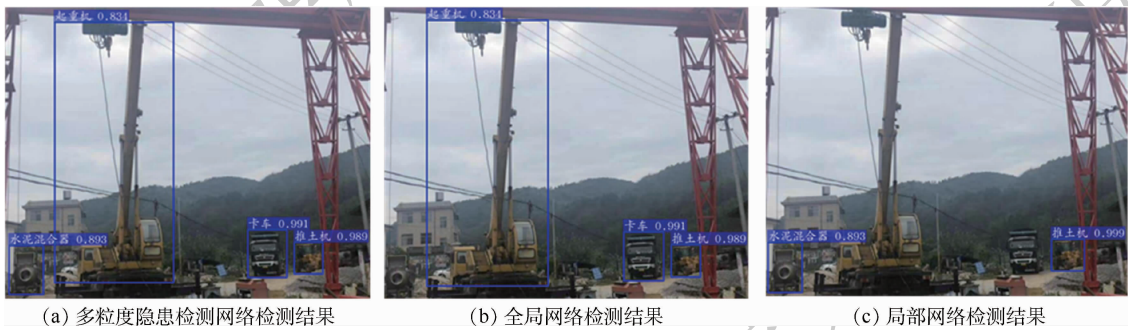


图 9 图 8 检测结果放大图
Fig. 9 Magnification of detection results in Fig. 8

3.5 参数实验

本节对本文模型 MGNet 进行了敏感度分析。特别地,研究了切块尺度因子 m 和融合阈值 λ 对于模型隐患检测效果的影响。

3.5.1 切块尺度因子

图 10 为本文模型 MGNet 在 2 个数据集中取不同切块尺度因子时对应的平均精度均值。本文将切块尺度因子 m 设置在 $[1, 5]$ 范围内,并以 1 为步长进行调节。从图 10 可以看出,整体而言,当模型 MGNet 对于切块尺度因子的敏感性较低,表现比较稳定。同时,当 $m = 2$ 时,模型 MGNet 在 2 个数据集中取得最优的效果。值得注意的是,当 $m = 1$ 时,即不对图像进行细粒度切割时,模型 MGNet 在 2 个数据集中均取得最差的效果。这一现象再次验证了多粒度切割电力图像的必要性。特别地,当切块尺度因子逐渐增大接近 5 时,模型

在 2 个数据集集中的效果表示均有所下降,这一现象表明盲目追求细粒度切割会损害隐患物体的特征,从而无法准确表示隐患物体,造成检测效果的下降。

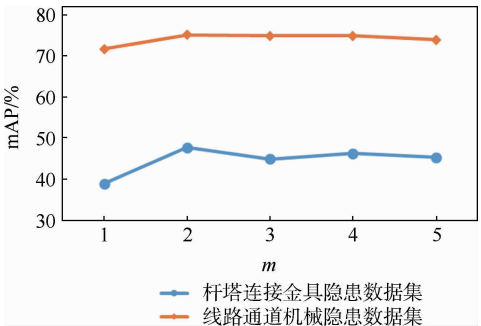


图 10 切块尺度因子在两个数据集中对模型 MGNet 平均精度均值的影响
Fig. 10 Effect of block scale factor on mean average precision of MGNet model in two datasets

3.5.2 融合阈值

为了探索多粒度协作式融合对于模型隐患检测效果的影响,对于融合阈值 λ 进行了敏感度分析。本文将融合阈值 λ 设置在 $[0.1, 0.7]$ 范围内,并以0.1为步长进行调节。如图11所示,当 $\lambda = 0.6$ 时,模型在杆塔连接金具隐患数据集中取得最优的效果,然而,当 λ 从0.6增长到0.7时,效果急剧下降。同时,如图12所示,当 $\lambda = 0.5$ 时,模型在线路通道机械隐患数据集中具有最好的性能,当 λ 从0.5增长到0.7时,模型性能也出现下降趋势。根据式(8),当 λ 趋近于0时,模型对于全局网络和局部网络具有重合面积的检测结果,会忽略全局网络检测结果,只考虑局部网络检测结果。当 λ 趋近于1时,模型会保留2个网络的全部检测结果。对应地,基于实验现象可以得出以下结论,忽略具有重合面积时的全局网络检测结果,会导致隐患检测不完全,但盲目融合2个网络的检测结果,会造成检测噪声过大,从而引起模型性能的下降。因此,协作式融合图像全局网络和局部网络的检测结果是至关重要的。

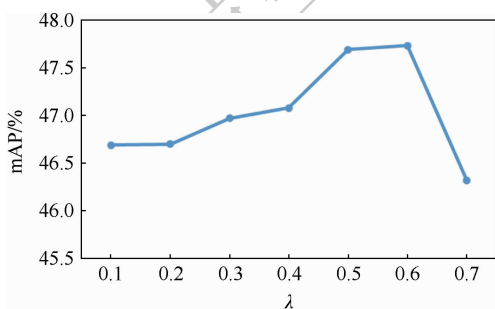


图11 融合阈值在杆塔连接金具隐患数据集中对模型MGNet平均精度均值的影响

Fig. 11 Effect of fusion threshold on mean average precision of MGNet model in tower connection fitting hazard datasets

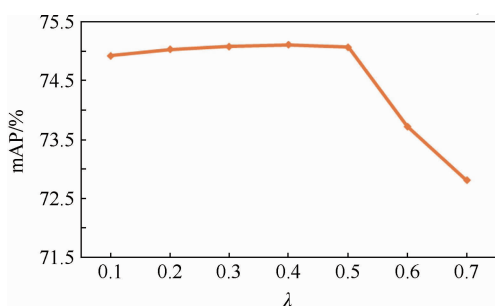


图12 融合阈值在线路通道机械隐患数据集中对模型MGNet平均精度均值的影响

Fig. 12 Effect of fusion threshold on mean average precision of MGNet model in transmission line channel mechanical hazard datasets

4 结论

1) MGNet 可实现较为优异的电力系统潜在隐患检测性能,在2个不同电力场景(杆塔连接金具隐患和线路通道机械隐患)的真实数据集,平均精度均值分别达到47.73%和75.11%,优于其他最优隐患检测基准方法。

2) MGNet 通过引入局部网络学习图像的局部细节表示,有利于提高模型效果,尤其是针对小尺寸隐患物体(如销钉)的检测。

3) MGNet 对于局部网络中切块尺度因子的参数敏感性较低,具有良好的稳定性。

MGNet 目前着眼于图像的多粒度信息融合。在未来的工作中,将尝试对比多种基本目标检测方法,引入时间、声音、红外等信息,以更全面立体的方式实现对电力系统的隐患检测。

参考文献 (References)

- [1] ZHAI Y J, WANG D, ZHANG M L, et al. Fault detection of insulator based on saliency and adaptive morphology[J]. Multimedia Tools and Applications, 2017, 76(9): 12051-12064.
- [2] PRASAD P S, RAO B P. LBP-HF features and machine learning applied for automated monitoring of insulators for overhead power distribution lines[C] // Proceedings of the International Conference on Wireless Communications, 2016: 808-812.
- [3] NORDENG I E, HASAN A, OLSEN D, et al. DEBC detection with deep learning[C] // Proceedings of the Scandinavian Conference. Berlin: Springer, 2017: 248-259.
- [4] 王万国, 田兵, 刘越, 等. 基于RCNN的无人机巡检图像电力小部件识别研究[J]. 地球信息科学, 2017, 19(2): 256-263. WANG W G, TIAN B, LIU Y, et al. Study on the electrical devices detection in UAV images based on region based convolutional neural networks[J]. Journal of Geo-information Science, 2017, 19(2): 256-263 (in Chinese).
- [5] 柴玉梅, 员武莲, 王黎明, 等. 基于双注意力机制和迁移学习的跨领域推荐模型[J]. 计算机学报, 2020, 43(9): 1-20. CHAI Y M, YUAN W L, WANG L M, et al. A cross-domain recommendation model based on dual attention mechanism and transfer learning[J]. Chinese Journal of Computers, 2020, 43(9): 1-20 (in Chinese).
- [6] 严星, 尤洪峰. Faster-RCNN 电力输送塔检测算法[J]. 计算机仿真, 2020, 37(2): 135-139. YAN X, YOU H F. Target detection algorithm for Faster-RCNN power transmission tower[J]. Computer Integrated Manufacturing Systems, 2020, 37(2): 135-139 (in Chinese).
- [7] 陈耀东, 李仁发, 李实英, 等. 面向目标检测与姿态估计的联合文法模型[J]. 计算机学报, 2014, 37(10): 2206-2217. CHEN Y D, LI R F, LI S Y, et al. A combined grammar for object detection and pose estimation[J]. Chinese Journal of Computers, 2014, 37(10): 2206-2217 (in Chinese).
- [8] 张冬明, 靳国庆, 代锋, 等. 基于深度融合的显著性目标检测

- 算法[J]. 计算机学报, 2019, 42(9): 2076-2086.
- ZHANG D M, JIN G Q, DAI F, et al. Salient object detection based on deep fusion of hand-crafted features [J]. Chinese Journal of Computers, 2019, 42(9): 2076-2086 (in Chinese).
- [9] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [C] // Proceedings of the International Conference on Learning Representation, 2015.
- [10] GIRSHICK R B. Fast R-CNN [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1440-1448.
- [11] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 770-778.
- [12] 陆永帅, 李元祥, 刘波, 等. 基于深度残差网络的高光谱遥感数据霾监测[J]. 光学学报, 2017, 37(11): 314-324.
- LU Y S, LI Y X, LIU B, et al. Hyperspectral data haze monitoring based on deep residual network [J]. Acta Optica Sinica, 2017, 37(11): 314-324 (in Chinese).
- [13] 朱超平, 杨艺. 基于YOLO2和ResNet算法的监控视频中的人脸检测与识别[J]. 重庆理工大学学报(自然科学), 2018, 32(8): 170-175.
- ZHU C P, YANG Y. Face detection and recognition in monitoring video based on YOLO2 and ResNet algorithm [J]. Journal of Chongqing University of Technology (Natural Science), 2018, 32(8): 170-175 (in Chinese).
- [14] 徐龙壮, 彭力, 朱凤增. 多任务金字塔重叠匹配的行人重识别方法[J/OL]. 计算机工程, 2020(2020-02-11) [2020-08-01]. <https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=JSJC20200209002&v=fpDLQvPDFGfsJ7%25mmd2BIY85YejXOe%25mmd2B%25mmd2FpNa24x8p%25mmd2FiBh3e5i3ALZRcuH3lF0Q0IM4Rgp>.
- XU L Z, PENG L, ZHU F Z. Person re-identification method based on multi-task pyramid overlapping matching [J/OL]. Computer Engineering, 2020(2020-02-11) [2020-08-01]. <https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ&dbname=CAPJLAST&filename=JSJC20200209002&v=fpDLQvPDFGfsJ7%25mmd2BIY85YejXOe%25mmd2B%25mmd2FpNa24x8p%25mmd2FiBh3e5i3ALZRcuH3lF0Q0IM4Rgp> (in Chinese).
- [15] UIJLINGS J R R, VAN DE SANDE K E A, GEVERS T, et al. Selective search for object recognition [J]. International Journal of Computer Vision, 2013, 104(2): 154-171.
- [16] 李文璞, 谢可, 廖道, 等. 基于Faster RCNN变电设备红外图像缺陷识别方法[J]. 南方电网技术, 2019, 13(12): 79-84.
- LI W P, XIE K, LIAO X, et al. Intelligent diagnosis method of infrared image for transformer equipment based on improved Faster RCNN [J]. China Southern Power Grid, 2019, 13(12): 79-84 (in Chinese).
- [17] 徐向前, 孙涛. Faster RCNN的交通场景下行人检测方法[J]. 软件导刊, 2020, 19(4): 67-70.
- XU X Q, SUN T. Pedestrian detection method in traffic scene based on Faster RCNN [J]. Software Guide, 2020, 19(4): 67-70 (in Chinese).
- [18] 李林升, 曾平平. 改进深度学习框架Faster-RCNN的苹果目标检测[J]. 机械设计与研究, 2019, 35(5): 24-27.
- LI L S, ZENG P P. Apple target detection based on improved Faster-RCNN framework of deep learning [J]. Machine Design and Research, 2019, 35(5): 24-27 (in Chinese).
- [19] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot MultiBox detector [C] // European Conference on Computer Vision. Berlin: Springer, 2016: 21-37.
- [20] REN S Q, HE K M, GIRSHICK R B, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(6): 91-99.
- [21] REDMON J, FARHADI A. YOLOv3: An incremental improvement [EB/OL]. (2018-04-08) [2020-08-01]. <https://arxiv.org/abs/1804.02767>.
- [22] CAI Z W, VASCONCELOS N. Cascade R-CNN: Diving into high quality object detection [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6154-6162.
- [23] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection [EB/OL]. (2020-04-23) [2020-08-01]. <https://arxiv.org/abs/2004.10934>.
- 作者简介:**
- 徐晓华 男, 硕士, 高级工程师。主要研究方向: 电力信息及通信管理。
- 钱平 男, 硕士, 高级工程师。主要研究方向: 电力系统自动化。
- 王一达 男, 硕士, 高级工程师。主要研究方向: 电力系统自动化。
- 周昕悦 女, 硕士, 工程师。主要研究方向: 电力信息系统。
- 徐汉麟 男, 工程师。主要研究方向: 电力信息系统。
- 徐李冰 男, 工程师。主要研究方向: 网络信息技术。

Multi-granularity hazard detection method for electrical power system

XU Xiaohua¹, QIAN Ping², WANG Yida^{1,*}, ZHOU Xinyue¹, XU Hanlin¹, XU Libing¹

(1. State Grid Zhejiang Electric Power Company Hangzhou Power Supply Company, Hangzhou 310020, China;

2. State Grid Zhejiang Electric Power Company, Hangzhou 310007, China)

Abstract: As the security hazard of the electrical power system can lead to serious economic damage and social impacts, the potential hazard detection has become an indispensable part for electrical power system. With the advances of artificial intelligence, intelligent deep learning based hazard detection methods for electrical power system have emerged. Although the existing methods have made promising progress, most of them only consider the global or local features of the image, which cannot thoroughly characterize the image and accurately conduct the hazard detection in the context of the electrical power system especially for the complex outdoor background. In the light of this, in this paper, we present a multi-granularity hazard detection network MGNet for the electrical power system. To be specific, we explore the multi-granularity representation of images with both the global and local representation learning networks. Based on that, we conduct the hazard detection at different granularity levels and finally collaboratively fuse the detection results to fulfill the precise hazard detection. Extensive experiments on two real-world datasets of hazard (i. e., tower connection fitting hazard dataset and transmission line channel mechanical hazard dataset) demonstrate the superiority of the detection performance of the proposed model. In particular, the mean average precision is improved by 2.74% and 2.77% on two datasets, respectively, compared with the existing optimal hazard detection benchmark method.

Keywords: hazard detection; multi-granularity; collaborative fusion; deep learning; electrical power system

Received: 2020-09-02; **Accepted:** 2020-09-04; **Published online:** 2020-11-05 09:28

URL: kns.cnki.net/kcms/detail/11.2625.V.20201104.0931.002.html

Foundation item: Science and Technology Project of State Grid Zhejiang Electric Power Company (5211HZ19014U)

* **Corresponding author.** E-mail: wang_yida@zj.sgcc.com.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0476

电力系统厂站接线图拓扑关系检测技术

李昊^{1,*}, 管冀², 王杉¹, 石玮¹, 刘子鑫³, 刘晓川³

(1. 国网山东省电力公司青岛供电公司 电力调度控制中心, 青岛 266002;

2. 国网山东省电力公司 电力调度控制中心, 济南 250001;

3. 山东大学(青岛) 计算机科学与技术学院, 青岛 266237)



摘 要: 厂站接线图中电气元件的拓扑关系是厂站接线图自动生成技术所需的核心数据。目前, 已知的厂站接线图自动生成技术仍然依靠人工获取图中的拓扑关系。通过利用基于深度学习的目标检测技术与传统的计算机图像处理技术相结合的方式, 能够实现厂站接线图拓扑关系检测。首先, 利用基于深度学习的目标检测方法对电气元件进行识别, 并利用计算机图像处理技术对标量格式接线图进行预处理, 完成电气元件与连接线的分割。然后, 利用轮廓跟踪算法对连接线连通区域进行检测标记。最后, 根据获取的电气元件信息与连接线信息获取图纸的拓扑关系。采用国家电网有限公司提供的数据集, 并设计了对比实验, 验证了所提方法的有效性。

关 键 词: 厂站接线图; 拓扑关系; 深度学习; 目标检测; 轮廓跟踪

中图分类号: TP399; TM734

文献标志码: A **文章编号:** 1001-5965(2021)03-0531-08

电力系统作为国家经济发展的支柱, 支撑着各行各业几乎所有环节的正常运转。自国家公布电网智能化发展计划以来, 中国电力系统得以快速发展, 完成了从旧时代传统电力系统向新时代智能电力系统的转变。随着电网规模的日益扩大, 电力系统的调度工作越来越难以进行, 这给电力系统的优化带来了新的挑战。

厂站接线图作为电力系统中的关键数据, 是电力系统调度工作的基础。无论从电能生产管理角度还是电能生产与消费不可间断角度来讲, 其被广泛应用于电力系统中, 如潮流计算、短路电流计算、可靠性分析和电网调度自动化等。传统的厂站接线图绘制与维护工作主要依靠电力调度人员的工作经验, 这种工作方式不仅增加了电网系统发生安全事故的风险, 而且给厂站接线图的维护工作带来了高昂的人工成本。因此, 国内外众

多机构对厂站接线图的自动生成技术展开了研究, 也取得了一定的成果。但目前已有的厂站接线图自动生成技术需要使用矢量格式的数据, 这种数据需要人工录入电气元件的拓扑关系, 并将以标量格式数据存储的接线图转换为矢量格式。这给厂站接线图的自动生成技术带来了一定人工成本, 也增加了人工录入信息出错的风险。

经过多年研究, 计算机图像处理技术已经发展成熟, 并已成功应用于各个领域。近年来, 深度学习技术也逐渐发展成熟, 一些研究机构已经开始尝试将深度学习技术应用到电力领域。在厂站接线图元件识别技术上, 国内不少机构已经做了不同的尝试, 并取得了一定的效果, 进一步证明了将深度学习技术应用到厂站接线图识别上的可行性。

本文旨在利用计算机图像处理技术与深度学习技术相融合的方式, 自动获取厂站接线图中电

收稿日期: 2020-08-31; 录用日期: 2020-09-05; 网络出版时间: 2020-10-28 16:03

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201028.1441.001.html

基金项目: 国网山东省电力公司科技项目(5206021900TW)

* 通信作者。E-mail: lihao-0717@163.com

引用格式: 李昊, 管冀, 王杉, 等. 电力系统厂站接线图拓扑关系检测技术[J]. 北京航空航天大学学报, 2021, 47(3): 531-538.

LI H, GUAN T, WANG S, et al. Topological relation detection technology of substation wiring diagram in electric power system [J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 531-538 (in Chinese).

气元件的拓扑关系。由于厂站接线图具有元件众多、线条复杂等特点,本文通过对电力系统标量格式接线图进行预处理,实现了高质量、高效率的拓扑关系检测,进一步提高了电力系统智能化与自动化水平,降低了人工成本,对国家电网智能化推进有着重要意义。

1 相关工作

一般来说,厂站接线图自动生成技术的核心部分可分为图元识别和拓扑关系获取 2 部分。其中,图元识别是指电气元件的识别及其相关信息的获取,拓扑关系获取是指获取电气元件之间的连接与邻接关系。

国际电工委员会(IEC)提出了规范系统间互联和模型交换的标准,即能量管理系统应用程序接口^[1]。该标准包含了电力系统公共信息模型(Common Information Model, CIM),其描述了电力系统运行对象的相关信息及对象与对象之间的相互关系。同时,该标准还推荐将矢量图形 SVG 应用至电力系统中。在实际应用中, CIM 模型及 SVG 矢量图中所需的关键信息由人工进行录入,这给厂站接线图自动生成技术带来了一定的人工成本。

目前,厂站接线图自动生成技术大多是基于 CIM 模型和矢量格式数据集实现的。沙树名和林峰^[2]提出了一种根据设备连接信息自动生成厂站接线图的经验方法,对 CIM 数据中的有关设备信息进行提取,进而实现图纸的生成。章坚民等^[3]提出了一种基于 CIM 模型实现不同规模变电站图形的自动生成方法。文献[4-9]描述了在电力领域基于拓扑关系模型的厂站接线图自动生成技术。以上方法均从 CIM 模型、矢量数据图中提取出图元信息、图元拓扑关系信息,进而生成厂站接线图。在此基础上,苏运光等^[10]对此做了改进,提出了基于存量图形拓扑分析的厂站接线图自动生成方法,该方法通过获取图中的特征单元,实现了图元识别,但仍需从 CIM 模型中获取元件的拓扑关系。

为了解决以上问题,本文对标量格式的厂站接线图拓扑关系检测进行了研究。标量图的拓扑关系检测类似于标量图矢量化过程,矢量化研究自 20 世纪末被提出后,已有大量研究人员投身其中,大量算法被相继提出。宋晓宇等^[11]提出了一种动态像素跟踪的直线段提取方法,该方法是基于像素跟踪的局部匹配矢量化方法,具有速度快、局部特征信息描述准确等优点,但在扫描过程

中会产生大量中间结果。Chang 等^[12]提出了一种基于线扫操作的图像骨架化算法,通过对图纸图像进行全局特征统计,获得所有元件的特征描述,从而完成图元识别,但图元之间的连接线难以正确处理,对于连接线相交、黏连、断裂等复杂情况识别准确率低,并且易受噪声干扰。Song 等^[13]提出了一种面向对象的单相矢量化模型,能根据图元的自然特征直接完整地进行像素级别识别,但该方法在识别短小图线时存在困难,并且不同种类图纸的处理方法难以统一。张琪和叶颖^[14]提出了一种基于对象图例及其拓扑关系识别的矢量化方法,在实现对象图例识别的基础上,进一步利用基于连通域标记方法实现了对象图例拓扑关系的识别。该方法对本文研究起到了启发性的作用,但由于厂站接线图具有元件种类多样、连接线错综复杂、绘图规则不一等特点,现有的矢量化技术仍然难以满足其拓扑关系识别的精度需求。

近年来,随着深度学习技术的发展,基于深度学习的目标检测技术已在各个领域取得了不错的效果,各研究机构也已经开始将目标检测技术应用于工程图的拓扑关系检测研究中。如何将传统技术与目标检测技术相融合,从而更精确地获取电力标量图的拓扑关系,是本文研究的重点。

2 厂站接线图拓扑关系检测

2.1 问题定义

在电力图纸中,拓扑关系是指电气元件之间相互空间上的连接与邻接关系,而与电气元件的具体位置无关。厂站接线图是指使用电气系统图形符号和文字符号表示电气装置中的各元件及其相互联系的工程图,按其在电力系统中的作用,可分为厂站一次接线图和厂站二次接线图。厂站一次接线图是电力系统中表示电能输送和电能分配线路的接线图,又称主接线图,主要由母线、变压器、断路器和互感器等一次设备构成。厂站二次接线图是表示控制、指示、测量和保护主电路及其设备运行的接线图,主要由测量仪表、继电保护装置、熔断器和安全自动装置等二次设备构成。本文主要关注厂站一次接线图的拓扑关系识别。

在厂站接线图拓扑关系检测问题中,输入数据为标量格式的厂站一次接线图,如图 1 所示。输出为接线图中各元件信息及代表元件拓扑关系的链接点号,如表 1 所示。其中,每一个链接点号都代表一段连接线,具有相同链接点号的元件相互连接。用链接点号表示元件连接关系的方式只考虑了元件相互空间上的连接与邻接关系,不受

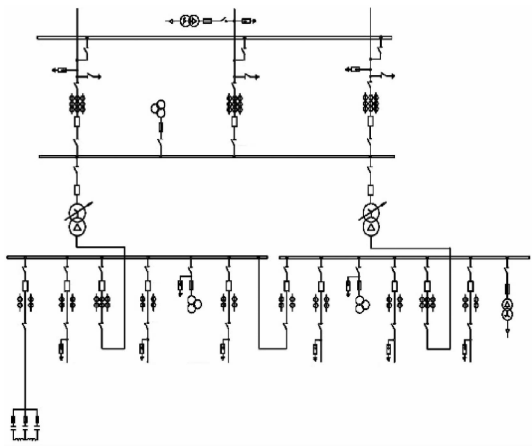


图1 标量格式接线图示例

Fig.1 An example of scalar-format wiring diagram

元件具体位置的影响。因此,元件的拓扑关系可抽象为元件的链接点号。

表1 拓扑关系检测结果

Table 1 Results of topological relation detection				
设备名称	x 坐标	y 坐标	电压等级/kV	链接点号
刀闸_25	1 298	1416	35	54,68
刀闸_10	1 273	182	35	9,12
变压器_2	2 010	1 168	35	74,76
避雷器_10	1 891	1 496	35	43,51
避雷器_5	873	1 806	35	23,31
熔断器_4	297	2 155	35	18,25
电容_2	297	2 197	35	15,18
电容_3	356	2 198	35	15,16

2.2 实现流程

厂站接线图拓扑关系检测技术实现流程如图2所示,主要包括目标检测、预处理、拓扑关系获取等部分。

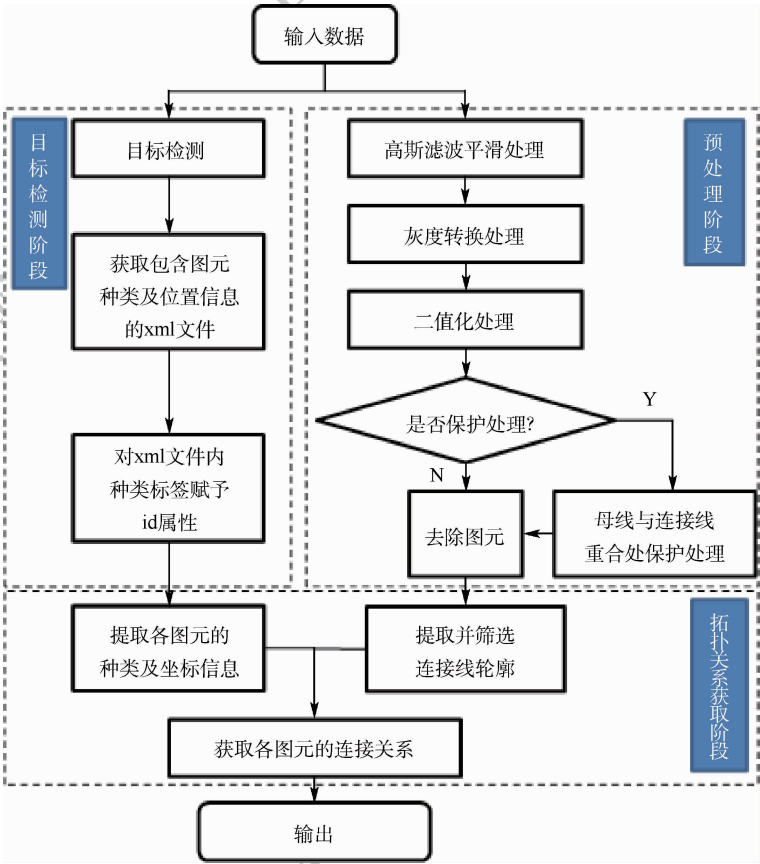


图2 厂站接线图拓扑关系检测流程

Fig.2 Flowchart of topological relation detection of substation wiring diagram

首先,输入数据经过目标检测技术,得到输入接线图中各电气元件的名称、坐标位置等信息,存储至 xml 文件中。然后,根据目标检测的 xml 文件,对输入接线图进行预处理。最后,根据 xml 文件中保存的图元信息及预处理后的输入接线图,进行拓扑关系提取工作,并输出最后结果。

2.3 目标检测

本文使用 Faster R-CNN 深度学习模型进行

目标检测工作^[15]。首先,通过一组基础的卷积层提取图片的特征图,以传递给后续的区域生成网络(Region Proposal Network, RPN)和全连接层。然后,感兴趣区域(Region of Interest, ROI)池化层对特征图和 ROI 进行综合处理,提取出 ROI 的特征图。最后,通过全连接层和 Softmax 回归模型进行目标分类,同时获得目标的精确位置。其中,RPN 网络是 Faster R-CNN 的最大亮点。在进

行候选框位置确定时, Faster R-CNN 抛弃了传统的滑动窗口和 SS (Selective Search) 等方法, 可以利用 RPN 网络直接生成 ROI, 极大地提升了检测框的生成速度。

与大多数检测任务的检测目标相比, 厂站一次接线图的图像分辨率较高, 但图中待检测元件分布密集、像素数少且很多不同种类的元件之间具有较高的相似性。为了提高小尺度目标元件的检测精度, 采用了图像分割技术, 将原图按元件分布分割成若干子图并放大, 再对原图和子图分别训练一个检测网络, 并将大小图网络的检测结果进行融合, 从而得到最终的检测结果。

2.4 预处理

2.4.1 高斯滤波

厂站接线图由于其存储格式的问题, 存在明显的噪声且包含大量的无用细节。采用高斯滤波技术, 对图像进行去噪, 并将高斯核大小设置为 3×3 。通过对图片进行高斯滤波处理, 消除了图中大部分噪声及无用细节。同时, 又最大程度地保留了有用信息。

2.4.2 灰度转换

厂站接线图属于 RGB 三通道图像, 包含的像素信息数量较多。为了简化矩阵, 提高运算速度, 将 RGB 三通道的接线图转换为灰度的单通道图像。本文采用的灰度转换公式是一种常见的转换公式, 如下:

$$\text{Gray} = \frac{R \times 30 + G \times 59 + B \times 11}{100} \quad (1)$$

式中: R 代表图像色彩通道中的红色通道; G 代表图像色彩通道中的绿色通道; B 代表色彩通道中的黑色通道。

式(1)是以国际电信联盟 (International Telecommunication Union, ITC) 提供的灰度转换公式为基础, 将 R 、 G 、 B 三个通道对应的权重由 0.299:0.587:0.114 变为 0.3:0.59:0.11。式(1)提升了运算速度, 又兼顾了人眼对颜色感知的生理特点^[16]。

2.4.3 二值化

将经由高斯滤波处理后的图像转换为如图 3(a) 所示的单通道灰度图像后, 仍存在一些无用细节且图中连接线轮廓不太清晰。为了解决上述问题, 又对厂站一次接线图采取了二值化处理。将 256 个等级的灰度图像根据适当的阈值, 变换为如图 3(b) 所示能够更加反映图像特征的二值化图像。

2.4.4 去除图元

通过目标检测技术已经获取了厂站一次接线

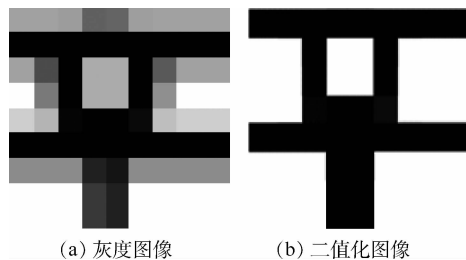


图 3 图像二值化

Fig. 3 Image binarization

图中各电气元件的信息。为了获取厂站接线图的拓扑关系, 还需要获取图中连接线的轮廓信息, 才能进而判断出各元件的连接关系。由于本文方法是对厂站接线图中的像素进行操作, 为了排除电气元件像素对寻找连接线轮廓的影响, 还需对电气图元和连接线进行分割, 去除图像中如图 4(a) 红框所示属于电气元件的像素信息, 使图中不存在属于任何元件的像素信息, 只保留除元件外的其他像素信息, 如图 4(b) 所示。

过程 1 去除图元。

步骤 1 从目标检测返回的 xml 文件中, 依次提取出 $\langle \text{object} \rangle$ 下的 $\langle \text{bndbox} \rangle$ 包含的 $\langle \text{xmin} \rangle$ 、 $\langle \text{ymin} \rangle$ 、 $\langle \text{xmax} \rangle$ 、 $\langle \text{ymax} \rangle$ 坐标信息。其中, $\langle \text{bndbox} \rangle$ 为检测出的元件坐标框, ($\langle \text{xmin} \rangle$, $\langle \text{ymin} \rangle$) 和 ($\langle \text{xmax} \rangle$, $\langle \text{ymax} \rangle$) 分别为该坐标框的左上角端点和右下角端点的像素坐标。

步骤 2 将厂站接线图中坐标 (x, y) 处的像素信息抹去, 其中 $\{(x, y) \mid x \in (\text{xmin}, \text{xmax}), y \in (\text{ymin}, \text{ymax})\}$ 。

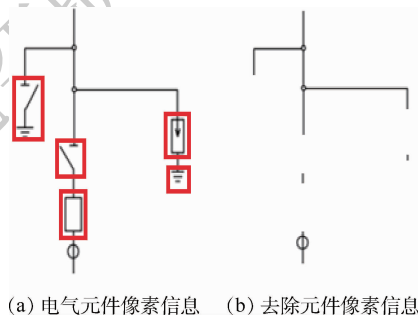


图 4 去除图元

Fig. 4 Removing graphics

2.4.5 保护处理

由于厂站一次接线图布线复杂, 图中线条较多, 存在如图 5(a) 所示母线与连接线交叉重合的情况。如果按照 2.4.4 节去除图元的方式, 如图 5(b) 所示, 会导致本应连续的连接线中断, 从而导致拓扑连接检测出错。因此, 在去除母线区域像素信息的同时需要对母线区域内的连接线进

行保护处理,使其不因去除母线而被消除,从而保持连接线轮廓信息的完整性。本文中,首先,利用类似匹配的方式,检测出母线区域内出现的连接线的位置信息;然后,对母线区域进行去除图元处理,保护处理后的效果如图 5(c)所示。假设母线区域为 $\{(x,y) \mid x \in (x_{\min}, x_{\max}), y \in (y_{\min}, y_{\max})\}$,过程如下。

过程 2 保护处理过程。

步骤 1 创建一个包含 $(y_{\max} - y_{\min} + 1) \times 1$ 维黑色像素的区域。

步骤 2 将此区域以 1 像素的步距,沿着母线行索引递增的方向运动。

步骤 3 运动一次便与当前所在母线区域包含的像素信息进行迭代比较一次,判断当前母线区域包含的像素信息是否与其包含的像素信息相同,相同则保存当前位置信息之后继续运动,否则继续运动,直至遍历完母线区域。最终可以得到母线区域内所有交叉重合连接线的行坐标。

步骤 4 结合去除图元的方法,抹去母线区域像素信息的同时避开母线内连接线区域,就可以达到保持连接线轮廓完整性的要求。

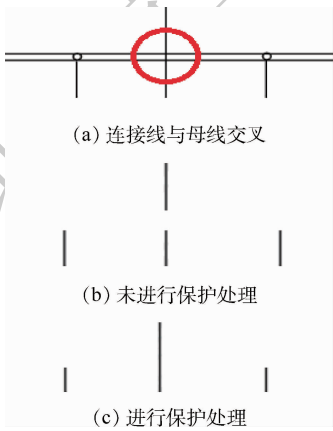


图 5 保护处理
Fig. 5 Protection processing

2.5 拓扑关系获取

步骤 1 针对厂站一次接线图像素信息的特点,使用基于 8 邻域连通区域标记的轮廓跟踪算法,对预处理后的图像进行连接线的轮廓提取,并将提取后的轮廓信息存储至列表中。

步骤 2 对保存至列表中的轮廓坐标信息进行筛选操作,将 L 中的坐标信息 (x,y) 与 2.4.4 节提取出的坐标信息 $\langle x_{\min} \rangle$ 、 $\langle y_{\min} \rangle$ 、 $\langle x_{\max} \rangle$ 、 $\langle y_{\max} \rangle$ 进行比较判断,只保留各轮廓内符合 $x_{\min} - 1 \leq x \leq x_{\max} + 1$ 和 $y_{\min} - 1 \leq y \leq y_{\max} + 1$ 条件的轮廓坐标信息,并对提取出的每个轮廓进行编号,以作为后续代表图元拓扑关系的链接点

号。其中,通过对 $x_{\min}$ 、 $y_{\min}$ 和 $x_{\max}$ 、 $y_{\max}$ 分别采取减 1、加 1 操作,消除了二值化处理与人工绘制图纸不规范导致的电气元件轮廓与连接线轮廓不相连的情况,保证了厂站接线图的完整性。

步骤 3 将筛选后的各连接线轮廓坐标信息和图元的坐标信息进行比较,将与同一连接线相连的图元放至集合中,并将该轮廓编号作为集合编号,同时使用集合操作,可以达到去重的效果。

至此,已经获得厂站接线图包含的所有连接关系。

3 实验

3.1 实验设定

为了验证本文方法的有效性,使用真实的厂站接线图数据集:青岛市某 70 个区域变电站接线图数据集。该数据集主要包含断路器、变压器、电抗和避雷器等 12 种电气元件,其数据规模如表 2 所示。

表 2 青岛市区县变电站接线图数据集
Table 2 Wiring diagram dataset of district and county substation in Qingdao

元件类别	数目
断路器	344
接地刀闸	1 093
电抗	208
刀闸	1 258
变压器	144
避雷器	829
母线	275
手车开关	991
隔离手车	308
所变	172
电力电感器	205
电容	974

共设计了 3 组实验,对本文方法进行验证。第 1 组验证分割技术对目标检测任务的有效性;第 2 组检验分割技术对拓扑关系检测的影响;第 3 组验证预处理技术对拓扑关系检测的有效性。

在第 1 组实验中,按照 8:1:1 的比例对数据集进行划分(即 80% 作为训练集,10% 作为验证集,10% 作为测试集),并用目标检测常用评价指标——平均精度(Average Precision, AP)、平均精度均值(mean Average Precision, mAP)对模型进行评估。

在第 2、3 组实验中,将第 1 组实验的测试集作为测试数据,并用精确率(Precision)、召回率(Recall)和 F1-score 指标进行评估,公式如下:

$$\text{Precision} = \frac{\text{CorrectRelations}}{\text{DetectedRelations}} \tag{2}$$

$$\text{Recall} = \frac{\text{CorrectRelations}}{\text{TotalRelations}} \tag{3}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

式中:CorrectRelations 为正确检测出的拓扑关系数;DetectedRelations 为检测出的拓扑关系总数;TotalRelations 为数据共包含的正确拓扑关系总数。

3.2 验证分割技术对目标检测任务的有效性

为了验证分割技术对目标检测任务的有效性,对比了分割前后各类元件的目标检测结果。图 6 展示了分割前后各类元件的 AP 结果对比,表 3 给出了分割前后目标检测的 mAP 结果对比。

从图 6 及表 3 中可以看出,采用分割技术的目标检测模型在数据集中的表现明显高于未采用分割技术的模型。该实验结果说明,分割技术增强了模型的检测效果,有利于实现更加精准的电气元件检测。同时还观察到,采用分割技术后,模型对于一些尺寸较小、特征不明显的电气元件(如刀闸、避雷器、隔离手车、电容等)的检测效果有了明显的提升,而对于一些尺寸较大、特征较明显的元件(如变压器、手车开关等)的检测效果相对提升较小。这可能是因为:分割技术使得目标检测模型更加注重图像中局部区域的特征提取,使得模型对较小元件的检测结果更加准确。

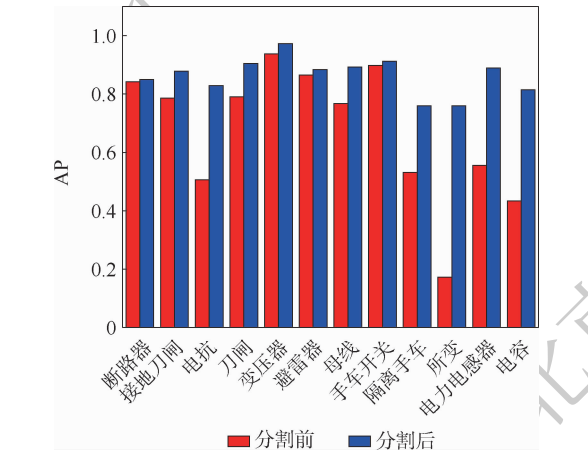


图 6 分割前后各类元件的平均精度

Fig. 6 Average precision of various components before and after segmentation

表 3 分割前后目标检测平均精度均值

Table 3 Mean average precision of object detection before and after segmentation

分割前后	mAP/%
分割前	62.91
分割后	84.37

3.3 分割技术对拓扑关系检测的影响

为了检验分割技术对拓扑关系检测的影响,利用未采用分割技术的目标检测模型与采用分割技术的目标检测模型,分别对测试数据进行拓扑关系检测。表 4 展示了采用分割技术前后的拓扑关系检测结果对比。

从表 4 可以看出,分割技术可有效提升拓扑关系检测效果。经过分析发现,目标检测存在漏检、误检等情况,会直接导致拓扑关系检测效果下降,目标检测效果的提升有助于实现更高效、精确的拓扑关系检测。本文使用的分割技术可以有效提升目标检测效果,进而提升拓扑关系检测效果。

表 4 分割前后拓扑关系检测结果对比

Table 4 Comparison of topological relation detection results before and after segmentation

分割前后	Precision/%	Recall/%	F1-score/%
分割前	81.27	74.09	77.41
分割后	85.02	85.86	85.42

3.4 验证预处理对拓扑关系检测的有效性

为了验证预处理对拓扑关系检测的有效性,对测试集数据分别进行了有无预处理的拓扑关系检测实验,表 5 展示了有无预处理阶段的拓扑关系检测的结果对比。由于去除图元是分割图元与连接线的关键操作,为了能够正确提取出连接线位置信息,保证实验结果的准确性,该环节的无预处理实验具有去除图元操作。

从表 5 可以看出,预处理工作对拓扑关系检测有关键性的作用。经过分析发现,未进行预处理的厂站接线图具有许多无用像素,这些像素导致了元件粘连、线条轮廓不清晰等情况的发生,进而使得轮廓跟踪算法无法正确检测出连接线轮廓,最终导致较差的拓扑关系检测结果。该实验结果进一步表明了预处理操作对厂站一次接线图数据的高度适用性。

表 5 预处理前后拓扑关系检测结果对比

Table 5 Comparison of topological relation detection results before and after preprocessing

有/无预处理	Precision/%	Recall/%	F1-score/%
无预处理	35.92	1.64	3.11
有预处理	85.02	85.86	85.42

4 结 论

1) 本文方法可以有效检测出厂站接线图中电气元件的拓扑关系,且实现了较优的检测效果。在 Precision、Recall、F1-score 三个指标上,分别达到了 85.02%、85.86%、85.42%。

2) 本文方法适用于标量格式的数据集,减少了人工干预,降低了人工成本。

3) 目标检测效果可直接影响拓扑关系检测效果,分割技术提升了目标检测效果,进而提升了拓扑关系检测的效果。

4) 标量格式数据集存在制图不规范、噪声多等问题,这会导致拓扑关系检测效果下降,预处理过程可降低这类问题对检测结果的不利影响。

为提升本文方法的检测效果,仍需优化目标检测模型和增加对数据集有效的预处理。

参考文献 (References)

- [1] W3C. Scalable vector graphics (SVG) 1.2 specification [EB/OL]. (2004-10-27) [2020-08-01]. <http://www.w3.org/TR/2004/WD-SVG12-20041027>.
- [2] 沙树名,林峰.一种基于CIM的厂站接线图自动生成技术[J].电力系统自动化,2008,32(21):68-71.
- SHA S M, LIN F. A technique of automatic generation of substation wiring diagram based on CIM [J]. Automation of Electric Power Systems, 2008, 32(21): 68-71 (in Chinese).
- [3] 章坚民,叶义,徐冠华.变电站单线图模数图一致性设计与自动成图[J].电力系统自动化,2013,37(9):84-91.
- ZHANG J M, YE Y, XU G H. Consistency design and automatic mapping of single line diagram modulus diagram of substation [J]. Automation of Electric Power Systems, 2013, 37(9): 84-91 (in Chinese).
- [4] 沈伟,吴文传,张伯明,等.能量管理系统中电网潮流单线图自动生成算法[J].电力系统自动化,2010,34(6):48-53.
- SHEN W, WU W C, ZHANG B M, et al. A power grid one-line diagram automatic generation method in the EMS [J]. Automation of Electric Power Systems, 2010, 34(6): 48-53 (in Chinese).
- [5] 章坚民,叶琳,孙维真,等.基于地理相对位置的省级输电网均匀接线图自动生成[J].电力系统自动化,2010,34(24):55-59.
- ZHANG J M, YE L, SUN W Z, et al. Automatic generation of the uniformly-distributed provincial transmission power grid diagram based on relative geographical location information [J]. Automation of Electric Power Systems, 2010, 34(24): 55-59 (in Chinese).
- [6] 周博曦,孟昭勇,王志臣,等.基于CIM的变电站与配电馈线一次接线图自动绘制算法[J].电力系统自动化,2012,36(11):77-81.
- ZHOU B X, MENG Z Y, WANG Z C, et al. Automatic graphing algorithm for primary electric wiring diagram of substation and distribution feeder based on CIM [J]. Automation of Electric Power Systems, 2012, 36(11): 77-81 (in Chinese).
- [7] 陈连杰,赵仰东,韩韬,等.基于层次结构及模型驱动的配电网图形自动生成[J].电力系统自动化,2015,39(1):226-232.
- CHEN L J, ZHAO Y D, HAN T, et al. Hierarchically structured and model driven automatic generation of distribution network graphs [J]. Automation of Electric Power Systems, 2015, 39(1): 226-232 (in Chinese).
- [8] 陈璐,陈连杰,欧阳文,等.基于环形结构的配电网网络图布局算法[J].电力系统自动化,2016,40(24):151-156.
- CHEN L, CHEN L J, OUYANG W, et al. Ring structure based layout algorithm for distribution network contact graph [J]. Automation of Electric Power Systems, 2016, 40(24): 151-156 (in Chinese).
- [9] 廖凡钦,刘东,闫红漫,等.基于拓扑分层的配电网电气接线图自动生成算法[J].电力系统自动化,2014,38(13):174-181.
- LIAO F Q, LIU D, YAN H M, et al. An automatic electrical diagram generation method for distribution networks based on hierarchical topology model [J]. Automation of Electric Power Systems, 2014, 38(13): 174-181 (in Chinese).
- [10] 苏运光,罗俊,熊浩,等.基于存量图形拓扑分析的厂站接线图自动生成[J].电力系统自动化,2020,44(3):162-168.
- SU Y G, LUO J, XIONG H, et al. Automatic generation of substation wiring graph based on inventory graph topology analysis [J]. Automation of Electric Power Systems, 2020, 44(3): 162-168 (in Chinese).
- [11] 宋晓宇,王永会,栾方军.基于动态跟踪步长的直线快速矢量化算法[J].计算机工程与应用,2002,38(2):92-93.
- SONG X Y, WANG Y H, LUAN F J. Straight lines vector algorithm based on dynamic trace step [J]. Computer Engineering and Applications, 2002, 38(2): 92-93 (in Chinese).
- [12] CHANG F, LUY C, PAVLIDIS T. Feature analysis using line sweep thinning algorithm [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999, 21(2): 145-158.
- [13] SONG J Q, SU F, TAI C L, et al. An object-oriented progressive-simplification-based vectorization system for engineering drawing: Model, algorithm, and performance [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(8): 1048-1060.
- [14] 张琪,叶颖.基于对象图例及其拓扑关系识别的二维工程CAD图纸矢量化方法[J].计算机与现代化,2018(11):40-45.
- ZHANG Q, YE Y. 2D engineering CAD drawings vectorization method based on recognition of object legends and their topology [J]. Computer and Modernization, 2018(11): 40-45 (in Chinese).
- [15] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.
- [16] PRATT W K, ADAMS J E. Digital image processing [J]. Journal of Electronic Imaging, 2007, 16(2): 131-145.

作者简介:

李昊 男,硕士,工程师。主要研究方向:电力调度自动化。

Topological relation detection technology of substation wiring diagram in electric power system

LI Hao^{1,*}, GUAN Ti², WANG Shan¹, SHI Wei¹, LIU Zixin³, LIU Xiaochuan³

- (1. Electric Power Dispatching & Control Center, State Grid Qingdao Power Supply Company, Qingdao 266002, China;
- 2. Electric Power Dispatching & Control Center, State Grid Shandong Electric Power Company, Jinan 250001, China;
- 3. School of Computer Science and Technology, Shandong University (Qingdao), Qingdao 266237, China)

Abstract: Topological relation of electrical components is the core-data required by substation wiring diagram automatic generation technology. At present, known technologies still rely much on artificial access to topological relations. By the combination of deep learning based object detection technology and traditional computer image processing technology, topological relation can be detected automatically. Firstly, to segment the electrical components and connection lines, deep learning based object detection technology was used to identify the electrical components, and the image processing technology was used to preprocess the scalar-format wiring diagram of power plants. Secondly, a contour tracking algorithm was adopted to detect and mark the connected area of the connection lines. Finally, the topological relation of the drawing was acquired according to the obtained information of electrical components and connection lines. Comparative experiments based on a dataset released by the State Grid Corporation of China indicate the effectiveness of the proposed method.

Keywords: substation wiring diagram; topological relation; deep learning; object detection; contour tracking

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

DOI: 10.13700/j.bh.1001-5965.2020.0478

基于深度学习和图匹配的接线图检测与校核



李昊^{1,*}, 王杉¹, 耿玉杰², 王黎¹, 孙文昌³, 苗纯源³

(1. 国网山东省电力公司青岛供电公司 电力调度控制中心, 青岛 266002;

2. 国网山东省电力公司 电力调度控制中心, 济南 250001;

3. 山东大学(青岛) 计算机科学与技术学院, 青岛 266237)

摘 要: 传统的厂站一次接线图的绘制和管理主要依靠电网运行人员, 费时费力且缺乏科学可校核的参考标准。提出了一种基于深度神经网络和数字图像处理相结合的厂站一次接线图的自动检测、识别和校核算法。首先, 使用目标检测 Faster R-CNN 模型检测厂站接线图中的电器元件, 并达到 92% 的检测准确率, 同时使用端到端的文字检测识别模型识别厂站接线图中的文字信息, 并达到 94.2% 的文字检测准确率和 92% 的文字识别准确率; 然后, 使用数字图像处理技术进行厂站接线图连接线、拓扑关系识别; 最后, 使用改进的 VF2 算法进行厂站一次接线图和人工维护的厂站一次接线图拓扑关系匹配校核, 将拓扑数据抽象为无向图, 通过轮廓序号得到元件的相对位置信息, 根据改进的 VF2 算法得到 2 张图的匹配率, 并通过匹配率与设定好的阈值来帮助核验, 相比于节点遍历的匹配方法, 核验准确率提高了 37.5%。基于某供电公司提供的部分变电站的厂站一次接线图标注了接线图电器元件, 贡献了一个小型接线图数据集。

关键词: 厂站一次接线图; 拓扑关系匹配; 自动生成; 自动校核; 深度学习; 数字图像处理

中图分类号: TP399; TM734

文献标志码: A

文章编号: 1001-5965(2021)03-0539-10

在电力领域, 传统的厂站一次接线图图纸绘制和管理对电网工作人员有很高的要求, 如需要对电网历史及当前厂站设备非常熟悉且有丰富的工作经验等, 其面临着两方面的问题: ①缺乏科学可校核的参考标准, 很有可能导致厂站接线图缺乏规范, 进而含有潜在风险。②由于网络新建、旧线改造、方式变更、检修等原因导致主接线的元件组成、连接方式经常发生变化, 要做到厂站接线图的实时更新需要付出高昂的人工成本和管理成本。针对以上问题, 其对自动化的需求越来越迫切。

电网主接线图是电力系统分析的基础, 其广泛用于电力系统中, 如潮流计算、短路电流计算、可靠性分析、电网调度自动化等。考虑到高昂的人工管理成本, 不少学者开发了电力系统专用的图形编辑软件, 使得电气工程图生成效率大大提高, 但其仍然依靠人工编辑, 并且大部分图形的生成、维护和更新都还是人工进行, 电力工作人员使用 CAD 等绘图软件绘制人工维护的厂站一次接线图, 并进行人工校核, 导致图纸绘制及一次接线图和人工一次接线图的校核缺乏规范性。因此, 厂站接线图自动检测、维护和智能校验仍然是国

收稿日期: 2020-08-31; 录用日期: 2020-09-11; 网络出版时间: 2020-11-02 11:18

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201030.1652.004.html

基金项目: 国网山东省电力公司科技项目(5206021900TW)

* 通信作者。E-mail: lihao-0717@163.com

引用格式: 李昊, 王杉, 耿玉杰, 等. 基于深度学习和图匹配的接线图检测与校核[J]. 北京航空航天大学学报, 2021, 47(3): 539-548. LI H, WANG S, GENG Y J, et al. Wiring diagram detection and check based on deep learning and graph matching[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 539-548 (in Chinese).

内外众多学者不断深入研究的课题。

针对现存问题,本文提出了一种厂站接线图自动检测、识别和校核算法,实现了厂站一次接线图的自动校核。算法分为 2 部分:①对厂站接线图中的元器件、文字和拓扑关系等利用改进的深度学习和数字图像处理相关的检测和分类算法进行目标定位和识别。②基于图匹配算法对电子厂站一次接线图和人工维护的厂站一次接线图进行元器件和拓扑关系匹配校核。

本文研究的内容突破了传统人工识别、人工校核的方式,基于人工智能和数字图像处理技术实现厂站一次接线图的自动识别、自动校核,对接线图中电器元件的名称、坐标、拓扑连接关系等进行精准识别和校核。本文研究的内容不仅可以应用于厂站接线图的自动识别、自动校核,也可以为其他领域的线路图识别、校核任务提供思路。本文研究内容成果投入运行后,可以有效地提高图纸绘制速度、准确度和标准化程度,广泛应用到调控云的基础数据输入,为电力行业引入计算机领域的前沿技术,促进电力智能化的发展,切合泛在电力物联网的建设要求。

本文的主要贡献如下:①将目标检测和识别算法应用到了电力厂站一次接线图上,并且提出了一种多尺度融合的改进方式,实现了较高的检测和识别准确率;②基于改进的图匹配算法对厂站一次接线图和人工维护的厂站一次接线图进行匹配校核,解决了接线图电器元件标号顺序的问题,实现了较高的匹配校核准确率;③标注了一个小型电力厂站一次接线图电器元件数据集。

1 相关工作

关于电力图纸的自动绘制,早期的大多研究内容集中在基于地理信息技术平台的图形自动生成算法方面,主要是基于数据库或空间数据引擎,实现数据搜索,进行拓扑分析和图形自动绘制^[1]。部分学者以典型接线形式的图形编排为依据,提出了一种根据设备连接信息自动生成接线图的经验方法,即借助于 Oracle 数据库提供的一套基于结构查询语言 (SQL) 的方案和函数 (Oracle Spatial) 的网络分析功能,对公共信息模型 (CIM) 数据中有关设备信息进行提取、编排实现厂站图的生成^[2],该方法有利于系统集成,但模板匹配算法依赖性强,不易扩展。后来又有部分学者提出了基于存量图形自动生成厂站接线图的原理。从基于 CIM/G 的拓扑分析出发,给出了存量图形特征识别、特征提取及建立图形特征库

的方法,结合特征评估,获取厂站接线图特征单元,并在特征单元的基础上完成厂站接线图的生成与绘制^[3]。部分学者提出了通过采用节点布局的数学模型和线路走线寻路的函数算法来实现图形布局^[4],该算法可保证线路完全不相交,但这使得线路严重弯折,有时甚至无法完成连线工作。

国内的几大电力自动化系统供应商在厂站一次接线图自动生成和校核方面都进行了一定的尝试,但能够成熟应用的产品不多见。河北电网曾率先研发了一种基于典型接线方式的智能系统来实现厂站图自动生成的软件^[5]。基于厂站主接线图是有规范的且接线类型有限,朱永利等^[5]提出了一种表示厂站拓扑结构的框架,接线图的整体布局思想为:先画面上下划分,分割轴线作为主变中心;之后布置母线,再布置变压器,最后布置线路等的布局,基于经验规则归纳了 20 多种常见的接线类型,并利用这些规则进行各种接线类型母线和相关的特殊支路(如母联、旁母、分段)的处理;主变布置有十几种规则^[5]。遗憾的是,该方法没有考虑到变电站的数据统一建模,同时,其是基于经验进行预知的布局布线,是一种穷举试错的智能绘图方法,然而由于变电站规模不一,拘泥于细节的接线方式十分繁多,该方法的规则数量十分庞大。之后,华中科技大学将 SVG 图形格式引入到了变电站图形系统中,改进了朱永利等^[5]的方法,并取得了很好的效果,但仍需要在作图过程中定义图元之间的连接关系,工作量依然很大。

经过多年的发展,各研究机构已经开始尝试将人工智能技术和图像处理技术应用于工程设计图纸的分析与处理。例如,黄若航^[6]利用基于 Inception ResNetV2 和 Faster R-CNN 算法的深度学习网络,对经过预处理的热力管道工程设计图进行特征提取、目标定位和目标分类,得到一个基于工程图纸元件数据集的目标检测模型,之后利用光学字符识别技术 (OCR) 对目标检测模型识别出的图纸区域进行数据提取。同时,电力系统基于人工智能技术已做出多种技术创新,并取得了一定的效果。例如,尹思宇^[7]在智能变电站的建设中,为解决海量电力设备红外图像数据的智能检测问题,对红外图像进行预处理,分离出存在缺陷的电力设备,以深度学习为理论基础提出了一种基于深度学习的电力设备红外图像缺陷识别分类的算法,进一步证明了基于人工智能的厂站一次接线图自动生成、自动校核技术的可行性。

2 关键技术

为了实现人工维护的厂站一次接线图自动校核,需要先进行厂站一次接线图的自动识别,再进行匹配校核。实现厂站一次接线图的自动识别,又分为接线图中的电器元件识别,接线图中的文字信息识别,接线图中的连接线识别、拓扑连接关系识别 3 部分。具体流程如图 1 所示。

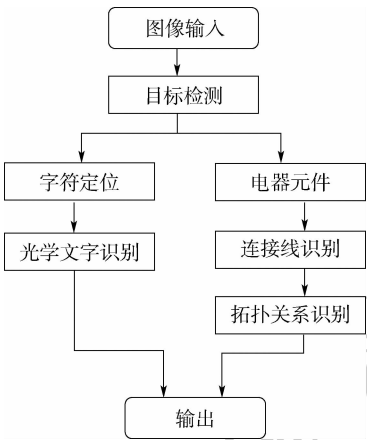


图 1 厂站一次接线图识别流程

Fig. 1 Identification flowchart of primary wiring diagram of plant and station

2.1 厂站接线图中的电器元件识别

电器元件的识别是图纸拓扑关系识别的基础,其准确率关乎厂站接线图自动生成和自动校核的精度。针对厂站一次接线图中电器元件的检测识别问题,经过分析发现,图纸中各电器元件有以下特征:具有规定的形状且接线图中常出现的有 17 种类别;部分电器元件过小且形状相似度极高;分布位置不定、方向不定、大小不一。要想做到厂站接线图的准确识别,要求模型能够精确找到元件所在位置并准确识别其类别。

2.1.1 基于 YOLO 的目标检测算法

本文先采用 One-Stage 检测的深度学习模型 YOLOv3^[8]来进行元器件的检测识别。由于厂站接线图的分辨率过大,直接将其作为网络输入会导致小目标的检测效果较差。针对此问题,本文考虑将厂站接线图切割成 4 份后进行目标检测。将厂站接线图输入网络,利用 darknet-53^[8-9] 深层网络来提取目标的特征,引入 3 个不同尺度的特征图预测应对目标形状的变化问题,优化提升目标检测准确率和位置精度,再将 4 份检测的结果进行拼接,得到电器元件的类别和坐标等属性信息。其坐标信息公式如下:

$$\begin{cases} b_x = \sigma(t_x) + c_x \\ b_y = \sigma(t_y) + c_y \\ b_w = p_w e^{t_w} \\ b_h = p_h e^{t_h} \end{cases} \quad (1)$$

式中: t_x 、 t_y 、 t_w 和 t_h 为模型的预测输出; c_x 和 c_y 为网格单元的坐标; σ 表示 Sigmoid 函数; p_w 和 p_h 表示预测前 bounding box 的大小; b_x 、 b_y 、 b_w 和 b_h 为预测得到的 bounding box 中心的坐标和尺寸。

对于超大分辨率的厂站接线图来说,由于元件众多且大小相差较大,直接使用 YOLOv3 来进行检测效果欠佳,故而进一步考虑基于多尺度融合的检测算法。

2.1.2 基于 Faster R-CNN 的多尺度融合检测算法

Faster R-CNN^[10] 作为一种 Two-Stage 目标检测算法,先使用特征提取网络提取输入图像的特征图,用于 RPN (Region Proposal Network) 和全连接层共享。再使用 RPN 网络生成候选区域。在 Faster R-CNN 算法中,ROI (Region of Interest) 池化层收集输入的特征图和候选区域,综合这些信息后对候选区域提取特征图,送入全连接层判定目标类别。最后利用候选区域的特征图计算候选区域的类别,并通过回归获得检测框的精确位置。

对于超大分辨率的厂站接线图,为使算法适用于不同尺度的电器元件,本文提出了一种多尺度融合的检测算法。使用整张接线图和其分割 4 份后的小图构建的数据集分别训练出对应的 Faster R-CNN 模型,即大图模型和小图模型。对于大型的电器元件,大图模型有更好的检测效果,而对于细小型电器元件,则容易出现错检和漏检的现象。通过计算两者结果的交并比 (IoU) 来融合 2 种模型检测的结果,使用小图模型检测后的细粒度的结果去弥补大图模型的错检和漏检,有效实现了对厂站接线图中大小不一的电器元件的精准检测和识别。图 2 对目标检测的模型框架进行了展示。

2.2 厂站接线图中的文字识别

厂站接线图中的信息复杂,字符信息是接线图中信息提取的重中之重。厂站接线图中的文字信息字体多样、长度不同、位置多变、方向不定,字符也不局限于字母和数字,还有各种器件标识、说明等,个别地方也会出现线条密集、字符与线条黏连等情况,想要利用通用的光学文字识别模型直接来检测厂站接线图中的文字信息难度极大,效果不佳。因此,需要研究针对厂站接线图的特定光学文字识别模型,以提升文字信息的识别准确率。

厂站接线图光学文字识别的流程如图 3 所示。先对厂站接线图进行图片预处理,使图片有

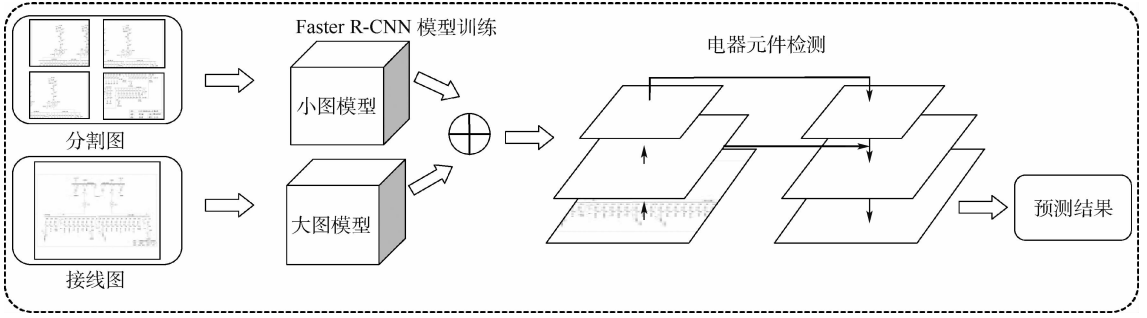


图2 厂站接线图多尺度融合检测算法

Fig. 2 Multi-scale fusion detection algorithm for wiring diagram of plant and station

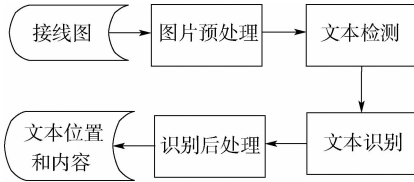


图3 厂站接线图文字识别流程

Fig. 3 Text recognition process of wiring diagram of plant and station

助于后续的文字识别,再进行文字检测定位,获取图片中所有的文本框的位置坐标,之后进行识别前处理和文字识别部分,最终对识别的内容进行后处理,以提高识别的准确率。

2.2.1 基于 Two-Stage 的文字识别算法

先采用 Advanced EAST^[11] 模型对文本区域进行检测,再采用 Tesseract OCR^[12] 模型对检测出的文本区域中的文本信息进行识别,即文字检测和文字识别相分离的两阶段算法。由于厂站接线图是大分辨率图片,采用分割为 4 份的做法来提高文字识别的准确率。在识别出结果后再进行合并。

Advanced EAST 以 VGG16^[13] 作为网络结构的骨干来提取特征图,之后将提取的特征图采用 U-Net^[14] 的算法进行合并,最终输出检测出文本框的置信度和坐标信息。Advanced EAST 算法对于长文本和方向不定的文本区域有着较好的检测效果。

Tesseract OCR 识别过程分为 2 步:单字切割和分类任务。对检测出的文本区域内的文本通过字符单元分割出单个字符。之后对每个字符进行分析,采用自适应分类器对分割出的字符进行分类。实现了对厂站接线图中拥有字体多样、长度不一、方向不定等特点文本的准确识别效果。

2.2.2 基于端到端的文字识别算法

端到端的文字识别模型融合文本检测和文本识别 2 个过程,简化文字识别的处理过程,同时加快识别速度。采用 Attention-OCR^[15] 进行端到端的文字识别。在厂站接线图输入到 Attention-OCR 网络之前,对文字区域使用 Cascade R-CNN^[16] 模型进行文字分割。

Cascade R-CNN 作为两阶段 R-CNN 检测的

多级扩展,通过级联不同的 IoU 阈值检测模型挑选对应质量的样本进行训练和回归。使用 Cascade R-CNN 进行厂站接线图的文本区域检测,并根据其检测出的结果分割出厂站接线图中的文字用于 Attention-OCR 的文本识别。Attention-OCR 主要利用了 spatial attention^[17] 进行端到端的文字识别。通过卷积网络提取特征并融合为一个大的特征矩阵 f 。计算图片中文本的 spatial attention 值 α_i ,其值越大代表该区域为文字区域的概率越大。通过 α_i 抽取特征矩阵 f 中的文本区域特征,后续送入循环神经网络(RNN)进行文本识别。

不同时间进行空间位置的加权公式如下:

$$u_{t,c} = \sum_{i,j} \alpha_{t,i,j} f_{i,j,c} \quad (2)$$

式中: i 和 j 为空间位置; t 为 RNN 的时间维度; c 表示特征图的通道; $u_{t,c}$ 为根据注意力模型给出的权重对不同位置的特征加权后的特征向量,作为模型的输入。

网络在时刻 t 的预测输出字符最终为 c_t ,其推导过程如下:

$$\begin{cases} x_t = W_c c_{t-1}^{\text{Onehot}} + W_{u1} u_{t-1} \\ (o_t, s_t) = \text{RNNstep}(x_t, s_{t-1}) \\ \hat{o}_t = \text{softmax}(W_o o_t + W_{u2} u_t) \\ c_t = \underset{k}{\text{argmax}} \hat{o}_t(k) \end{cases} \quad (3)$$

式中: x_t 为 RNN 在时刻 t 的输入; c_{t-1} 为前一刻的预测字符; s_t 为时刻 t 的 RNN 隐藏层的值; $\hat{o}_t$ 为结合 RNN 输出 o_t 和注意力特征向量 u_t 的时刻 t 的预测输出; W_c 为时刻 $t-1$ RNN 权重矩阵; W_o 为时刻 t RNN 权重矩阵; W_u 为注意力特征向量的权重矩阵。最终通过计算得出预测的字符 c_t 。

由于厂站接线图的分辨率过大,影响算法的处理速度和精度,同样对厂站接线图进行分割处理。将图片分别切成 4 份和 9 份,由于图片切割成 4 份后的切割处容易造成文本信息的缺失,而切割为 9 份则计算耗时较长。为了提高识别的准确率和速度,如图 4 所示,考虑使用切成 9 份后的第 2、4、5、6、8 这 5 张图片的信息来进行弥补。具

体为使用 Attention-OCR 对 5 张图片进行文字识别,根据其识别的结果和切割处的文本进行交并比的判断来决定是否添加该图片的信息,以对切割处遗失的信息进行增补。

得到初步识别结果后,根据先验知识对结果进行后处理,以提高识别的准确率。

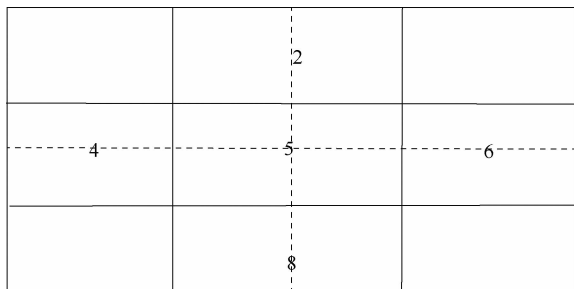


图 4 厂站接线图文字识别分割

Fig. 4 Text recognition segmentation of wiring diagram of plant and station

2.3 厂站接线图中的连接线、拓扑关系识别

在之前的基础之上,可以得到厂站接线图中所有电器元件的个体信息,但是元件与元件之间的连接关系却无法得到,想要获取整幅图像的拓扑关系,需要依据元件之间的连接关系,因此,需要进行连接线识别及拓扑关系识别来找到元件与元件之间的联系,完成厂站接线图的识别任务。

关于厂站接线图拓扑关系识别的工作,主要思想为:利用 OpenCV 的轮廓跟踪技术^[18],检测出每段连接线所连元件,进而得到每个元件的连接关系。使用数字图像处理技术,并与目标检测技术相结合,对厂站接线图进行拓扑连接关系检测。

首先,对目标检测的结果进行提取,将厂站接线图目标检测获取到的各电器元件的位置、种类等信息保存至文件中。然后,对文件进行预处理,为区分开多个同种类部件,对同名元件赋予 id 属性,并提取出各元件的位置和种类信息。

之后对厂站接线图进行图像预处理。使用高斯滤波平滑处理去除厂站接线图中的无用信息和噪声。通过灰度转换将厂站接线图由三通道图像转换为单通道图像,提升对接线图的处理速度。最终通过二值化处理使线条轮廓更加清晰,提高后续检测的精确度及效率。其中,高斯滤波平滑处理中高斯函数的公式如下:

$$G(x,y) = \frac{1}{2\pi\theta^2} e^{-\frac{x^2+y^2}{2\theta^2}} \quad (4)$$

式中:\$(x,y)\$代表像素值的坐标信息;\$\theta\$为正态分布的标准偏差。

对厂站接线图数据的特征进行预处理即去除图像中的目标元件,只保留除图元外的其他像素信息,以便后续提取出元件间连接线的轮廓信息。

针对电力厂站接线图像素信息的特点,使用轮廓跟踪算法,将提取后的各轮廓坐标信息保存。同时,对保存的各轮廓进行筛选判断,只保留与目标元件连接的连接线轮廓信息。即比较轮廓信息和提取的目标检测的结果,只保留符合如下公式条件的轮廓坐标信息,并删去进行判断后不包含任何坐标信息的轮廓:

$$x_{\min} \leq x \leq x_{\max} \cap y_{\min} \leq y \leq y_{\max} \quad (5)$$

式中:\$x_{\min}\$、\$x_{\max}\$、\$y_{\min}\$和\$y_{\max}\$为目标检测的坐标信息;\$x\$和\$y\$为各轮廓坐标信息。

最终根据提取的连接线外轮廓坐标信息和电器元件的种类、位置信息,解析出各连接线的连接关系,将连接线的连接关系转换为元件的拓扑连接关系。

2.4 人工维护的厂站一次接线图自动校核

本模块将厂站一次接线图和人工维护的厂站一次接线图的拓扑检测结果进行提取,将元件的连接关系进行抽象,抽象成 2 张无向图,再通过子图同构匹配算法对 2 张无向图进行匹配,得到 2 张图的映射关系,来对手工图的错误进行核验。同时,通过提取拓扑检测时的线条轮廓序号来确定每个元件连接的相对位置关系,从而解决了 2 张图在拓扑关系检测时的序号命名问题。最终将子图匹配算法与节点遍历匹配方法的结果进行对比,验证本文方法具有更高的准确率。

2.4.1 子图匹配算法

将 2 张图抽象化为无向图后再进行匹配,最常见的方法是将 2 张图中的节点进行遍历,并根据节点值进行一一匹配,这种遍历方法在数据规模较大时,时间开销过大,并且不能很好地反映 2 张图之间的拓扑关系。

在对厂站一次接线图进行核验时,使用的是子图匹配算法,子图匹配是无向图中一种常见且实用的查询方法,其优势是:可以很好地利用节点间的连接关系,并可以大大减少遍历所需的时间消耗。这里采用的子图匹配算法是 VF2 算法^[19],其是一种子图同构算法,子图同构算法用于识别图中节点对之间的结构关系,属于精确匹配。Foggia 等^[20]对 Ullmann、SD、Nauty、VF 和 VF2 等算法进行了比较,认为 VF2 算法对稀疏图或小型图的效率较高,对于二维网状的规则图,VF2 算法的效率明显高于其他算法。由于厂站接线图的结构属于规则图,本文采用 VF2 算法会有更好的效率^[21]。VF2 算法通过在深度优先的搜索过程中利用高效剪枝的方法实现了子图匹配,最优情况下时间复杂度为 \$\Theta(N^2)\$,最差情况下时间复杂度为 \$\Theta(N! \cdot N)\$^[22]。

由于原始的 VF2 算法对于检测结果没有很

好的容忍度,需要检测精度足够高时才会有效,对 VF2 算法进行改进,将查询图划分为若干个不同的子图,将划分的子图分别与原图进行匹配。其中,划分的子图数量取决于前一步的拓扑关系识别得到的拓扑关系条数目。

2.4.2 接线图校核过程

1) 将 2 张图(target 与 query)拓扑结构检测的结果进行提取,将不同的元件命名为不同的节点,根据节点间的连接关系抽象为 2 张无向图。

2) 将 query 图划分为若干个子图,所有子图合并为一整个查询图集(querySet)。

3) 提取在拓扑关系检测中得到的线条轮廓序号,分别确定每个元件的相对位置关系。

4) 使用 VF2 算法对 querySet 中的每一个子图与 target 图进行一一匹配,根据匹配的子图数计算 2 张图的匹配率。

5) 将匹配率与最初设定好的匹配率阈值进行比较,若超过则手工图无误;若未得到 2 张图的映射关系或匹配率未达到阈值,则手工图有误,2 张图不匹配。

2.4.3 分析与总结

本模块主要对 2 张图的拓扑关系进行匹配,来辅助手工图的核验工作,其中使用了 VF2 算法来对 2 张接线图进行图匹配。本文对算法进行了改进,解决了原 VF2 算法对数据容忍度差的问题;提取了拓扑检测时的线条轮廓序号信息来确定每个元件的相对位置关系;解决了 2 张图中元件序号命名不统一的问题。最终将该方法与传统的遍历方法进行对比,可以看出准确率有明显提升。

3 变电站一次接线图数据集构建

本文的数据集是选取某供电公司部分变电站的一次接线图和人工维护的一次接线图作为研究样本。

3.1 数据集整理及标注

数据集共有 93 张变电站一次接线图,其中电子图 79 张,人工图 14 张,在图 5 和图 6 中展示了电子图和人工图的样例。对数据集图片进行人工标注,标注为 pascal voc 格式。标注的内容是厂站一次接线图中的电器元件类型和电器元件位置。通过数据标注软件 labelImg 对图片中各类电器元件进行标记框选,整理为 xml 文件作为接线图图片训练的标签。标注类别共 17 类,分别为避雷器(blq)、变压器(byq)、电抗(dk)、电力电感器(dldgq)、断路器(dlq)、电容(dr)、刀闸(dz)、隔离手车 1(glsc1)、隔离手车 2(glsc2)、接地(jd)、接地刀闸(jddz)、母线(mx)、电压互感器(pt)、熔断

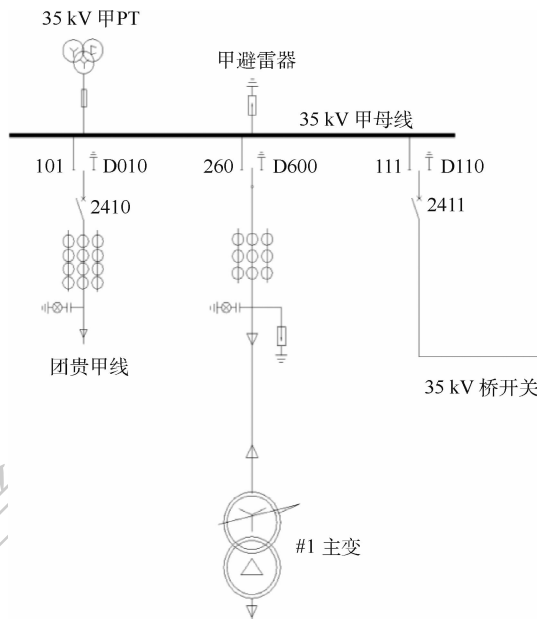


图 5 电子图示例

Fig. 5 Example of electronic wiring diagram

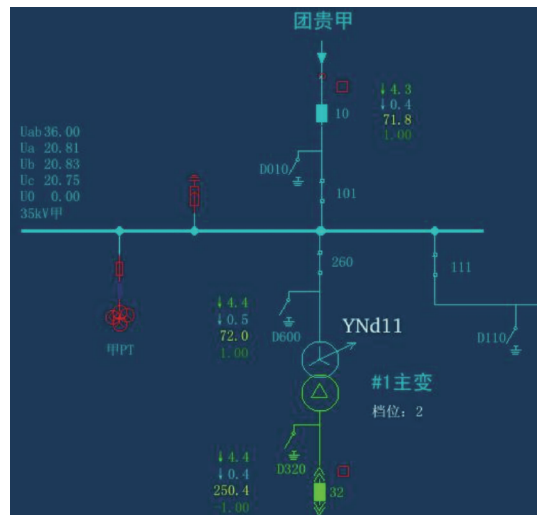


图 6 人工图示例

Fig. 6 Example of manual wiring diagram

器(rdq)、所变(sb)、手车开关(sckg)、忽略部件(ignore)。数据集大约共有 11 500 个标注框。

3.2 数据集处理

标注好的数据集,厂站一次接线图分辨率过大需要做分割处理,因此对于数据集中的接线图,采取分割图片及其对应 xml 文件的方法构建出新的分割图数据集。其中,分割图数据集共有 316 张电子图和 56 张人工图,类别是和接线图相同的 17 类。因此,数据集可以分为 2 部分,即接线图原图数据集和分割图数据集,以便应用于改进的检测算法中。

4 实验结果与分析

实验环境包括:Ubuntu 16.04.6 LTS 操作系统、NVIDIA TITAN XP 显卡、12196MiB 显存、

CUDA10.1、OpenCV2.4.9 等。

4.1 厂站一次接线图电器元件识别的实验结果

厂站一次接线图中的电器元件识别部分分别采用基于 Paddle 深度学习框架的 YOLOv3 算法和基于 Tensorflow 框架的 Faster R-CNN 模型。

对于厂站一次接线图中的电子图,采用 2 种数据集的形式,分别为接线图原图数据集和每张接线图分割为 4 份后的分割图数据集。对于 YOLOv3,直接使用分割图的数据集进行训练,训练集 244 张,测试集 72 张。对于 Faster R-CNN 模型,先采用原图数据集进行训练,其中训练集 63 张,测试集 16 张;再采用多尺度融合的检测算法,使用分割图的数据集训练出小图模型,并将两者的结果进行融合。

由于电子图和人工图略有差异,为了提高核验的效果,需对模型进行 finetune,使其更加适用于人工图的电器元件检测。对于大图模型,采用 11 张人工图作为训练集,3 张作为测试集。对于小图模型,采用 44 张分割的人工图作为训练集,12 张作为测试集。在基于训好的电子图的模型权重条件下,使用较少的人工图数据集进行 finetune。为了防止过拟合,使用更低的学习率 0.000 1 进行训练。

在训练的过程中,初始学习率设为 0.001, momentum 设为 0.9。YOLOv3 使用 darknet-53 的预训练权重,Faster R-CNN 选择 ResNet101 的预训练权重。YOLOv3 在迭代 140 000 epoch 后,loss 趋于稳定。Faster R-CNN 在迭代 100 000 epoch 后,loss 趋于稳定,数值在 0.3 左右,finetune 迭代 70 000 epoch 后 loss 稳定在 0.3 左右。表 1 为

表 1 电器元件识别实验结果对比

类别	检测准确率/%		
	YOLOv3	Faster R-CNN	Faster R-CNN 融合算法
断路器	100	84	85.6
接地刀闸	92.7	90.5	96.6
接地	85.6	72.7	87.6
电抗	85.2	89.3	97.9
刀闸	95.3	86.7	94.8
变压器	96.9	100	100
避雷器	93.1	80	94.4
母线	87.5	85.9	97.2
手车开关	89.1	95.1	96
熔断器	93.9	92.7	93.3
电容	82.6	66.5	88.8
隔离手车 1	87.2	0.72	76
隔离手车 2	92.2	58.4	81.8
所变	94.6	86.1	88.9
电压互感器	96.4	88.6	85.7
电力电感器	75.4	83.3	94.4
忽略部件	80	66.7	80
平均	89.5	83.5	92

YOLOv3 和 Faster R-CNN 实验结果的对比,采取的评判标准是检测准确率。将 IoU 大于 0.5 且类名正确的样本视为正样本,通过计算正样本和总样本的比作为检测准确率。其中,17 种待识别的器件在数据集中的数量分布并不均衡,表 1 中的平均检测准确率是通过计算测试集中图片的所有种类器件正样本和总样本中的比得来的。虽然 YOLOv3 的检测准确率结果中有 7 种器件优于 Faster R-CNN 融合算法,如隔离手车、所变、电压互感器等器件,但这些器件在接线图中出现的频率并不高,所以 YOLOv3 的检测效果并不如 Faster R-CNN 融合算法。根据得出的平均检测准确率结果可以看出,多尺度融合后的 Faster R-CNN 算法效果最佳。

4.2 厂站一次接线图文字识别的实验结果

厂站一次接线图中的文字、数字识别部分分别采用了 2 种方法。两阶段的识别方法有 Advanced EAST 和 Tesseract OCR。端到端的识别方法有 Attention-OCR 方法。

首先,采用 Advanced EAST 模型和 Tesseract OCR 模型。采用标注文本框的数据集训练基于 Keras 框架的 Advanced EAST 模型;获取出接线图文字区域图片,通过 jTessBoxEditor 对文字图片中的每个字符进行标注,使用 Tesseract OCR 构建接线图的专属字库。

然后,采用端到端的 Attention-OCR 方法,其基于 Tensorflow 框架,其中 Cascade R-CNN 采用 ResNet101-FPN 作为骨干网络,预训练权重来自 ICDAR2017RCTW^[23]数据集。Attention-OCR 使用 InceptionV4^[24]作为骨干网络,预训练模型由 ImageNet 数据集得来,之后通过多个包含不规则多方向的文本数据集(如 LSVT、COCO-Text、ICDAR2017 等)训练模型。

表 2 为 2 种方法的结果对比。文本区域检测采用目标检测的评判标准,以 IoU 大于 0.5 的样本判断为正样本。文本识别部分以每一个字符的准确率作为评判标准。可以看出,无论是文本区域检测还是文本识别,Attention-OCR 方法的效果都要优于 Advanced EAST 和 Tesseract-OCR 的结果。

表 2 文字识别实验结果对比

检测算法	文本区域 检测准确率/%	文本识别 准确率/%
	检测准确率/%	准确率/%
Advanced EAST + Tesseract OCR	76.6	78
Attention-OCR	94.2	92

4.3 厂站一次接线图拓扑关系识别的实验结果

厂站一次接线图拓扑关系识别主要采用的是图像处理技术,其拓扑关系识别的准确率较为依靠电器元件识别的准确率。对于电子图和人工图,在进行图像预处理的过程中选择了更加适用自身的不同参数。对每张测试图片的拓扑关系进行人工标注,在对测试图片进行拓扑关系准确率计算后得出准确率约为 87.7%。其具体情况如表 3 所示。

表 3 拓扑关系识别测试结果

Table 3 Test results of topological relation recognition			
图片名称	正确的连接关系数	连接关系总数	准确率/%
10097	71	85	83.53
10099	115	119	96.64
10125	93	103	90.29
10129	67	80	83.75
20008	88	92	95.65
20009	68	104	65.38
20014	149	151	98.68

4.4 厂站一次接线图自动校核的实验结果

人工维护的厂站一次接线图自动校核技术采用的是子图匹配算法,选用的子图匹配算法是 VF2 算法。由于子图匹配算法本身存在对数据容忍度差的问题,因此对算法进行了一定程度的改进。通过自动校核技术来对接线图及其对应的人工图的人工校验提供辅助,输入为 2 张图的拓扑关系识别结果,输出为 2 张图的匹配率。这里没有使用改进前的 VF2 算法的匹配结果,因为原 VF2 算法用于拓扑关系匹配时,只能用于判断查询图是否为目标图的子图。由于拓扑关系识别不能完全精准地还原 2 张图的拓扑关系,一旦有一个节点不匹配,VF2 算法输出的结果便是 0,这也是对 VF2 算法进行改进的原因之一。从测试数据中挑选出 4 组清晰且完好的测试图,分别对这 4 组数据进行测试,分别使用遍历的方法及改进的 VF2 算法 2 种方法,通过计算匹配的子图数与总子图数的比例得到其对应的匹配率,结果如表 4 所示。可以看出,改进的 VF2 算法在 8 组数据上的匹配率均在 50% 以上,明显高于遍历方法匹配的结果。将匹配的阈值设定为 0.6,即匹配率超过该阈值则判定这一组电子图与人工图正确匹配,否则将判定为错误匹配。此时,改进的 VF2 算法对于厂站一次接线图的自动校核对数据有更好的稳定性和更好的准确率,其准确率为 62.5%,而遍历方法的准确率较低只有 25%(即 8 组结果中只有 2 组的匹配率高于阈值 0.6)。

将电子图和人工图的匹配结果可视化,绿色框表示匹配部分,未匹配的元件在原图中用红框的形式标出,匹配可视化结果如图 7 和图 8 所示。

表 4 两种方法的匹配率结果

Table 4 Matching rate results of two methods

测试组别	匹配率/%	
	节点遍历方法	改进的 VF2 算法
第 1 组	27.97	55.95
第 2 组	37.50	51.61
第 3 组	0.74	54.74
第 4 组	62.07	63.49
第 5 组	24.42	76.67
第 6 组	9.33	92.00
第 7 组	93.24	92.86
第 8 组	0	71.69

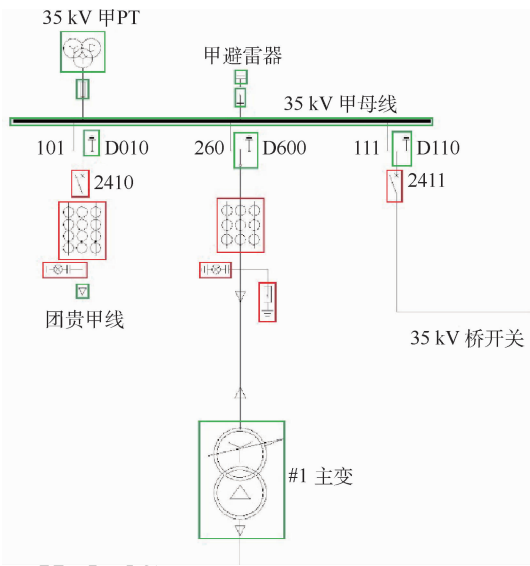


图 7 电子图匹配结果

Fig. 7 Matching results of electronic wiring diagram

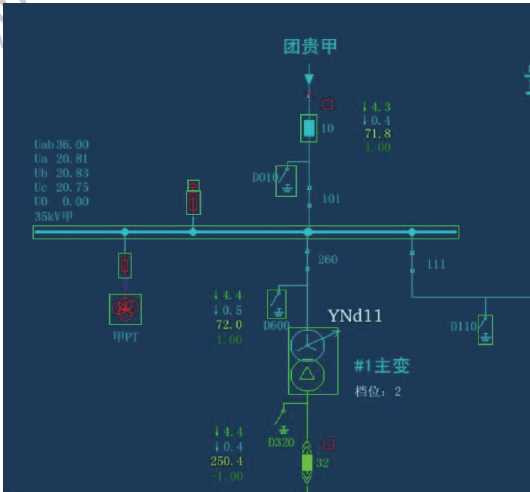


图 8 人工图匹配结果

Fig. 8 Matching results of manual wiring diagram

5 结 论

1) 本文实现了厂站一次接线图的自动识别和自动校核,极大提高了电力厂站一次接线图绘制和管理工作的自动化程度。

2) 对厂站一次接线图中电器元件的目标检测、文字识别和拓扑关系检测均可达到较高的准确率。

3) 改进的 VF2 算法相比于节点遍历的匹配方法,可以提高 37.5% 的准确率,大大降低了人工干预的成本。

目前,对于厂站一次接线图的自动识别和自动校核工作仍有些许不足。对接线图电器元件的目标识别和文字识别仍需继续提高准确率。对接线图的拓扑关系识别及自动核验需要加快对图片的处理速度,提高厂站一次接线图的自动识别和自动校核的效率。未来将及时关注最新的目标检测、文本检测和文本识别等算法的发展,改进方法提高检测和识别的准确率。对于拓扑关系识别和自动核验,将继续改进算法提高对检测结果的容忍度,提高识别和核验的效率。

参考文献 (References)

[1] 刘建,吴媛,刘凯权. 配电馈线地理图到电气接线图的转换[J]. 电力系统自动化,2005,29(14):73-77.
LIU J, WU Y, LIU G Q. Conversion form GIS based feeder maps to electric diagrams[J]. Automation of Electrical Power Systems,2005,29(14):73-77 (in Chinese).

[2] 沙树名,林峰. 一种基于 CIM 的厂站接线图自动生成技术[J]. 电力系统自动化,2008,32(21):68-71.
SHA S M, LIN F. A technique of automatic generation of substation wiring diagram based on CIM[J]. Automation of Electric Power Systems,2008,32(21):68-71 (in Chinese).

[3] 苏运光,罗俊,熊浩,等. 基于存量图形拓扑分析的厂站接线图自动生成[J]. 电力系统自动化,2020,44(3):162-168.
SU Y G, LUO J, XIONG H, et al. Automatic generation of plant and station wiring diagram based on stock graphic topology analysis[J]. Automation of Electric Power Systems,2020,44(3):162-168 (in Chinese).

[4] KLUMP R, SCHOOLEY D, OVERBYE T. Advanced visualization platform for real-time power system operations[C]// Proceedings of the 14th Computation,2002:1-8.

[5] 朱永利,栗然,刘艳. 电力系统厂站主接线图的自动生成[J]. 中国电力,1999,32(11):56-58.
ZHU Y L, LI R, LIU Y. Automatic generation of main wiring diagram of power system plant[J]. China Electric Power, 1999, 32(11):56-58 (in Chinese).

[6] 黄若航. 深度学习在热力管道工程设计上的研究和应用[D]. 成都:电子科技大学,2019.
HUANG R H. Research and application of deep learning in

thermal pipeline engineering design[D]. Chengdu: University of Electronic Science and Technology,2019 (in Chinese).

[7] 尹思宇. 深度学习在变电站电力设备缺陷检测中的应用研究[D]. 成都:电子科技大学,2020.
YIN S Y. Research on application of deep learning in defect detection of power equipment in substation[D]. Chengdu: University of Electronic Science and Technology,2020 (in Chinese).

[8] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL]. (2018-04-08) [2020-08-01]. <https://arxiv.org/abs/1804.02767>.

[9] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press,2017:17355115.

[10] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1149.

[11] ZHOU X, YAO C, WEN H, et al. EAST: An efficient and accurate scene text detector[C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press,2017:17355294.

[12] SMITH R. An overview of the Tesseract OCR engine[C]// Ninth International Conference on Document Analysis and Recognition. Piscataway: IEEE Press,2007:9870037.

[13] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04) [2020-08-01]. <https://arxiv.org/abs/1409.1556>.

[14] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[EB/OL]. (2015-05-18) [2020-08-01]. <https://arxiv.org/abs/1505.04597>.

[15] WOINA Z, GORBAN A N, LEE D S, et al. Attention-based extraction of structured information from street view imagery[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE Press, 2017,1:844-850.

[16] CAI Z, VASCONCELOS N. Cascade R-CNN: Delving into high quality object detection[EB/OL]. (2017-12-03) [2020-08-01]. <https://arxiv.org/abs/1712.00726>.

[17] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// Proceedings of the 32nd International Conference on Machine Learning,2015,37:2048-2057.

[18] SUZUKI S, BE K. Topological structural analysis of digitized binary images by border following[J]. Computer Vision Graphics and Image Processing,1985,30(1):32-46.

[19] CORDELLA L P, FOGGIA P, SANSONE C, et al. A (sub) graph isomorphism algorithm for matching large graphs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2004,26(10):1367-1372.

[20] FOGGIA P, SANSONE C, VENTO M. A database of graphs for isomorphism and sub-graph isomorphism benchmarking[C]// Proceedings of the 3rd IAPR-TC-15 International Workshop on Graph-based Representations,2001:176-188.

[21] 李欣,宋婷婷,何险峰. 基于 VF2 算法的分子二维子结构检

索[J]. 计算机与应用化学, 2007, 24(11): 1551-1554.

LI X, SONG T T, HE X F. Molecular two-dimensional substructure retrieval based on VF2 algorithm [J]. Computer and Applied Chemistry, 2007, 24(11): 1551-1554 (in Chinese).

[22] 楼昀恺, 王朝坤. 使用社区结构信息的子图匹配算法优化方法[J]. 计算机科学与探索, 2019, 13(1): 1-22.

LOU Y K, WANG C K. Optimization method of subgraph matching algorithm using community structure information[J]. Computer Science and Exploration, 2019, 13(1): 1-22 (in Chinese).

[23] SHI B, YAO C, LIAO M, et al. ICDAR2017 competition on reading Chinese text in the wild (RCTW-17) [C] // 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). Piscataway: IEEE Press, 2017, 1: 1429-1434.

[24] CHRISTIAN S, SERHEY I, VINCENT V, et al. Inception-v4, inception-resnet and the impact of residual connections on

learning[C] // Thirty-First AAAI Conference on Artificial Intelligence, 2017: 1-12.

作者简介:

李昊 男, 硕士, 工程师。主要研究方向: 电力调度自动化。

王杉 男, 技师。主要研究方向: 电力调度自动化。

耿玉杰 男, 硕士, 高级工程师。主要研究方向: 电力调度自动化。

王黎 女, 硕士, 高级工程师。主要研究方向: 电力调度自动化与继电保护。

孙文昌 男, 硕士研究生。主要研究方向: 计算机视觉。

苗纯源 男, 硕士研究生。主要研究方向: 计算机视觉。

Wiring diagram detection and check based on deep learning and graph matching

LI Hao^{1,*}, WANG Shan¹, GENG Yujie², WANG Li¹, SUN Wenchang³, MIAO Chunyuan³

- (1. Electric Power Dispatching & Control Center, State Grid Qingdao Power Supply Company, Qingdao 266002, China;
- 2. Electric Power Dispatching & Control Center, State Grid Shandong Electric Power Company, Jinan 250001, China;
- 3. School of Computer Science and Technology, Shandong University (Qingdao), Qingdao 266237, China)

Abstract: The drawing and management of the traditional primary wiring diagram of plant and station mainly depend on the operators of power grid, which wastes time and labor and lacks scientific and verifiable reference standards. Towards this problem, based on the deep neural networks and digital image processing, we propose an algorithm for automatic detection, recognition and verification. Specifically, Faster R-CNN is first adopted to detect the electrical components of wiring diagram with 92% detection accuracy. In the meantime, an end-to-end text detection and recognition model is used to recognize the text with 94.2% detection accuracy and 92% character recognition accuracy. Then we take advantage of digital image processing technique to identify the connection of wiring diagram and topological relation. Finally, improved graph matching algorithm VF2 is used to check the difference between the electronic and manually maintained diagrams. The topological data is abstracted into an undirected graph, and the relative position information of components is obtained through the outline number. Based on the improved VF2 algorithm, we can compute the matching rate of two graphs to help the verification. Compared with the matching method of node traversal, the verification accuracy can be improved by 37.5%. Based on the first wiring diagram of some substations provided by a power supply company, this paper marks the electrical components of wiring diagram and contributes a small wiring diagram dataset.

Keywords: primary wiring diagram of plant and station; topological relation matching; automatic generation; automatic check; deep learning; digital image processing

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

DOI: 10.13700/j.bh.1001-5965.2020.0498

基于记忆关联学习的小样本高光谱图像分类方法



王聪^{1,2}, 张锦阳², 张磊², 魏巍^{1,2,*}, 张艳宁²

(1. 西北工业大学 深圳研究院, 深圳 518057; 2. 西北工业大学 计算机学院, 西安 710129)

摘

要: 高光谱图像(HSI)分类是遥感领域的基础应用之一。该任务旨在根据部分带类别标签的像素样本训练分类器,预测图像中剩余像素对应的类别标签。在实际应用中,由于人工标记样本成本过高,只能获得少量带标签的样本。针对少量样本无法准确描述数据分布从而导致训练过程过拟合的问题,提出一种基于记忆关联学习的小样本高光谱图像分类方法。考虑到无标签样本中包含大量与数据分布相关的信息,构建基于有标签样本记忆模块,并根据样本间的特征关联,利用不断更新的记忆模块学习无标签样本的潜在类别分布,构建无监督分类模型,并传统的有监督分类模型进行联合学习。在多个高光谱图像分类数据集上的实验结果表明,所提方法能有效提升小样本高光谱图像分类的准确性。

关键词: 记忆关联学习; 半监督; 小样本; 高光谱图像(HSI); 分类

中图分类号: V221⁺.3; TB553

文献标志码: A

文章编号: 1001-5965(2021)03-0549-09

高光谱图像(Hyperspectral Image, HSI)是一种通过高光谱传感器探测地物反射的电磁波而获得的数字图像。在常规的彩色或者灰度图像中,每一个像素仅包含一个或者三个离散的能量值,而高光谱图像中的每一个像素包含几十甚至数百个光谱波段,每一个像素记录了对应地物的光谱曲线^[1]。由于不同地物通常呈现出不同的光谱曲线,使得高光谱图像有效地集合了成像场景中的空间信息和光谱信息,利用高光谱图像中丰富的空间及光谱信息有助于更加准确地区分成像场景中的不同物质。随着遥感技术和传感器技术的进步,获取的高光谱遥感图像数据量在不断增加,使得高光谱图像被广泛应用于资源勘探^[2]、环境监测^[3]、伪装目标检测^[4]等诸多民用和军事领域。

在众多遥感应用中,高光谱图像分类是其中的一项重要应用^[4-6]。高光谱图像分类旨在根据部分带标签的像素样本训练分类器,预测图像中

其他像素样本对应的标签,获得图像中不同地物的空间分布情况。传统的高光谱图像分类方法包括K近邻(K-Nearest Neighbor, KNN)^[7]、支持向量机(Support Vector Machine, SVM)^[8]和基于稀疏表示(Sparse Representation)^[9]的方法等。这些模型结构浅、速度快,但是表征能力弱,无法充分挖掘像素光谱中隐含的非线性、可区分特征,在复杂情况下的泛化能力有限。

近期研究表明,深度神经网络具有强大的非线性表征能力,能够有效提升分类识别性能。因此,大量工作开始研究基于深度神经网络的高光谱图像分类方法。例如,基于一维卷积神经网络(1D-CNN)的方法^[10]利用端到端的卷积神经网络提取每一个像素光谱维度的特征并训练分类器。基于二维卷积神经网络(2D-CNN)的方法^[11]使用二维卷积网络对局部的空间像素块进行高维特征提取。基于三维卷积神经网络(3D-CNN)的方

收稿日期: 2020-09-07; 录用日期: 2020-09-11; 网络出版时间: 2020-11-04 17:23

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201103.1825.002.html

基金项目: 深圳市科技创新委员会基金(JCYJ20190806160210899); 国家自然科学基金(61671385, U19B2037)

*通信作者: E-mail: weiweinwu@nwpu.edu.cn

引用格式: 王聪, 张锦阳, 张磊, 等. 基于记忆关联学习的小样本高光谱图像分类方法[J]. 北京航空航天大学学报, 2021, 47(3): 549-557. WANG C, ZHANG J Y, ZHANG L, et al. Small sample hyperspectral image classification method based on memory association learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 549-557 (in Chinese).

法^[12]同时对高光谱图像的空间和光谱维进行特征表示。在此基础上,基于残差网络(Residual Network)^[13]、密集连接网络(Dense Convolutional Network)^[14]和胶囊网络(Capsule Network)^[15]的模型也相继被应用到高光谱图像分类任务中。虽然基于深度神经网络的方法显著提升了分类性能,但是需要大量的带有类别标签的像素样本进行模型训练。

在实际高光谱图像分类应用中,人工标注像素样本耗时、费力、成本高昂,致使能够获取的带标签像素样本极少。少量带标签的训练样本无法充分地描述数据分布,从而导致基于深度神经网络的方法产生过拟合现象,分类精确度大幅下降。近年来,虽然有少量的工作提出了一些解决方案^[16-18],但是未能充分地挖掘无标签样本中的结构信息,分类效果依然有限。因此,研究小样本高光谱图像分类问题是非常有必要且具有挑战性的。

为了解决这一问题,本文提出一种基于记忆关联学习的小样本高光谱图像分类方法。首先,利用有标签样本构建记忆模块,并根据样本间的特征关联,学习无标签样本的潜在类别分布。然后,在深度神经网络模型的基础上,利用学习的类别分布在无标签样本上构建无监督学习任务,并与有标签样本上的有监督学习任务进行联合优化。由于无监督学习任务能够有效约束有监督学习过程,避免了小样本情况下训练深度神经网络模型导致的过拟合现象。在多个高光谱图像分类数据集上,本文方法有效提升了小样本高光谱图像分类的准确性。

本文贡献主要包括以下3点:

- 1) 将记忆模块引入高光谱图像分类任务中,基于有标签样本,通过记忆关联学习,预测无标签样本的潜在类别分布。
- 2) 构建联合学习分类模型,其中根据预测的类别分布在无标签样本上构建无监督分类模型,同时有标签样本上构建有监督分类模型,并对2个模型进行联合求解。
- 3) 在3个高光谱图像分类数据集上的实验结果证明了本文方法能有效提升小样本高光谱图像分类的准确性。

1 基于记忆关联学习的高光谱图像分类

本节详细介绍提出的基于记忆关联学习的小样本高光谱图像分类方法,整体流程如图1所示,共包括3个模块:特征提取模块、记忆模块及学习

模块。

在介绍具体的模块之前,先对本文中使用的符号进行介绍。 $\mathbf{X}_l = \{\mathbf{x}_{1,l}, \dots, \mathbf{x}_{n_l,l}\} \in \mathbf{R}^{n_b \times n_l}$ 表示高光谱图像标签样本集, n_b 为波段数, n_l 为标签样本数量, $\mathbf{Y}_l = \{y_{1,l}, \dots, y_{n_l,l}\} \in \mathbf{R}^{1 \times n_l}$ 表示标签样本的标签, $y_{i,l} \in \{1, 2, \dots, C\}$, C 为数据集中的类别个数, $\mathbf{X}_u = \{\mathbf{x}_{1,u}, \dots, \mathbf{x}_{n_u,u}\} \in \mathbf{R}^{n_b \times n_u}$ 表示无标签样本集, n_u 为无标签样本数量, $\mathbf{Y}_u = \{y_{1,u}, \dots, y_{n_u,u}\} \in \mathbf{R}^{1 \times n_u}$, $y_{i,u} \in \{1, 2, \dots, C\}$ 。其中,标签样本数量远远小于无标签样本数量 $n_l \ll n_u$ 。为方便,下文用 $\mathbf{x} \in \{\mathbf{X}_l, \mathbf{X}_u\}$ 表示标签样本或者无标签样本, $y \in \{\mathbf{Y}_l, \mathbf{Y}_u\}$ 表示样本对应标签。

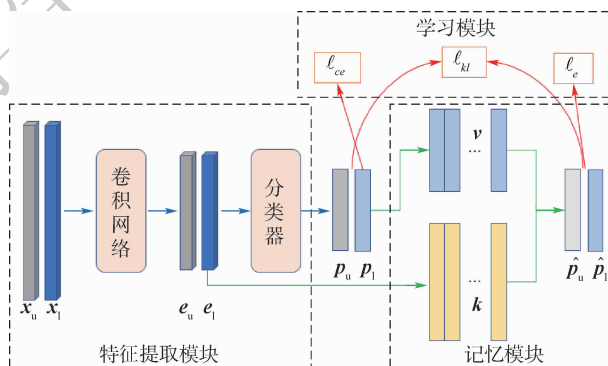


图1 基于记忆关联学习的小样本高光谱图像分类方法流程

Fig. 1 Flowchart of small sample hyperspectral image classification method based on memory association learning

1.1 特征提取模块

特征提取模块由卷积网络和分类器2部分组成。卷积网络将高光谱图像通过卷积神经网络映射到高维特征空间,分类器将高维特征映射到类别概率空间。具体来说,卷积网络包括6个卷积层(Convolutional Layer),其中每一个卷积层由3个连续的子模块构成,分别为卷积子模块、批标准化模块(Batch Normalization, BN)及非线性激活函数模块(Rectified Linear Unit, ReLU)。为了同时捕获高光谱图像的空间和光谱特征,使用三维卷积对图像进行特征提取。分类器层包含2个级联的子模块,分别为全连接子模块和归一化指数(Softmax)激活函数模块。

对于输入的有标签或者无标签高光谱图像数据 $\mathbf{x} \in \{\mathbf{X}_l, \mathbf{X}_u\}$,通过卷积网络对其进行特征提取,获得对应的特征向量。

$$\mathbf{e} = f(\mathbf{x}) \quad (1)$$

式中: $f(\cdot)$ 表示卷积神经网络模型; $\mathbf{e} \in \{\mathbf{E}_l, \mathbf{E}_u\}$, $\mathbf{E}_l = \{\mathbf{e}_{1,l}, \dots, \mathbf{e}_{n_l,l}\} \in \mathbf{R}^{d \times n_l}$ 为有标签样本对应的特征, $\mathbf{E}_u = \{\mathbf{e}_{1,u}, \dots, \mathbf{e}_{n_u,u}\} \in \mathbf{R}^{d \times n_u}$ 为无标签样本

对应的特征, d 为特征维度。

通过卷积网络获得特征后,进一步引入一个分类器层,输出每个样本的类别预测概率向量。

$$\mathbf{p} = g(\mathbf{e}) \quad (2)$$

式中: $g(\cdot)$ 表示分类器模型; $\mathbf{p} \in \{\mathbf{P}_l, \mathbf{P}_u\}$, $\mathbf{P}_l = \{\mathbf{p}_{1,l}, \dots, \mathbf{p}_{n_l,l}\} \in \mathbf{R}^{C \times n_l}$ 为有标签样本的预测概率, $\mathbf{P}_u = \{\mathbf{p}_{1,u}, \dots, \mathbf{p}_{n_u,u}\} \in \mathbf{R}^{C \times n_u}$ 为无标签样本的预测概率。

1.2 记忆模块

记忆关联学习通过存储学习到的知识,并将其抽象成对任务有益的结构化信息,提升当前任务的性能。为了充分挖掘无标签样本中的结构信息,构建了基于有标签样本的记忆模块,在每一次训练过程中,对类间信息进行动态地更新和存储,并根据样本间的特征关联,预测无标签样本的类别分布。

本节构建的记忆模块包括 2 部分,分别为类别特征空间 $\mathbf{K} = \{\mathbf{k}_1, \mathbf{k}_2, \dots, \mathbf{k}_c\}$ 和类别概率分布特征空间 $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_c\}$, $\mathbf{k}_i \in \mathbf{R}^d$ 表示第 i 类的类别特征中心, $\mathbf{v}_i \in \mathbf{R}^C$ 表示第 i 类的概率分布特征中心。本文中仅使用带标签的数据对记忆模块进行更新,以避免无标签数据的类别预测不确定性对模型的干扰。对于有标签样本,通过特征提取模块得到对应的特征向量和类别概率向量,记忆模块中 $\mathbf{k}_i$ 由 n_i 个第 i 类样本的特征向量更新得到, $\mathbf{v}_i$ 由 n_i 个第 i 类样本的类别概率向量更新得到。具体更新公式为

$$\mathbf{k}_i \leftarrow \mathbf{k}_i - \eta \nabla \mathbf{k}_i \quad (3)$$

$$\nabla \mathbf{k}_i = \frac{\sum_{j=1}^{n_i} \mathbf{1}[y = i] (\mathbf{k}_i - \mathbf{e}_{j,l})}{\text{sum}(\mathbf{1}[y = i])} \quad (4)$$

$$\mathbf{v}_i \leftarrow \frac{\mathbf{v}_i - \eta \nabla \mathbf{v}_i}{\sum_{j=1}^c (\mathbf{v}_{i,j} - \eta \nabla \mathbf{v}_{i,j})} \quad (5)$$

$$\nabla \mathbf{v}_i = \frac{\sum_{j=1}^{n_i} \mathbf{1}[y = i] (\mathbf{v}_i - \mathbf{p}_{j,l})}{\text{sum}(\mathbf{1}[y = i])} \quad (6)$$

式中: η 为更新系数; $\mathbf{1}[y = i]$ 表示指示函数。为了公平对待所有类别,同时不引入任何类别先验, $\mathbf{K}$ 初始化为 0, $\mathbf{V}$ 初始化为 $1/C$ 。

对于有标签样本或者无标签样本 $\mathbf{x}$ 对应的类别概率向量 $\mathbf{p}$, 利用更新的记忆模块进一步获得预测的类别概率向量 $\hat{\mathbf{p}}$:

$$\hat{\mathbf{p}} = \sum_{i=1}^c w(\mathbf{k}_i | \mathbf{x}) \mathbf{v}_i \quad (7)$$

式中: $w(\mathbf{k}_i | \mathbf{x})$ 表示样本 $\mathbf{x}$ 到记忆模块中每个类

别中心 $\mathbf{k}_i$ 的权重系数。

对于有标签样本,权重 $w(\mathbf{k}_i | \mathbf{x})$ 定义为

$$w(\mathbf{k}_i | \mathbf{x}) = \begin{cases} 1 & i = y \\ 0 & i \neq y \end{cases} \quad (8)$$

对于无标签样本,通过计算样本的特征向量 $\mathbf{e}$ 到每个类别中心 $\mathbf{k}_i$ 的距离得到权重 $w(\mathbf{k}_i | \mathbf{x})$ 。

$$w(\mathbf{k}_i | \mathbf{x}) = \frac{e^{-\text{dist}(\mathbf{e}, \mathbf{k}_i)}}{\sum_{j=1}^c e^{-\text{dist}(\mathbf{e}, \mathbf{k}_j)}} \quad (9)$$

式中: $\text{dist}(\cdot)$ 表示余弦距离。

1.3 学习模块

根据预测的类别分布,在有标签样本上构建有监督损失,同时,在无标签样本上设计无监督损失,通过联合求解有监督及无监督损失函数,优化分类模型参数。

1.3.1 监督损失

对于有标签样本 $\mathbf{x}$, 经过特征提取模块得到对应的类别概率向量 $\mathbf{p}$ 后,通过交叉熵损失:

$$\ell_{ce} = - \sum_{i=1}^c \mathbf{1}[y = i] \ln p_i \quad (10)$$

1.3.2 无监督损失

本文中无监督损失由信息熵和相对熵共同定义。信息熵用来衡量信息量的大小,事件发生的概率越低,其发生时所能给出的信息量越大,即如果模型不能以较高的置信度预测样本类别,那么该模型的样本类别信息熵较大。通过最小化记忆模块预测的类别概率向量的熵,可以对模型的预测不确定性进行惩罚。

对于有标签或者无标签样本 $\mathbf{x}$, 通过特征提取模块得到对应类别概率向量 $\mathbf{p}$, 利用记忆模块得到类别概率向量 $\hat{\mathbf{p}}$, 信息熵损失定义为

$$\ell_e = H(\hat{\mathbf{p}}) = - \sum_{i=1}^c \hat{p}_i \ln \hat{p}_i \quad (11)$$

相对熵又称 KL 散度,是 2 个概率分布间差异的非对称性度量。在信息论中,相对熵等价于 2 个概率分布的信息熵的差值,若其中一个概率分布为真实分布,另一个为拟合分布,则此时相对熵等于交叉熵与真实分布的信息熵之差,表示使用拟合分布拟合真实分布时产生的信息损耗。为了使特征提取模块预测的类别概率向量 $\mathbf{p}$ 和记忆模块预测的类别概率向量 $\hat{\mathbf{p}}$ 的分布趋于一致,通过最小化相对熵,对 2 个分布之间的差异进行惩罚。相对熵损失定义如下:

$$\ell_{kl} = \max(\hat{\mathbf{p}}) \text{KL}(\mathbf{p} \| \hat{\mathbf{p}}) \quad (12)$$

$$\text{KL}(\mathbf{p} \| \hat{\mathbf{p}}) = \sum_{i=1}^c p_i \ln \left(\frac{p_i}{\hat{p}_i} \right) \quad (13)$$

式中: $KL(\boldsymbol{p}||\hat{\boldsymbol{p}})$ 为KL散度; $\max(\hat{\boldsymbol{p}})$ 表示KL散度的动态调节系数,该系数可以减小2个分布之间的差异,降低KL散度在总损失中的比重,使模型更专注于分类任务。

因此,总的无监督损失定义为

$$\ell_u = \mu_1 \ell_e + \mu_2 \ell_{kl}$$

(14)

式中: μ_1 和 μ_2 表示信息熵损失和相对熵之间的均衡系数,实验中 μ_1 和 μ_2 均设置为1。

综上,学习模型总的损失函数定义为

$$\ell = \ell_{ce} + \ell_u$$

(15)

在模型的训练过程中,需要利用记忆模块及学习模块进行记忆关联学习,但是在测试阶段只需要特征提取模块得到的类别概率向量,不再需要记忆模块和学习模块,进而减少额外的存储和计算消耗。

2 实验结果及分析

本节在多个高光谱图像分类数据集上设计实验,对比本文方法和现有的其他分类方法以验证提出方法的有效性。

2.1 数据集

本文实验中采用3个标准的高光谱图像分类数据集作为实验数据,分别为Indian Pines数据集、Pavia University数据集和Salinas数据集,数据集中样本类别和样本数如表1~表3所示,数据和真实分类结果(Ground Truth,GT)如图2所示。Indian Pines数据集由AVIRIS卫星在美国Indiana州西北部拍摄地面场景得到。该数据空间包含 145×145 个像素,在波长 $0.38 \sim 2.5 \mu\text{m}$ 的光谱范围内,共捕获220个不同的光谱波段,空间分辨率为20 m,该数据场景中共包含16种不同的地物。在实验中,参考文献[19],移除该数据中的20个水吸收波段,将余下的200个波段作为实验数据。Pavia University数据集由ROSIS卫星在意大利Pavia University拍摄得到。该数据空间包含 610×340 个像素,共包含115个光谱波段,覆盖的光谱波长范围为 $0.43 \sim 0.86 \mu\text{m}$,空间分辨率为1.3 m。图像场景内包含9种不同的地物类别。本文移除该数据中的12个水吸收波段,将余下的103个波段作为实验数据。Salinas数据集也是由AVIRIS卫星拍摄,场景内容为美国California州的Salinas峡谷。该数据集总共包含224个波段,空间上共有 512×217 个像素,空间分辨率为3.7 m。这些像素分别来自于16种不同的地物类别。实验中,移除该数据中的20个水吸收波段,将剩下的204个波段作为实验数据。

表1 Indian Pines数据集中样本类别和样本数
Table 1 Name and pixel numbers of per adopted class on Indian Pines datasets

序号	样本类别名称	样本数
1	Alfalfa	46
2	Corn-notill	1 428
3	Corn-mintill	830
4	Corn	237
5	Grass-pasture	483
6	Grass-trees	730
7	Grass-pasture-mowed	28
8	Hay-windrowed	478
9	Oats	20
10	Soybean-notill	972
11	Soybean-mintill	2 455
12	Soybean-clean	593
13	Wheat	205
14	Woods	1 265
15	Buildings-Grass-Trees-Drives	386
16	Stone-Steel-Towers	93
总计		10 249

表2 Pavia University数据集中样本类别和样本数
Table 2 Name and pixel numbers of per adopted class on Pavia University datasets

序号	样本类别名称	样本数
1	Asphalt	6 631
2	Meadows	18 649
3	Gravel	2 099
4	Trees	3 064
5	Painted metal sheets	1 345
6	Bare Soil	5 029
7	Bitumen	1 330
8	Self-Blocking Bricks	3 682
9	Shadows	947
总计		42 776

表3 Salinas数据集中样本类别和样本数
Table 3 Name and pixel numbers of per adopted class on Salinas datasets

序号	样本类别名称	样本数
1	Brocoli_green_weeds_1	2 009
2	Brocoli_green_weeds_2	3 726
3	Fallow	1 976
4	Fallow_rough_plow	1 394
5	Fallow_smooth	2 678
6	Stubble	3 959
7	Celery	3 579
8	Grapes_untrained	11 271
9	Soil_vinyard_develop	6 203
10	Corn_senesced_green_weeds	3 278
11	Lettuce_romaine_4wk	1 068
12	Lettuce_romaine_5wk	1 927
13	Lettuce_romaine_6wk	916
14	Lettuce_romaine_7wk	1 070
15	Vinyard_untrained	7 268
16	Vinyard_vertical_trellis	1 807
总计		54 129

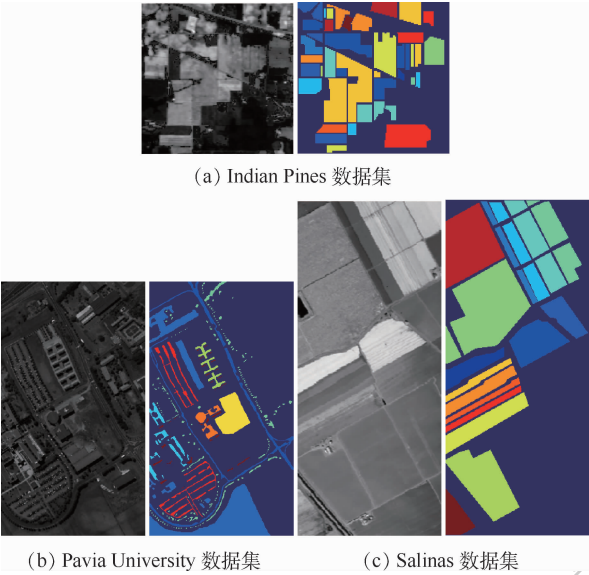


图 2 三个高光谱图像分类数据集的代表性波段图像和真实分类结果

Fig. 2 Representative band image and ground truth classification map of three HSI classification datasets

2.2 对比方法及实验设置

为了验证本文方法能够有效减轻深度神经网络在面对少量训练数据时产生的过拟合问题,对比了 6 种不同的分类方法,分别为 KNN 方法^[7]、SVM 方法^[8]、基于深度学习的一维卷积模型(1D-CNN)和三维卷积模型(3D-CNN)的方法,以及半监督卷积神经网络分类方法(SS-CNN)^[20]和基于类内相似结构表示的高光谱图像小样本分类方法(ISSR-HIC)^[18]。本文提出的基于记忆关联学习的高光谱图像分类方法的特征提取模块中卷积网络的整体结构如表 4 所示, s 表示第 5 个卷积单元光谱维的维度,valid 表示不进行特征补齐。分类器层如表 5 所示。一维卷积模型与本文方法的特征提取模块不同的是:一维卷积模型使用一维卷积对图像进行特征提取。三维卷积模型的网络结构和层数与本文方法的特征提取模块均相同。

表 4 卷积网络

Table 4 Convolution network

卷积层	卷积核	卷积步长	特征数	补齐
conv1	(3,1,8)	(1,1,3)	64	valid
conv2	(1,1,3)	(1,1,2)	64	valid
conv3	(1,3,3)	(1,1,2)	128	valid
conv4	(1,1,3)	(1,1,2)	128	valid
conv5	(1,1,3)	(1,1,2)	256	valid
conv6	(1,1, s)	(1,1,1)	128	valid

表 5 分类器层

Table 5 Classifier layer

名称	神经元节点	激活函数
fc	128	ReLU
logits	C	Softmax

实验中,取每个样本周围大小为 3 邻域的像素块,训练 batch size 设置为 180,其中包括无标签样本 90,有标签样本 90,使用带动量的随机梯度下降法进行参数优化,动量大小设置为 0.9,优化器初始学习率为 0.001,对于学习率,本文采用指数衰减的方式进行更新,衰减率为 0.95,衰减步长为 5 000。对于记忆模块的更新系数 η ,根据具体的数据设定,实验数据集为 Indian Pines 时设置为 0.8,实验数据集为 Pavia University 和 Salinas 时设置为 0.5。

2.3 评价指标

为了客观比较不同方法的分类性能,在实验中采用了 3 种不同的量化评价指标,分别为总体分类精度(Overall Accuracy, OA)^[8]、类别平均分类精度(Average Accuracy, AA)^[8]和 Kappa 系数^[21],这 3 个指标的取值范围均为 $[0,1]$ 。对于每一个评价指标,得分越高表示分类结果越好。

2.4 实验结果

对于所有的数据集,随机采样 5 次,每次每类数据随机采样 10 个有标签样本,以及一共 $10 \times C$ 个无标签样本用于训练,其余用作测试。并将 5 次采样结果的平均值作为对应方法的最终性能评价结果。将不同方法在 3 个数据集上的分类数值结果汇总到表 6~表 8 中(表中加粗数据表示最好结果)。具体来说,在 Indian Pines 数据集上,当数据量较少时,本文方法的总体分类精度优于除 ISSR-HIC 方法以外的其他对比方法,对比 KNN、SVM、1D-CNN、3D-CNN、SS-CNN 方法,本文方法分别提升了 11.95%、3.14%、4.34%、2.68%、1.50%,对比 ISSR-HIC 方法降低 3.69%。在 Pavia University 数据集和 Salinas 数据集上,本文方法明显优于其他对比方法,对比 ISSR-HIC 方法,本文方法的总体分类精度分别提高 13.38%、3.59%。相较于 Pavia University 数据集和 Salinas 数据集,Indian Pines 数据集的空间分辨率更低,因此 Indian Pines 数据集中存在更多的噪声。本文通过记忆模块对无标签数据赋予伪标签,大量噪声的存在会导致记忆模块的预测不准确,无监督学习有偏移,影响整体的分类精度。因此,实验结果表明本文方法在大多数情况下能在训练样本有限时提升模型的分类性能。

表 6 不同方法在 Indian Pines 数据集上的分类结果

Table 6 Classification results of different methods on

Indian Pines datasets							%
序号	KNN	SVM	1D-CNN	3D-CNN	SS-CNN	ISSR-HIC	本文方法
1	86.11	82.50	79.44	79.44	76.09	90.76	63.89
2	44.63	41.70	31.71	29.28	18.00	55.25	10.65
3	36.05	47.74	47.63	44.15	6.86	52.56	11.59
4	48.02	57.31	49.16	42.29	70.04	64.98	46.26
5	69.98	76.77	74.97	77.59	17.18	76.09	64.38
6	64.39	81.19	89.44	87.72	80.41	85.82	92.36
7	92.22	92.22	83.33	98.89	92.86	92.86	91.67
8	66.03	82.46	76.03	77.82	78.87	78.09	97.44
9	68.00	98.00	94.00	100	100	95.00	90.00
10	39.33	49.95	49.29	54.51	55.33	55.14	17.05
11	29.87	39.75	38.45	46.74	68.73	47.82	89.65
12	26.83	42.04	38.46	36.50	31.20	48.06	23.50
13	90.15	93.23	92.00	96.62	93.17	95.61	92.56
14	61.96	66.22	70.17	70.80	96.05	74.35	84.22
15	14.31	36.91	46.65	38.24	38.95	41.97	28.99
16	86.02	91.57	88.43	98.07	100	93.55	96.99
AA	57.74	67.47	65.57	67.42	63.98	71.74	62.57
OA	44.99	53.80	52.60	54.26	55.44	60.63	56.94
Kappa	38.69	48.46	47.16	48.71	48.96	55.87	49.28

表 7 不同方法在 Pavia University 数据集上的分类结果

Table 7 Classification results of different methods on

Pavia University datasets							%
序号	KNN	SVM	1D-CNN	3D-CNN	SS-CNN	ISSR-HIC	本文方法
1	58.07	68.75	68.15	68.84	75.19	69.44	96.55
2	66.60	68.17	62.95	74.63	86.05	65.22	91.94
3	48.22	66.40	66.18	68.43	61.99	66.33	62.83
4	85.38	82.27	86.64	90.22	93.81	85.13	83.72
5	92.27	99.28	99.43	98.97	99.18	99.57	99.91
6	29.25	67.02	71.30	50.37	32.85	67.21	39.74
7	89.97	87.83	86.50	87.88	74.06	91.24	90.71
8	56.00	81.55	71.14	73.40	72.41	80.75	92.46
9	99.89	99.91	99.96	99.59	97.99	98.76	100
AA	70.07	80.13	79.14	79.15	77.06	80.41	84.21
OA	62.83	72.47	69.98	73.30	76.47	71.56	84.94
Kappa	53.02	65.14	70.32	65.79	68.95	64.38	79.65

表 8 不同方法在 Salinas 数据集上的分类结果

Table 8 Classification results of different methods on

Salinas datasets							%
序号	KNN	SVM	1D-CNN	3D-CNN	SS-CNN	ISSR-HIC	本文方法
1	98.27	94.92	95.55	91.23	99.29	97.55	99.28
2	81.15	97.47	99.77	97.34	98.09	97.82	99.60
3	74.38	87.60	92.33	89.70	69.38	86.29	98.19
4	94.45	99.29	98.71	95.55	97.99	98.05	99.47
5	88.54	92.83	90.41	93.64	99.18	95.44	94.61
6	96.58	98.34	99.21	96.28	99.09	99.25	99.26
7	98.55	98.83	99.00	96.02	99.80	99.46	99.46
8	62.39	53.00	68.02	58.28	25.30	58.08	79.12
9	91.36	96.03	95.60	92.80	99.00	96.02	99.03
10	58.92	83.71	83.48	76.08	81.96	82.82	74.96
11	78.11	86.60	93.72	87.75	86.48	95.13	96.96
12	91.46	99.74	99.45	99.46	99.90	98.62	99.96
13	98.72	98.10	95.74	95.81	99.56	96.97	99.58
14	88.17	90.91	94.11	92.87	94.53	95.35	94.83
15	52.84	70.70	54.26	69.76	93.09	67.23	56.91
16	43.17	90.04	79.67	90.16	77.30	92.17	98.25
AA	81.32	89.88	89.94	88.92	88.75	91.02	92.74
OA	76.20	82.61	83.57	82.31	79.73	83.72	87.31
Kappa	73.56	80.76	81.74	80.41	77.69	81.97	85.84

上述实验结果说明,本文方法能够有效减轻基于深度神经网络的分类方法在仅有少量训练样本的情况下产生的过拟合问题。为了进一步说明这一问题,在图 3 ~ 图 5 中绘制了不同方法在 3 个数据集上获得的分类结果视觉图。可以看出,相对于其他方法,在绝大多数情况下,本文方法更加接近真实的分类结果,从而再次验证了上述结论。

2.5 有效性验证

本节验证训练样本量不同及损失函数变化对本文方法性能的影响。

2.5.1 训练样本有效性

本节对比本文方法与 3D-CNN、ISSR-HIC 方法在训练样本量不同时的分类表现。实验中分别设置有标签训练样本量为每类 1、5、10、50、100 和 200,其余用作测试。3 个数据集上的数值结果如表 9 ~ 表 11 所示(表中加粗数据表示最好结果)。可以看出,随着有标签训练样本量不断增加,本文

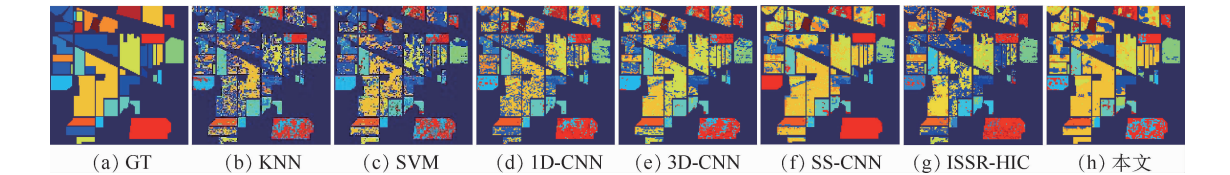


图 3 Indian Pines 数据集上的分类结果

Fig. 3 Classification results of different methods on Indian Pines datasets

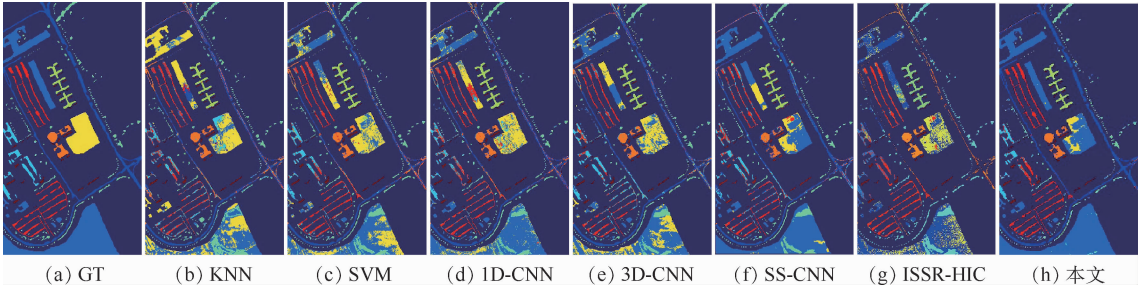


图 4 Pavia University 数据集上的分类结果

Fig. 4 Classification results of different methods on Pavia University datasets

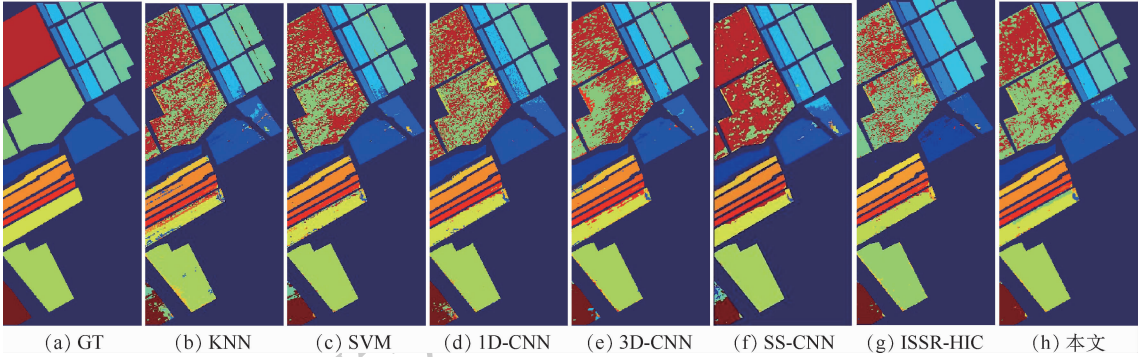


图 5 Salinas 数据集上的分类结果

Fig. 5 Classification results of different methods on Salinas datasets

表 9 Indian Pines 数据集上不同数量训练样本下的分类结果

Table 9 Classification results of different methods with changed numbers of training samples on Indian Pines datasets

方法	评价 指标	训练样本数						%
		1	5	10	50	100	200	
3D-CNN	AA	32.96	54.23	67.42	87.42	91.09	92.64	
	OA	22.70	43.42	54.26	77.01	85.54	87.71	
	Kappa	15.33	36.90	48.71	72.32	83.47	85.68	
ISSR-HIC	AA	40.93	66.35	71.74	84.31	88.73	91.01	
	OA	30.55	55.05	60.63	79.09	81.14	84.22	
	Kappa	23.31	49.64	55.87	75.79	78.80	82.22	
本文方法	AA	33.77	48.18	62.57	89.45	92.49	95.80	
	OA	39.60	49.94	56.94	83.92	86.27	92.46	
	Kappa	29.67	41.64	49.28	81.54	84.16	91.14	

方法及 3D-CNN、ISSR-HIC 方法的分类准确率均有所提高。当有标签样本充足时(每类有标签样本为 200),对比 3D-CNN 和 ISSR-HIC 方法,本文方法在 Indian Pines 数据集上的总体分类精度提升 4.75%、8.24%,在 Pavia University 数据集上提升 2.88%、8.79%,在 Salinas 数据集上提升 3.79%、1.68%。在有标签样本不足时(每类标签样本数为 5),对比 3D-CNN 和 ISSR-HIC 方法,本文方法在 Pavia University 数据集上的总体分类精度分别提升 10.08%、11.17%,在 Salinas 数据集上提升 8.52%、5.56%,在 Indian Pines 数据集上,本文方法比 3D-CNN 方法总体分类精度提升 6.52%,比 ISSR-HIC 方法下降 5.11%。

表 10 Pavia University 数据集上不同数量训练样本下的分类结果

Table 10 Classification results of different methods with changed numbers of training samples on Pavia University datasets

方法	评价 指标	样本数						%
		1	5	10	50	100	200	
3D-CNN	AA	55.03	68.27	79.15	92.70	95.74	96.04	
	OA	49.97	63.81	73.30	92.15	94.84	96.43	
	Kappa	37.48	53.36	65.79	89.65	93.18	95.22	
ISSR-HIC	AA	65.94	76.94	80.41	86.10	89.89	92.51	
	OA	53.46	62.72	71.56	81.55	86.84	90.52	
	Kappa	43.11	54.70	64.38	76.26	82.93	87.63	
本文方法	AA	55.80	72.21	84.21	95.83	97.87	98.80	
	OA	55.81	73.89	84.94	97.32	98.42	99.31	
	Kappa	41.88	65.14	79.65	96.41	97.89	99.07	

表 11 Salinas 数据集上不同数量训练样本下的分类效果

Table 11 Classification results of different methods with changed numbers of training samples on Salinas datasets

方法	评价 指标	样本数						%
		1	5	10	50	100	200	
3D-CNN	AA	62.02	85.24	88.92	94.08	94.76	95.82	
	OA	57.24	76.97	82.31	87.25	88.13	89.86	
	Kappa	53.04	74.61	80.41	85.83	86.79	88.68	
ISSR-HIC	AA	77.67	87.39	91.02	94.47	96.01	96.63	
	OA	70.65	79.93	83.72	90.21	91.52	91.97	
	Kappa	67.44	77.67	81.97	89.13	90.56	91.08	
本文方法	AA	78.30	90.99	92.74	94.56	96.76	97.27	
	OA	78.30	85.49	87.31	89.01	92.78	93.65	
	Kappa	74.39	83.83	85.84	87.73	91.94	92.89	

实验结果证明,在使用不同训练样本时,相比基准方法,本文方法在大多数情况下能有效提升分类任务的性能。

2.5.2 损失函数有效性

本节验证不同损失函数对分类结果的影响。将本文方法($\ell_{ce} + \ell_u$)与去除部分无监督损失($\ell_{ce} + \ell_{kl}$)及去除全部无监督损失(ℓ_{ce})的方法进行对比,在 3 个数据集上数值结果如表 12 所示(表中加粗数据表示最好结果)。可以看出,每部分无监督损失均能有效提升模型性能,如在 Pavia University 数据集上,去除信息熵损失,总体分类精度降低 3.3%,去除信息熵和相对熵损失,总体分类准确率下降 16.26%。该结果证明了无监督损失函数的有效性,记忆模块对无标签数据的预测可以有效减少噪声的引入,使无监督模型更加稳定。

表 12 不同损失函数在不同数据集上的分类结果

Table 12 Classification results of the proposed methods with different loss function on three datasets

数据集	评价指标	损失函数			/%
		$\ell_{ce} + \ell_u$	$\ell_{ce} + \ell_{kl}$	ℓ_{ce}	
Indian Pines	AA	62.57	55.58	61.20	
	OA	56.94	52.26	48.33	
	Kappa	49.28	45.28	42.28	
Pavia University	AA	84.21	83.08	77.42	
	OA	84.94	81.64	68.68	
	Kappa	79.65	75.38	60.88	
Salinas	AA	92.74	90.78	90.08	
	OA	87.31	85.49	83.29	
	Kappa	85.84	83.78	81.46	

3 结 论

1) 针对少量训练样本导致基于深度神经网络的分类方法产生过拟合的问题,本文提出一种基于记忆关联学习的小样本高光谱图像分类方法。

2) 首先,通过特征提取模块对特征进行提取,构建记忆模块学习和存储有标签样本的类别特征信息,同时利用更新后的记忆模块预测无标签数据的类别分布。然后,根据预测的类别分布在大量的无标签数据上构建无监督分类任务,同时在少量的有标签样本上构建有监督分类任务。最后,通过联合求解 2 个学习任务,优化分类模型参数,使得模型在有标签样本较少的情况下,避免了因监督信息不足而造成的过拟合问题。

3) 实验结果表明,本文方法能够有效地处理高光谱图像的小样本学习问题,提升分类准确率。

参考文献 (References)

[1] LANDGREBE D. Hyperspectral image data analysis[J]. IEEE Signal Processing Magazine,2002,19(1):17-28.

[2] BISHOP C A,LIU J G,MASON P J. Hyperspectral remote sensing for mineral exploration in Pulang, Yunnan Province, China [J]. International Journal of Remote Sensing,2011,32(9):2409-2426.

[3] ZHANG B,WU D,ZHANG L,et al. Application of hyperspectral remote sensing for environment monitoring in mining areas [J]. Environmental Earth Sciences,2012,65(3):649-658.

[4] ZHANG Y X,DU B,ZHAGN L P,et al. Joint sparse representation and multitask learning for hyperspectral target detection [J]. IEEE Transactions on Geoscience and Remote Sensing,2017,55(2):894-906.

[5] WANG C,ZHANG L,WEI W,et al. When low rank representation based hyperspectral imagery classification meets segmented stacked denoising auto-encoder based spatial-spectral feature [J]. Remote Sensing,2018,10(2):284.

[6] PAN E T,MEI X G,WANG Q D,et al. Spectral-spatial classification for hyperspectral image based on a single GRU[J]. Neurocomputing,2020,387:150-160.

[7] GUO G D,WANG H,BELL D A,et al. KNN model-based approach in classification[C] // OTM Confederated International Conferences “On the Move to Meaningful Internet Systems”. Berlin:Springer,2003:986-996.

[8] MELGANI F,BRUZZONE L. Classification of hyperspectral remote sensing images with support vector machines [J]. IEEE Transactions on Geoscience and Remote Sensing,2004,42(8):1778-1790.

[9] LI W,DU Q. Joint within-class collaborative representation for hyperspectral image classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2014,7(6):2200-2208.

[10] HU W,HUANG Y Y,LI W,et al. Deep convolutional neural networks for hyperspectral image classification[J]. Journal of Sensors,2015,2015:258619.

[11] LIU X F,SUN Q Q,LIU B,et al. Hyperspectral image classification based on convolutional neural network and dimension reduction[C] //2017 Chinese Automation Congress. Piscataway: IEEE Press,2017:1686-1690.

[12] LEE H,KWON H. Going deeper with contextual CNN for hyperspectral image classification[J]. IEEE Transactions on Image Processing,2017,26(10):4843-4855.

[13] ZHONG Z L,LI J,LUO Z M,et al. Chapman: Spectral-spatial residual network for hyperspectral image classification: A 3-D deep learning framework[J]. IEEE Transactions on Geoscience and Remote Sensing,2018,56(2):847-858.

[14] FANG B,LI Y,ZHAGN H K,et al. Hyperspectral images classification based on dense convolutional networks with spectral-wise attention mechanism[J]. Remote Sensing,2019,11(2):159.

[15] PAOLETTI M E,HAUT J M,BELTRAN R F,et al. Capsule networks for hyperspectral image classification[J]. IEEE Transac-

- tions on Geoscience and Remote Sensing, 2019, 57 (4): 2145-2160.
- [16] SAMIAPPAN S, MOORHEAD R J. Semi-supervised co-training and active learning framework for hyperspectral image classification[C]//IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Piscataway: IEEE Press, 2015: 401-404.
- [17] RASMUS A, VALPOLA H, HONKALA M, et al. Semi-supervised learning with ladder networks[EB/OL]. (2015-07-09) [2020-08-01]. <https://arxiv.org/abs/1507.02672>.
- [18] WEI W, ZHAGN L, LI Y, et al. Intraclass similarity structure representation-based hyperspectral imagery classification with few samples[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 1045-1054.
- [19] LI W, WU G D, ZHANG F, et al. Hyperspectral image classification using deep pixel-pair features[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55 (2): 844-853.
- [20] LIU B, YU X C, ZHAGN P Q, et al. A semi-supervised convolutional neural network for hyperspectral image classification[J]. Remote Sensing Letters, 2017, 8 (9): 839-848.
- [21] PONTIUS R G, MILLONES M. Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment[J]. International Journal of Remote Sensing, 2011, 32 (15): 4407-4429.

作者简介:

王聪 女, 博士研究生。主要研究方向: 高光谱图像分类。

张锦阳 男, 硕士。主要研究方向: 高光谱图像分类。

张磊 男, 博士, 教授。主要研究方向: 图像处理及机器学习。

魏巍 男, 博士, 副教授, 博士生导师。主要研究方向: 计算机视觉、图像处理。

张艳宁 女, 博士, 教授, 博士生导师。主要研究方向: 图像处理与计算机视觉、模式识别与人工智能、机器学习等。

Small sample hyperspectral image classification method based on memory association learning

WANG Cong^{1,2}, ZHAGN Jinyang², ZHANG Lei², WEI Wei^{1,2,*}, ZHANG Yanning²

(1. Research & Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China;

2. School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, China)

Abstract: Hyperspectral Image (HSI) classification is one of the fundamental applications in remote sensing domain. Due to the expensive cost of manual labeling in HSIs, in real applications, only small labeled samples can be obtained. However, limited samples cannot accurately describe the data distribution and often cause the training of classifiers to be overfitting. To address this problem, we present a small sample hyperspectral image classification method based on memory association learning. First, considering that the unlabeled samples also contain a lot of information related to the data distribution, we construct a memory module based on the labeled samples. Then, according to the feature association among labeled and unlabeled samples, we learn the label distribution of the unlabeled sample with the continuously updated memory module. Finally, we build an unsupervised classifier model and a supervised classifier model, and jointly learn these two models. Extensive experimental results on multiple hyperspectral image classification datasets demonstrate that the proposed method can effectively improve the accuracy of small sample HSI classification.

Keywords: memory association learning; semi-supervised; small sample; Hyperspectral Image (HSI); classification

Received: 2020-09-07; Accepted: 2020-09-11; Published online: 2020-11-04 17:23

URL: kns.cnki.net/kcms/detail/11.2625.V.20201103.1825.002.html

Foundation items: Foundation of Science, Technology and Innovation Commission of Shenzhen Municipality (JCYJ20190806160210899); National Natural Science Foundation of China (61671385, U19B2037)

* Corresponding author. E-mail: weiweinwu@nwpu.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0463

低延迟视频编码技术



宋利^{1,2,*}, 刘孝勇¹, 武国庆¹, 朱辰¹, 黄琰¹, 解蓉¹, 张文军^{1,2}

(1. 上海交通大学 图像通信与网络工程研究所, 上海 200240;
2. 上海交通大学 未来媒体网络协同创新中心, 上海 200240)

摘 要: 随着视频编码和视频传输技术的广泛应用, 视频需求量剧增, 实时视频通信成为视频行业的一项重要研究内容, 核心目标是提供更好的用户体验和更低的延迟。低延迟视频编码是实时视频通信应用的关键部分, 通过降低编码延迟可以有效地降低系统的整体延迟。首先, 分析了视频传输系统的延迟来源, 从通用的视频编码框架出发着重介绍了编码延迟的产生机制; 其次, 概述了国内外主流的视频编码标准, 介绍了率失真优化技术的原理和模型, 为低延迟视频编码器的设计提供了理论基础; 最后, 从参考结构、流水线设计、编码模式搜索、码率控制和硬件加速多个维度描述了优化编码延迟的技术手段, 并总结了业界具有代表性的低延迟视频编码方案, 简要说明了现有低延迟视频编码技术的局限性, 并对未来的发展方向做了展望。

关 键 词: 低延迟; 视频编码; 实时视频通信; 率失真优化; 编码用例

中图分类号: TN919.8

文献标志码: A **文章编号:** 1001-5965(2021)03-0558-14

娱乐、社交媒体、医学图像、用户拍摄的图像/视频等多媒体数据量剧增, 而视频数据在其中占主导地位, 同时人们对高分辨率、高帧率、高动态范围视频的需求也日益增加, 互联网智能解决方案提供商 Sandvine 最新报告显示, 2019 年视频流量占全球互联网下行总流量的 60% 以上^[1]。

此外, 视频通话、视频会议、远程教育、互动直播等实时通信场景不断涌现, 促进了实时通信技术的发展, 同时随着智能终端设备的广泛使用以及 5G 网络的普及, 用户对实时交互的需求也越来越高。

数字化后的视频数据存在大量的冗余, 在统计特征层面包括空间冗余、时间冗余、信息熵冗余; 在人类认知层面包括视觉冗余、知识冗余和结构冗余^[2]; 视频分辨率、帧率、动态范围的增加也导致视频数据量不断增长, 通过预测编码和变换

编码可以显著减少这些冗余, 从而减少视频的数据量, 为了规范各种视频编码方法生成的码流格式, 国际国内制定了相应的视频编码标准。

实时视频通信的关键环节在于实时视频编解码和实时视频传输, 视频编码、传输、解码等各模块需要协调配合, 确保能够在给定的时间段内完成。为了减小系统延迟, 编解码环节需要合理地权衡编码效率和延迟的关系, 同时考虑到压缩码流中会因视频内容的差异而出现突发式数据流。进行网络传输时容易引起拥塞, 因此在编码端也需要设计合理的码率控制算法以保证码率的平稳, 从而实现在受限带宽网络中的实时流传输, 并平滑视频的整体质量。

视频传输系统的延迟主要包括 3 部分: 编码器延迟、传输延迟和解码器延迟^[3]。编码器延迟由视频采集延迟、帧重新排序延迟和编码器处理

收稿日期: 2020-08-26; 录用日期: 2020-09-19; 网络出版时间: 2020-11-05 09:21
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201104.0859.001.html
基金项目: 国家重点研发计划(2019YFB1802701); 国家自然科学基金(61671296)
* 通信作者. E-mail: song_li@sjtu.edu.cn

延迟组成;传输延迟由编码缓冲区延迟、网络传输延迟和解码器缓冲延迟组成;解码器延迟由解码器处理延迟和帧重新排序延迟组成。

传输延迟主要由网络抖动和缓冲区大小决定,解码器延迟由客户端设计决定,本文主要从编码角度阐述低延迟视频编码器设计思路,以适配端到端的低延迟应用场景。具体包括减少编码处理的延迟,避免出现帧重新排序延迟以及降低编解码缓冲区的延迟。互联网行业的迅速发展催生了各式各样的低延迟应用,而不同的应用场景所要求达到的编码延迟不同,体育赛事、互动直播等直播类场景的目标编码延迟在秒级;无线投屏、视频会议等场景要求的编码延迟在百毫秒级别;对于云VR和第一人称云游戏(射击类和MOBA类云游戏)等应用,用户体验极易受到系统延迟的影响,编码延迟一般维持在十毫秒级;另外在无人驾驶和无人机等超低延迟场景下,编码延迟基本控制在毫秒级以实现对外部信息的及时反馈。

在视频编码领域,目前暂未查阅到总结低延迟视频编码技术的综述类论文,本文从编码延迟的来源出发,系统地归纳了学术界和工业界在低延迟视频编码方面的优化手段,对低延迟场景下编码方案的选择和编码器的设计具有重要的指导意义。

1 视频编码标准

视频编码技术解决了视频数字化产生的海量数据的存储和传输问题,但是不同的编码方法得到的码流格式可能存在差异,为了让编码码流可以在不同的平台上解码播放,就需要制定统一的视频编码标准。

1.1 标准简介

当前国际上主流的视频编码标准主要包含H. 264/AVC、H. 265/HEVC和H. 266/VVC,由ITU-T的VCEG(Video Coding Experts Group)和ISO/IEC的MPEG(Moving Picture Experts Group)联合制定,比较流行的标准还有Google发布的VP9和AV1。在国内,AVS工作组发布了AVS标准,现在已经走到了AVS3。

上述编码标准以不断提升编码性能为导向,追求更高的压缩效率和更好的观看质量,往往较少地考虑编码延迟对系统的影响,虽然在实际应用中已经衍化出了x264、x265、rav1e、xAVS2等实时编码器,但是仍然无法满足低延迟场景的需求。

1.2 编码工具

混合编码框架主要包括块划分、帧内预测、帧

间预测、变换、环路滤波等核心模块。

将图像划分为块可以更好地适应视频图像的纹理特征,随着视频编码标准地不断演进,编码块的划分结构越来越灵活,编码树单元(Coding Tree Unit,CTU)的尺寸越来越大,满足了复杂多变的视频场景的需要。

帧内预测利用图像中空间相邻像素的相关性来减少空间冗余,主要采用角度预测的方式,使用当前编码块邻近的重建像素预测当前块,通过增加帧内预测模式的数量和帧内插值滤波器的抽头数提升预测精度。

帧间预测利用视频时间域的相关性去除数据的时间冗余,在已编码的重建图像中寻找当前块的最佳匹配块以实现基于块的运动补偿,根据视频中的运动类型构建合适的运动模型以及使用更高精度的运动矢量可以有效提升预测精度。

变换编码可以将像素的能量集中到频域中少数几个低频系数上,并将图像的高频信息转化为较小的高频系数值,从而更紧凑地表示图像信息,对变换后的低频系数使用小的量化步长,对高频系数采用大的量化步长可以有效降低视频信号的空间相关性。

对视频图像进行块划分后,编码块的边界处会产生失真和不平滑,同时基于块的运动预测、变换量化等处理也会导致边缘效应,对重建图像进行环路滤波可以平滑边界处的视频质量,被用作参考图像后也可以提高编码效率。

2 率失真优化技术

率失真理论要解决的基本问题是在给定码率下,如何让失真最小,香农的“保真度准则下的信源编码定理”证明了率失真函数的存在性。在视频编码中,率失真优化(Rate-Distortion Optimization,RDO)的任务是为每一块图像区域选择最高效的编码参数(划分结构、预测模式、运动矢量、量化等级等),由于不同的编码选项在不同的码率(失真等级)、不同的场景内容下具有不同的压缩效率,所以率失真优化的过程比较复杂,在低延迟视频编码中,应该增加编码延迟作为一个新的约束条件。

2.1 基本原理

率失真优化的目的是在一定的码率约束下最小化失真,由于图像不同区域所分配的编码模式具有不同的率失真特性,所以编码器的目标是优化整体的保真度,这个约束优化问题可以表示为

$$\min D \quad \text{s.t.} \quad R < R_c \quad (1)$$

式中: D 为编码视频的失真; R 为所消耗的比特数; R_c 为最大允许的比特数。式(1)中的有约束优化问题可以通过拉格朗日乘数法转化为无约束优化问题^[4], 最小化问题的拉格朗日公式为

$$\min J = D + \lambda R \quad (2)$$

式中: J 为率失真代价; λ 为拉格朗日乘子, 可以用于调整优化目标。在给定 λ 下, 式(2)的每一个解对应式(1)的一个最优解, 所以率失真优化技术可以从优化的角度对大量可能的编码选项进行评价。

在实际进行视频编码优化时, 编码决策之间的一些相互影响是被忽略的, 而在进行运动估计和预测模式决策时, 当前编码区域和空间以及时间上的后续区域存在级联效应, 所以会带来一定的性能损失。假设不同编码决策之间的失真和码率互不相关时, 式(2)可以分解为多个子任务的率失真函数:

$$\min J = \sum_{i=1}^N D_i + \lambda \sum_{i=1}^N R_i \quad (3)$$

$$\min J_i = D_i + \lambda_i R_i \quad (4)$$

式中: N 为编码决策的数量; J_i 、 D_i 、 R_i 和 λ_i 分别为第 i 个编码决策的率失真代价、编码失真、消耗比特数和拉格朗日乘子。

考虑到编码单元内部和编码单元之间的编码参数广泛存在依赖关系, 许多研究对率失真优化中的空域和时域依赖关系进行了分析。在空域依赖方面, Bichon 等^[5]在进行帧内预测时对垂直方向上的 2 个编码块进行了联合率失真优化 (Joint Rate-Distortion Optimization, JRDO), 从编码参数集中选择全局率失真代价最小的编码参数; 后来 Bichon 等^[6]对上述文献中的方法进行了改进, 根据图像内容自适应地进行 JRDO 和 RDO, 从而显著降低了计算复杂度; Pang^[7]和 Wu^[8]等建立了边界像素失真和编码块失真的空间预测依赖关系, 通过空域失真传播模型进行 RDO。在时域依赖方面, 后续编码单元的率失真性能受到被参考编码单元的编码决策的影响, Yang 等^[9]提出了一种信源失真时域传播模型, 通过估计当前编码单元对时域传播链上的后续编码单元的影响获得了对应的拉格朗日乘子; Li 等^[10]分析了当前编码单元的性能对后续帧的影响, 通过调整拉格朗日乘子实现了时域依赖 RDO 方案; González-de-Suso 等^[11]通过对视频中静态背景和非静态背景的预分析自适应地估计拉格朗日乘子, 此外, 有的研究根据实验统计^[12-13]和理论分析^[14]的方法建立时域依赖的率失真模型。上述方法在不显著增加计算复

杂度的条件下可以取得 4% ~ 15% 的性能增益。

2.2 延迟约束的率失真优化

在跨层无线视频通信系统中, 为了实现最优的端到端服务质量, 需要根据端到端延迟约束和给定的最大电源功率为视频编码器合理地分配编码时间和编码功率。对视频编码器的延迟和功率消耗的约束可以从 2 个方面进行说明: 一方面, 高效的视频压缩可以降低视频数据的码率, 从而减少在物理层的传输功率和传输延迟, 或者减少数据链路层的传输速率和传输误码率; 另一方面, 高效的视频压缩具有更高的计算复杂度, 从而在应用层产生大的编码器功率消耗和编码延迟, 这 2 个矛盾的方面表明, 在设计视频编码器时需要权衡延迟、功率消耗、码率和视频失真的关系。为了找到最优的权衡方案, Li 等^[15]提出了一种理论分析框架用于建模视频编码器的延迟-功率-码率-失真 (delay-Power-Rate-Distortion, dPRD) 特性, 推导出了码率 R 、失真 D 、延迟 d 和功率 P 关于运动搜索范围和量化参数的分析模型, 并把信源码率控制问题表示为 dPRD 优化问题。

在上述文献的基础上, Li 等^[16]提出了基于延迟-码率-失真优化 (delay-Rate-Distortion Optimization, dRDO) 模型的码率控制方案, 使用信源编码延迟和缓冲区延迟、信源码率和信道编码产生的冗余码率这 2 种权衡最小化端到端总失真的平均值, 即相对于无线视频通信系统中已有的码率控制方案增加了端到端延迟的约束条件。

随着编码工具的增多, 编码器的压缩效率显著提升, 同时计算复杂度也大大增加, Huang 等^[17]在率失真性能损失和复杂度之间进行了权衡, 通过建立编码模式的子集得到了码率-失真-复杂度优化 (Rate-Distortion-Complexity Optimization, RDCO) 模型, RDCO 不仅适用于不同编码模式决策的算法设计, 而且可以在不同的码率-失真和复杂度的权衡要求下灵活地调整编码器的计算复杂度。

3 视频编码延迟优化技术

用户对 4K/UHD 等高分辨率视频的观看需求推动着视频编码标准不断更新换代, 与以往单纯地追求高压缩性能不同, 编码延迟逐渐成为设计编码器时需要考虑的重要因素之一, 虽然率失真优化技术可以选择最优的编码参数, 但是求解过程复杂, 不适合用在实际编码器中, 因此还需要考虑其他的技术手段。编码延迟具体是指从视频帧进入编码器到编码完成生成码流所消耗的时间,

编码过程中的参考关系、编码流水线的设计和编码模块的复杂度是导致编码延迟的 3 个主要因素。

在设计低延迟编码技术之前,首先需要分析编码的哪些部分会影响视频传输系统的端到端延迟。在视频采集端,视频输入编码器的顺序为时序,但是时序与编码/解码顺序并不一致,一些帧需要等待,带来了编码前延迟。在编码器内部,编码任务调度差,编码各模块计算量大导致了编码本身的延迟。对于传输过程,在固定带宽的网络中传输码率波动的视频流会导致拥塞,导致后续帧的延迟和缓冲延迟。

3.1 低延迟参考结构

在离线编码中,一般会使用双向预测(前向预测和后向预测)B 帧,B 帧一方面能够使当前帧被很好地预测,另一方面采用了较高的量化参数,因此压缩率较高,但是双向预测 B 帧由于需要在编码前对帧进行重排序,会带来不可避免的延迟。Carballeira 等^[18]描述了参考关系对编码延迟的影响,可用式(5)~式(9)进行说明。

$$L = \max(t_{\text{cod}_i} - t_{\text{capt}_i}) \quad i = 0, 1, \dots, M - 1 \quad (5)$$

$$t_{\text{cod}_i} = t_{\text{start}_i} + \Delta t_{\text{proc}_i} \quad (6)$$

$$t_r = \max(0, t_{\text{capt}_r} + \Delta t_{\text{proc}_r} - t_{\text{capt}_{r+1}}) \quad r \in R(i) \quad (7)$$

$$t_{\text{start}_i} = t_{\text{capt}_i} + \sum_{r \in R(i)} t_r \quad (8)$$

可以进一步推导出编码延迟的形式为

$$L = \max \left(\sum_{r \in R(i)} t_r + \Delta t_{\text{proc}_i} \right) \quad (9)$$

式中: L 为编码延迟; t_{cod_i} 为第 i 帧编码完成的时刻; t_{capt_i} 为第 i 帧的采集时间; M 为帧数; t_{start_i} 为第 i 帧开始编码的时刻; Δt_{proc_i} 为第 i 帧的处理时间; $t_r \geq 0$ 为参考路径中前一帧带给后一帧的延迟; $R(i)$ 为第 i 帧的参考路径所包含的参考图像集; t_{capt_r} 、 Δt_{proc_r} 和 $t_{\text{capt}_{r+1}}$ 分别为 $R(i)$ 中第 r 帧的采集时间、第 r 帧的处理时间和第 $r + 1$ 帧的采集时间。可以看出,编码延迟受预测结构和每一帧的处理时间的影响。

图 1 描述了 HEVC 的 RA(Random Access)模式的参考关系,图像编号(Picture of Count, POC)为 1 的帧会在 I 帧之后首先到来,但是由于参考关系,该帧需要在 POC 为 2 和 4 的帧编码完成之后才能进行编码,而在实时视频源中,POC 位置与采集时间相对应,因此,在不考虑编码计算延迟的前提下,该参考架构会直接带来额外的 3 帧延迟。在 50 帧/s 的场景中,3 帧延迟对应 60 ms,对于延迟不敏感的场景,例如离线编码,B 帧可以保

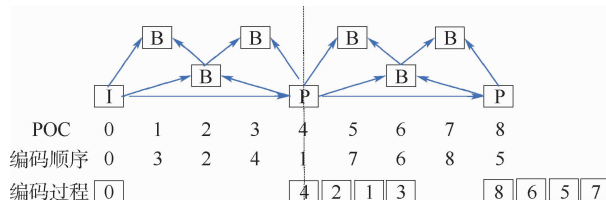


图 1 HEVC 的 RA 模式的参考关系

Fig. 1 Reference structure of RA mode for HEVC

留,以追求最大化压缩性能。

为了适配低延迟场景,HEVC 也设计了 LDP (Low-delay P) 模式,该模式的帧间参考关系如图 2 所示,Wenger^[19]研究了帧结构对编码性能的影响,文献基于 H. 263 + 比较了不同结构,包括含有 B 帧的结构和纯 P 帧结构的编码性能,实验证明,只使用 P 帧虽然节省了编码延迟,但是编码性能损失从 9% 到 42% 不等。事实上,鉴于不同应用场景的延迟要求,各家标准/私有编码器都设计了不同的参考帧结构,在延迟允许的条件下最大化编码性能。

多参考帧结构会增加运动估计的时间,同时不同帧之间的依赖关系会对编码延迟产生影响,Pan 等^[20]根据父预测单元和子预测单元的内容相似性获取子预测单元的参考帧选择信息,在 LD 和 RA 模式下分别减少了 54.29% 和 43.46% 的编码时间,性能损失分别为 1.07% 和 0.90%;Park 等^[21]利用了 QTBT (Quadtree plus Binary Tree) 结构的相关性,基于父节点的预测信息最小化运动估计中每一个编码单元的参考帧搜索范围,以 0.29% 的性能损失为代价减少了 13% 的编码时间。

通过上述帧级别的延迟优化,理论上输入的每一帧都可以立即编码,不需要等待,本质上不会引入新的延迟,但是仍存在后续条(slice)需要等待前面的条编码完成后才能启动编码,这种情况下编码延迟为前面的条的编码时间。对于一些超低延迟场景,为了降低这种延迟,出现了条级甚至宏块(macroblock)级的编码方案,在本条/宏块编码完成后立即将码流传输、解码并显示,可将编码延迟从 1 帧降为条/宏块,如图 3 所示^[22],同时也可以进一步通过并行化计算提升处理速度。

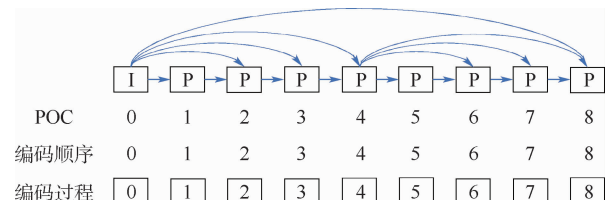


图 2 HEVC 的 LDP 模式的参考关系

Fig. 2 Reference structure of LDP mode for HEVC

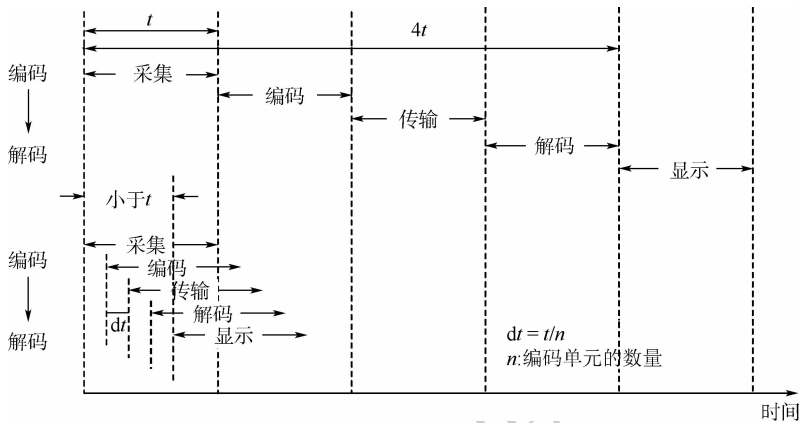


图 3 帧级编码和条/宏块级编码的端到端延迟^[22]

Fig. 3 End-to-end latency of frame-level encoding and slice/macroblock-level encoding^[22]

3.2 编码流水线优化

对于单线程离线编码,编码效率完全由编码标准决定,但是对于多线程/实时编码器,编码框架也有重大的影响。编码器内部的架构设计,一方面决定了编码器的延迟,另一方面决定了编码器的可并行程度。编码器的整体架构可以分为 3 种,如图 4^[23]所示。图 4(a)中,编码器不同编码模块,包括运动估计、运动补偿、变换(Discrete Cosine Transform, DCT)、量化(Quantization, Q)、编码、反变换和反量化,都是以帧为单位进行同步。图 4(b)中,则以宏块为单位同步。类似地,图 4(c)中,以条为单位进行同步,一个条包含多个宏块。每次同步之后,编码器就能把已

编码的数据输出,这也意味着,对于当前帧架构图 4(a)的延迟为帧级,图 4(b)为块级,图 4(c)为条级。

虽然理论上按照上述思路可以将延迟降为块级别,但是实际应用中,当一帧被切的越小,每一块编码时间差异越大,Schreier 和 Rothermel^[3]描述了不同宏块的计算量相差 2~3 倍,会导致一些同步问题和等待的额外开销,所以很多低延迟方案仍然采用条带级方案。

随着硬件性能的提升,CPU 具备了多线程并行处理的能力,这为编解码器的并行化操作提供了基础,在实际应用中主要分为任务级分解和数据级分解 2 部分^[24]。

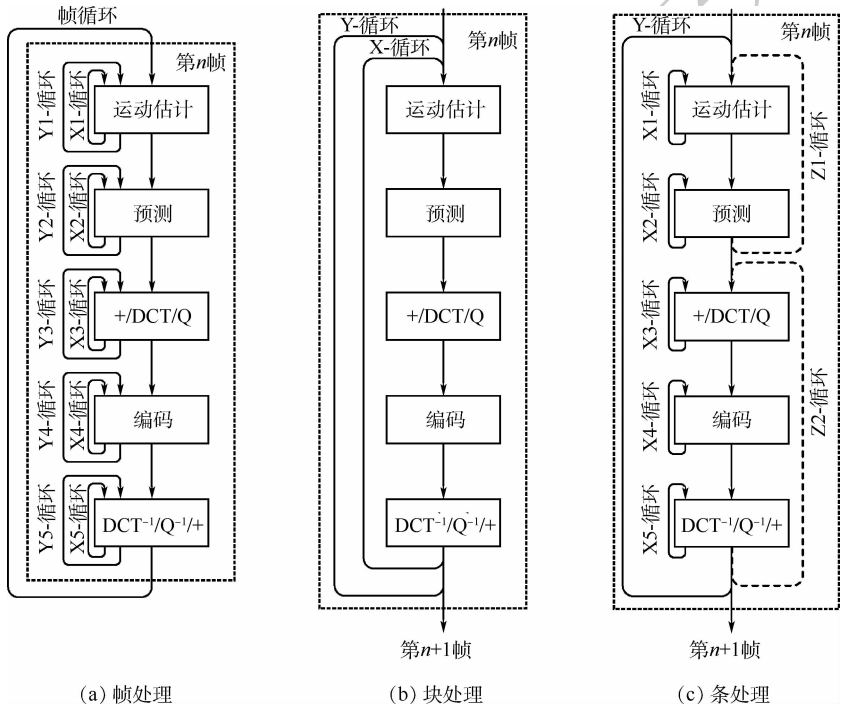


图 4 帧、条、块级编码任务分解^[23]

Fig. 4 Encoding processes at frame, slice and block levels^[23]

在任务级分解中,运动补偿、变换和量化等任务被分配给不同的处理器,处理器之间需要交流和同步,在帧、条、块级编码方案中(基于图4),可以通过不同编码任务的流化处理降低编码延迟。在数据级分解中,数据被切割并分配给不同的处理器,主要分为帧级并行、条级并行、片(tile)级并行、波前并行处理(Wavefront Parallel Processing, WPP)等。片级并行和WPP可以用于低延迟操作中,把图像细分为片或者WPP子图像,如图5所示。图5(a)是片的编码过程,叉号标记的3个片属于同一个编码单元,图5(b)包含了6个WPP子图像,每一个子图像为一个编码单元,则延迟量级由编码单元的数量决定^[22]。

并行方法主要通过横向扩展的方式增加编码速度,通过并行运算间接实现延迟的降低,不会降低因为参考关系导致的延迟^[25]。Mohamed等^[26]对HEVC帧级并行进行了性能评估,采用8个线程并行实现HEVC编码器(HEVC test Model, HM),每个线程被分配给一个核心处理器来编码一个单独的帧,使用All Intra配置打破连续帧之间的编码依赖,结果显示在没有编码性能损失的情况下,编码速度提升到4.35倍。

近几年WPP受到越来越多的关注,Chi等^[25]提出了一种基于WPP的重叠波前方法,实现了比片级并行和WPP更高的性能和效率。然而在计算资源充足的情况下,仍可以改进波前编码顺序设计,以进一步提高编码速度。Chen等^[27]提出集数据级和任务级方法于一体的波前高并行度算法。在任务级别,通过有效降低波前并行处理的依赖性,开发了一种新颖的帧间波前方法,利用三级线程管理方案,以最好地利用帧间波前方法的并行性实现相应的编码速度。Wang等^[28]提出了

一种基于协同调度的HEVC编码并行解决方案,包括自适应并行模式决策、异步帧级像素插值和多粒度任务调度过程,以克服WPP的缺点,进一步提高在多核平台上HEVC编码的并行性。然而基于协同调度的并行解决方案专注于帧内编码的并行化,将此方案与基于GOP或基于帧的多线程并行方案集成在一起可以进一步加快编码速度。

3.3 编码模块简化

使用并行的手段对编码框架进行优化,本质上是进行编码的调度优化,并没有减少编码模块本身的搜索,事实上,编码过程中,为了确定最终的编码模式,进行了大量冗余的搜索。同时为了适应视频内容的不同局部特征,更加高效地表示不同场景的纹理细节,图像需要被划分为一系列的编码块,AVC中使用了固定大小的宏块,HEVC中采用了四叉树的划分方式,VVC中使用了嵌套多类型树的四叉树划分,通过不断增加划分类型细化了对图像细节的处理,相应地计算复杂度也显著提升。HEVC与AVC相比,码率大约节省50%^[29],复杂度成倍增加^[30];VVC相对于HEVC约有30%的增益,编码复杂度约增加为7倍^[31],因此可以通过权衡编码效率与计算复杂度的关系以降低编码延迟。

如1.2节所述,当前的视频编码标准普遍采用了基于块的混合编码框架,包括预测编码和变换编码2部分^[32],计算复杂度主要来自帧内预测、帧间预测(运动补偿)、变换、量化、去块效应滤波、熵编码等模块。HEVC参考软件的复杂度测试结果显示,在采用未经优化的RA模式编码的情况下,各主要模块编码时间约为:参考块搜索占7%,量化占10%,运动估计占39%,运动补偿中的插值滤波占20%,熵编码占1.5%^[30],可见编码时间主要集中于帧间预测过程。

在复杂度优化层面,现有研究工作主要包括传统的基于统计信息的快速算法^[33-36]以及最新的利用机器学习、数据挖掘等新理论降低计算复杂度的方法^[37-40],传统方法易于实现,但是无法准确建模复杂的图像划分过程,基于学习的方法可以利用大规模数据充分发掘图像划分的广泛特征,提高预测精度,缺点是训练参数增加了额外的开销。通过以上方法将编码模式降到最简后,计算瓶颈主要来自熵编码和数据访存。基于参考软件,现有的低复杂度方法普遍同时考虑在性能损失合理的情况下尽可能降低编码复杂度,结合不同模块的加速算法,在压缩性能损失5%及以下的前提下,降低20%~70%的编码复杂度。

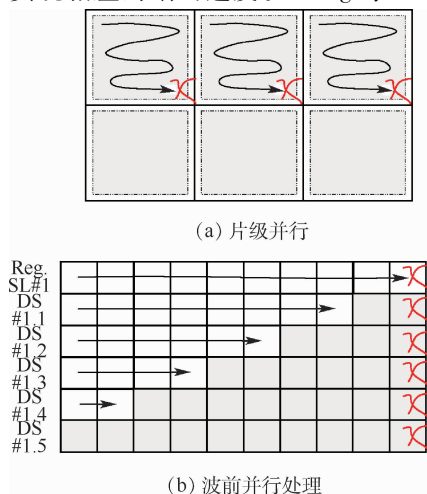


图5 片级并行和波前并行处理^[22]

Fig.5 Tile parallelization and wavefront parallel processing^[22]

3.4 低延迟码率控制

码率控制的目标是在满足码率约束的同时提高视频质量,通过选择合适的编码参数完成比特分配过程,因为传统方法要求编码的视频整体性能最优,而低延迟场景需要保证分配比特数的平稳,因此需要改进传统的码率控制算法以适应场景需要。

一方面,对于基本的视频编码方式,I帧通常以周期性整帧的方式插入,然而I帧一般比P、B帧多出数倍的数据量。这种编码结构对编码本身而言更为简洁,但会导致编解码缓冲区的较大延迟,这种延迟同样会影响许多实际应用中的传输质量。因而考虑到系统整体的低延迟,更好的做法是进行帧内刷新编码。

帧内刷新技术是指将原有I帧的码流分配到其他非I帧中,对于编码而言,即以特定方式将非I帧中的无重叠区域以帧内预测方式编码,共同分担I帧参考的任务。帧内刷新通过多帧帧内编码替代单个关键帧,可以起到降低缓冲区延迟、平稳码率的作用,避免I帧峰值码率带来的带宽压力。

在各类帧内刷新方法中,周期性帧内刷新(PIR)^[41]由于其简洁性、有效性被广泛采用。PIR是在限定的帧周期内,对每一帧固定位置的编码块进行帧内编码。除此之外,还有随机帧内刷新(RIR)^[42]、自适应帧内刷新(AIR)^[43-44]等方法。

Hendry等^[45]提出了一种新的图像类型GRA(Gradual Random Access)/GDR(Gradual Decoding Refresh)图像,图像中包含刷新区域和未刷新区域,在刷新条内的编码块只能参考属于同一GRA周期的参考图像中刷新区域的参考块,在未刷新

区域内的编码块可以参考属于不同GRA周期的参考图像中刷新区域的参考块或者属于同一GRA周期的参考图像中的任何参考块。图6给出了对刷新区域和未刷新区域的编码块的参考约束,GRA图像既可以平稳码率又有利于错误图像的解码恢复^[45]。

另一方面,超低延迟场景下编解码端的缓冲区很小,而采用条/块级编码后,子图像(宏块集)的比特数目会出现分配不均的情况,导致码率波动较大,因此在实际应用中需要避免出现缓冲区上溢的现象^[46]。子图像级的比特分配策略如下:

$$B_{\text{SubPic}} = \frac{B_{\text{Pic}} - B_{\text{Coded}}}{\sum_{\text{NotCoded}} \omega_{\text{SubPic}}} \omega_{\text{SubPicCur}} \quad (10)$$

$$B_{\text{SubPic}} < f_{\text{Size}} - f_{\text{Used}} \quad (11)$$

式中: B_{SubPic} 为每个子图像分配的比特数; B_{Pic} 为每帧图像分配的比特数; B_{Coded} 为已编码的子图像分配的比特数; ω_{SubPic} 为每个子图像的权重,可由前面已经编码的子图像的内容特性确定; $\omega_{\text{SubPicCur}}$ 为当前子图像的权重; f_{Size} 为缓冲区的大小; f_{Used} 为缓冲区的充盈度。

近年来对码率控制的研究主要以 Q 域^[47-48]、 ρ 域^[49]和 λ 域^[50]的码率控制算法为基础展开,码率控制过程包含GOP级、帧级、基本编码单元级的比特分配和编码参数选择2部分,现有码率控制算法对平均码率的控制精度较高,但是对基本单元级的比特控制不够精准,无法适用于缓存区较小的情况,而且不能有效处理包含快速运动物体、存在显著物体遮挡和场景变换的动态视频,所以需要改善现有的码率控制模型,实现不同编码

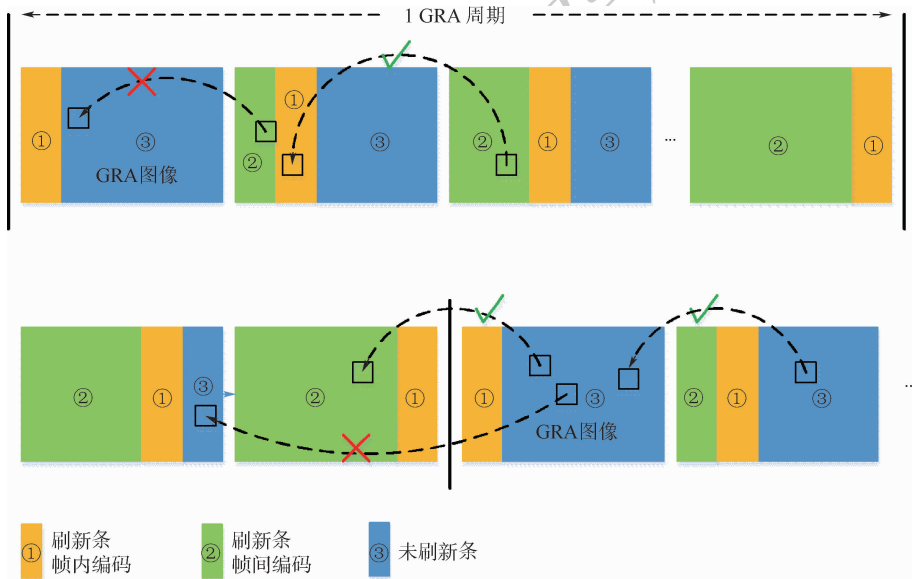


图6 刷新和未刷新区域的编码块参考^[45]

Fig.6 Encoding block reference structure for refreshed and unrefreshed regions^[45]

级别的比特的精确控制。最新的码率控制研究中使用了机器学习^[51]和深度强化学习^[52]理论,可以显著提升码率控制的精度,但是训练模型参数耗时过多,这 2 种算法以及 HM16.19 的原始算法可以实现的帧级(F)和序列级(Seq)码率控制精度的对比情况如表 1 所示,使用式(12)中的失配误差进行衡量^[52]:

$$E = \frac{|R_{tar} - R_{act}|}{R_{tar}} \times 100\%$$

(12)

式中: R_{act} 和 R_{tar} 分别为实际消耗的比特数和目标比特数,误差越小代表算法的性能越好。从实验结果可以看出,基于学习的方法在精确控制码率方面具有更大的潜力。

表 1 不同方法的码率控制精度对比^[52]

Table 1 Rate control accuracy comparison of different methods^[52]

序列	HM16.19		文献[51]		文献[52]	
	F/%	Seq/%	F/%	Seq/%	F/%	Seq/%
Mobisode	2.55	0.010	1.98	0.010	1.01	0.003
RaceHorses	1.53	0.008	1.43	0.007	0.66	0.004
PartyScene	2.33	0.009	2.11	0.008	0.29	0.003
Chromakey	2.55	0.010	2.32	0.009	0.33	0.003
Wave	2.22	0.009	2.01	0.008	0.28	0.002
Parkrun	2.73	0.011	2.38	0.009	0.41	0.003
Kimono	1.22	0.007	1.34	0.008	0.67	0.004
Tennis	2.76	0.011	2.41	0.009	0.44	0.004
Average	2.24	0.009	2.00	0.009	0.51	0.003

3.5 硬件编码加速

不同于 CPU 计算平台,硬件计算平台具有更强的并行数据处理能力,主要包括现场可编程门阵列(Field Programmable Gate Array,FPGA)和图形处理器(Graphical Processing Unit,GPU)等。

FPGA 芯片内部的逻辑门单元达到了千万量级,在并行处理和可扩展性方面具有明显的优势,可同时支持硬件和软件编码算法实现,已有针对帧内/帧间预测、变换、量化和熵编码等模块的 FPGA 实现,相关技术已经比较成熟。

GPU 采用了 SIMT(Single Instruction Multiple Threads)模型,通过对数千个线程的有效调度提升了指令的执行效率,广泛用于多种视频编码工具的加速中,适合插值、计算 SAD(Sum of Difference)等编码过程中分布集中的运算,基于 GPU 的视频编码加速是近年来的研究热点。

在软件算法的基础上,可以利用异构处理平台的协同计算对计算复杂度进行优化,利用 GPU 完成可并行化的编码模块,如视频预分析、码率分配、帧内预测模式选择、帧间预测运动估计、样点

自适应补偿等,利用 CPU 完成率失真优化决策、环路滤波及熵编码等难以并行化的过程。

随着 GPU 在通用计算领域的广泛应用,已有许多针对以往视频编码并行优化的研究工作。这些工作大部分集中在并行化特定的编码模块,这些模块可以被 SIMT 模型加速,比如像素插值滤波和运动估计。Jiang 和 Nooshabadi^[53]利用 GPU 的运算能力和对数据使用的动态规划,开发了一种可扩展的并行搜索方法来加速多视点视频编码的运动估计和视差估计。Shahid 等^[54]提出了一种并行运动估计算法,最大程度地减少了宏块之间的数据依赖性,除了 GPU,此并行运动估计算法也在 ASIC 和 FPGA 上实现了。然而这些方案侧重于利用 GPU 对一个或多个特定模块进行并行化,而缺乏对 CPU 和 GPU 之间协作的研究,这不可避免地降低了异构系统在视频编码中的优势。也有一些研究开始利用异构处理平台的协同计算对计算复杂度进行优化。Xiao 等^[55]提出了一种多粒度并行解决方案,使用大规模并行运动估计算法来并行化 GPU 上的部分编码模块,同时通过 WPP 和自适应并行模式决策算法的配合,在 CPU 上加速了其他几种编码模块。Momcilovic 等^[56]提出了任务分配和动态负载平衡的方法,在混合 CPU+GPU 平台上并行化 H.264/AVC 帧间编码,与仅使用 GPU 的编码方案相比,该方法可以实现高达 2.5 倍的编码速度,然而 Momcilovic 也指出编码效率的损失一直是并行化视频编码器的一个重要考虑因素,它并没有得到很好的研究。相对于参考软件,现有 GPU 加速方案在压缩性能损失 5% 以内的前提下可以减少 40%~80% 的处理时间。

4 低延迟视频编码方案

本节总结了业界具有代表性的几种低延迟编码方案,分别对 SVT-HEVC、TPCast、JPEG-XS 和 WHDI 的编码实现方式做了分析。

SVT-HEVC 编码器由英特尔于 2018 年发布至开源社区^[57],支持 HEVC Main 和 Main10 profile 文件,并支持高达 8K@60 fps、4:2:0、8 位和 10 位的视频输入格式,和多达 13 个预置(M0~M12)的档次,在选择质量与数据密度之间进行权衡时提供了良好的粒度,用在基于视频云的直播场景中。

TPCast 为 VR(Virtual Reality)、AR(Augmented Reality)和 MR(Mixed Reality)提供无线解决

方案,编码方案基于 H. 264,可以实时传输 2K 画面^[58],在质量上是视觉无损的。

JPEG-XS 标准是 TICO SMPTE RDD35 的演进标准^[59],是一种图像编码标准,具有目前良好的轻量级编码功能,可用于直播、电视/移动设备、AR/VR 系统等场景的图像编码中。

WHDI 技术是以色列 Amimon 公司最早开发的一种无线高清接口技术^[60],专门为视频通信设计无线调制解调的方案,支持 1080p@60 fps 视频的浅压缩传输,主要用于家庭视频源设备与显示设备的连接。

表 2 列出了 H. 265(SVT-HEVC)、H. 264(TP-Cast)、JPEG-XS、WHDI 4 种低延迟编码方案的配置。由于所采用的基础编码器和面向的应用场景不同,各类标准的方案和延迟差异较大,从 H. 265 到 WHDI 的延迟依次降低,覆盖了 1~1 000 ms 的延迟范围,如图 7 所示,视频质量相近时,不同编码方案对编码延迟和压缩率的权衡力度不同,编码延迟越小表示编码速度越快,压缩率越大表示编码器的性能越好,本节将对表中几种标准的优化方案进行详细介绍与分析。

表 2 编码方案配置

Table 2 Configuration of different encoding schemes		SVT-HEVC	TPCast	JPEG-XS	WHDI
编码模块	帧间/帧内预测	✓	✓	×	×
	变换	✓	✓	✓	✓
	量化	✓	✓	✓	✓
	熵编码	✓	✓	极简	×
延迟优化技术	参考关系简化	✓	✓	×	×
	硬件加速	✓	✓	✓	×
	帧内刷新	×	✓	×	×
	并行	✓	×	✓	×

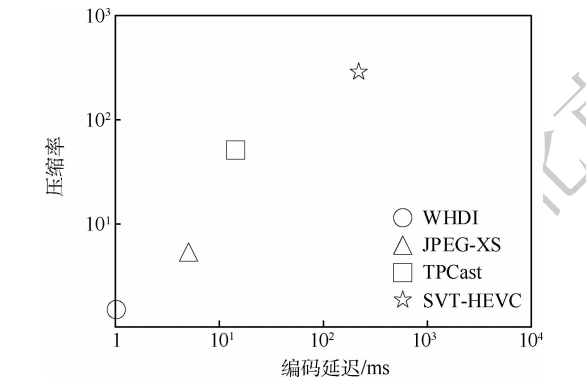


图 7 编码方案对比

Fig. 7 Comparison of different encoding schemes

4.1 H. 265 低延迟方案

目前,H. 265 已经拥有了针对不同平台的实现方案,如基于 CPU 的 x265、SVT-HEVC、NETINT 基于专用芯片设计的 Codensity 编码器^[61],以及

NVIDA 基于 GPU 设计的 NVENC 编码器等。在硬件编码器中,Codensity 和 NVENC 都在 H. 265 编码标准的基础上获得了超低的编码延迟,其中 Codensity 的编码延迟约 5 ms,NVENC 的编码延迟约为 3~10 ms。在软件编码器中,SVT-HEVC 相比于 x265 优势更加明显,SVT-HEVC 的 M11 视觉模式与 x265 的 very slow 档次的视觉质量相似,加速因子为 176 倍,SVT-HEVC 的 M9 视觉模式与 x265 的 medium 档次的视觉质量相似,加速因子为 15 倍,同时 SVT-HEVC 的使用成本更低,与硬件编码器相比可塑性更好。

SVT-HEVC 基于可伸缩视频技术(Scalable Video Technology,SVT)架构开发而来。SVT 是一种基于软件的视频编码技术,允许编码器在可伸缩处理器上实现性能、延迟和视觉质量之间的最佳折中^[62]。SVT-HEVC 是 H. 265 标准面向低延迟场景被广泛认可的版本,其最快档次的延迟在百毫秒级别,压缩比在 300:1 左右,配合其他低延迟技术可以到 100 ms 甚至几十毫秒级别。

SVT 架构较为灵活,其将原有视频编码流程拆解为资源调度、图像预分析、运动估计、码率控制、核心编解码、熵编码等模块,并基于此形成流水线,如 3.2 节所述。SVT 的流水线分为帧级和帧内编码超级块(Super Block,SB)级,2 种程度的流水线因编码模块/任务的特点均有 2 种方式,以 SB 级流水线为例,一种方式为:编码模块开启多个进程等待上一个模块输出数据,相应数据输入便开始作业。另一种方式为:编码模块开启单个进程,等待足够多的 SB 数据输入方可进行按序的作业。

基于上述流水线方式,SVT-HEVC 在 H. 265 标准上实现了进程级、帧级和块级三维并行的低延迟编码,同时根据整体复杂度目标和编码单元的人类视觉系统(Human Visual System,HVS)重要性决策 SB 划分模式,实现了编码速度和视觉质量的联合优化。

4.2 H. 264 低延迟方案

H. 264 标准已经被工业界广泛认可和应用,其作为 H. 265 的上一代标准,本身的编码复杂度相对较低,开源软件编码器 x264 虽然侧重实用性,但是对高压缩效率的追求导致编码延迟过高,现有低延迟方案大都基于硬件设计,以 TPCast 方案为例,通过采用帧内刷新等优化,可以达到 50:1 的压缩比和 10 ms 级别的延迟。

TPCast 使用 CAST 公司的 H264-E-BPF IP 核编码器^[63]。其基于 H. 264 Baseline Profile,因而

与常规的混合视频编码结构相同,由预测、变换、量化、熵编码等步骤组成。为实现低延迟的无线VR,TPCast编码方案采用了CAVLC熵编码选项以降低熵编码复杂度。

除熵编码外,该编码器最重要的是采用了帧内刷新技术克服由编码方式引起的延迟。H264-E-BPF可以使用CQP模式得到码流,同时也可以通过宏块级别的码率控制得到恒定比特率(Constant Bit Rate, CBR)的视频流。在CBR模式中,编码器可以快速响应视频内容在时域和空域上的变化,并与帧内刷新技术相结合,有效降低比特率峰值、编解码缓冲区延迟,同时保留周期性帧内编码的参考关系。

4.3 JPEG-XS 低延迟方案

JPEG-XS是一种低延迟图像编码标准,没有考虑帧间/帧内参考关系,同时简化了去相关变换、熵编码等模块,官方称标准延迟控制在毫秒甚至微秒的级别,同时压缩比与前述2种标准差距也较大,视觉无损情况下通常为2:1~6:1。

联合图像专家组(JPEG)在2017年初推出了第一版JPEG-XS测试模型(JPEG-XS test Model, XSM)^[64],同年发布JPEG-XS的第一个“委员会草案”版本,基本上确定了核心技术。在XSM测试模型设计方案中,输入图像依次经过拉伸、DC偏移去除、可逆颜色变换、小波变换、预量化(8位右偏移)、常规量化、熵编码后得到压缩信号。

JPEG-XS标准在提案征集阶段便对编码延迟作出了严格要求。从定稿的XSM模型中可以看到,JPEG-XS没有引入视频编码中有效提升效率的基于参考块的预测过程,而是以去相关变换和熵编码为核心实现低延迟、高质量、压缩比适中的图像压缩。

低延迟要求同样对变换等模块的设计和选取有限制。与基于块的编码方案中采用的变换方法不同,JPEG-XS最终选取了低复杂度、保证质量、更为灵活的DWT变换作为主要的去相关方式,并在5/3、13/7等不同长度的滤波器中选择了长度短、延迟低的5/3小波,其可以在1~2个垂直分解级别中运行。依据垂直分解级别及其产生的延迟、质量的差异,标准定义了Main、Light、Light-subline、High 4个配置文件用于不同内容、不同传输环境的编码,以支持不同级别的延迟和复杂度设置。

JPEG-XS对压缩比的要求相对宽松,因而设计的熵编码模块极简单。其将变换量化后的4个水平相邻小波系数组合成编码组,对于每个编码

组,计算组合内系数的数据密度最大的位平面,记录位平面特征的数在标准中称为“MSB(Most Significant Bit)位置”,熵编码进而在此基础上传递MSB位置和量化小波系数的绝对值等信息。可见熵编码过程产生的延迟较小,因而作为毫秒级甚至微秒级的图像编码标准,JPEG-XS产生的延迟主要集中在去相关变换上。

4.4 WHDI 低延迟方案

WHDI标准是高清数字家电接口协议,具体实施时可以提供足够的无线传输带宽,通过采用浅压缩视频编码方案保证了毫秒级的编码延迟,压缩比约为1.7:1。由于其总体方案面向短距离无线传输,因而需要编码器同时具备低延迟和配合系统发送数据的能力。WHDI中信源编码端发送压缩数据的系统采用视频数据分组的思想,根据视频数据的重要性将数据打包为不同的分组,通过适当地选择发送时间窗的长度来传输分组,可以权衡视频质量和系统延迟两者的关系,高质量视频对应着更大的延迟,低质量视频对应着更小的延迟^[65]。

WHDI标准中为了追求低编码延迟,在编码端不涉及编码预测模块,仅采用DCT变换进行空间去相关,同时,还有块交织等模块用于增强系统的抗差错能力^[66],减少了数据重传带来的传输延迟。

5 结 论

在实时视频通信中,以用户的QoE(Quality of Experience)为优化目标,即提供更低的系统延迟和更好的视频质量。低延迟视频编码器是实时视频通信系统的重要组成部分,本文总结了常用的低延迟视频编码技术,并分析了这些技术在业界低延迟编码方案中的应用,为低延迟视频编码器的设计提供了借鉴思路。

随着云游戏等超低延迟场景的出现,低延迟视频编码技术的重要性日益凸显,但同时也存在标准兼容性差、缺乏完整的理论支撑、延迟与性能失调等问题,仍然有进一步的研究空间:

1) 完善传统的率失真优化模型,现有的延迟约束的率失真优化模型应用在H.264标准上,没有充分考虑影响编码延迟的因素,理论分析框架比较简单,随着视频编码标准的不断演进,出现了许多新的编码工具,编码延迟的来源也越来越多,需要新的标准下继续拓展率失真优化模型。

2) 在编码器的计算复杂度方面利用机器学习等新方法充分发挥数据的潜力,提高编码单元

划分和编码参数选择的准确性,同时采用轻量高效的方案减少数据训练的额外开销。

3) 在码率控制方面提升细粒度编码单元的比特控制精度,结合强化学习等新思路维持码率的平稳,从而降低缓冲区延迟。

4) 以低延迟编码器为基础构建实时视频传输系统,通过系统的整体架构优化,在视频观看质量和系统延迟之间进行权衡。

未来的低延迟编码方案应能够按照不同应用场景的需求实现编码延迟的弹性配置,并通过与低延迟传输协议配合,最终达到超低延迟视频通信的目标。

参考文献 (References)

- [1] CULLEN C. The global internet phenomena report [EB/OL]. [2020-08-26]. <https://www.sandvine.com/press-releases/sandvine-releases-2019-global-internet-phenomena-report>.
- [2] 高文,赵德斌,马思伟. 数字视频编码技术原理[M]. 2版. 北京:科学出版社,2018:17.
GAO W, ZHAO D B, MA S W. Principles of digital video coding technology [M]. 2nd ed. Beijing: Science Press, 2018: 17 (in Chinese).
- [3] SCHREIER R M, ROTHERMEL A. A latency analysis on H. 264 video transmission systems [C] // Proceedings of the 2008 Digest of Technical Papers-International Conference on Consumer Electronics. Piscataway: IEEE Press, 2008: 307-308.
- [4] SULLIVAN G J, WIEGAND T. Rate-distortion optimization for video compression [J]. IEEE Signal Processing Magazine, 1998, 15(6): 74-90.
- [5] BICHON M, LE T J, ROPERTOPERT M, et al. Inter-block dependencies consideration for intra coding in H. 264/AVC and HEVC standards [C] // Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2017: 1537-1541.
- [6] BICHON M, LE T J, ROPERT M, et al. Low complexity joint RDO of prediction units couples for HEVC intra coding [C] // Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2018: 1733-1737.
- [7] PANG C, AU O C, ZOU F, et al. Optimal distortion redistribution in block-based image coding using successive convex optimization [C] // Proceedings of the 2011 IEEE International Conference on Multimedia and Expo Workshops. Piscataway: IEEE Press, 2011: 1-5.
- [8] WU Q, XIONG J, LUO B, et al. A novel joint rate distortion optimization scheme for intra prediction coding in H. 264/AVC [J]. IEEE Transactions on Information and Systems, 2014, 97(4): 989-992.
- [9] YANG T, ZHU C, FAN X, et al. Source distortion temporal propagation model for motion compensated video coding optimization [C] // Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops. Piscataway: IEEE Press, 2012: 85-90.
- [10] LI S, ZHU C, GAO Y, et al. Lagrangian multiplier adaptation for rate-distortion optimization with inter-frame dependency [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(1): 117-129.
- [11] GONZÁLEZ-DE-SUSO J L, MARTÍNEZ-ENRÍQUEZ E, DÍAZ-DE-MARÍA F. Adaptive Lagrange multiplier estimation algorithm in HEVC [J]. Signal Processing: Image Communication, 2017, 56: 40-51.
- [12] YANG K, WAN S, GONG Y, et al. An efficient Lagrangian multiplier selection method based on temporal dependency for rate-distortion optimization in H. 265/HEVC [J]. Signal Processing: Image Communication, 2017, 57: 68-75.
- [13] ZHANG F, BULL D R. An adaptive Lagrange multiplier determination method for rate-distortion optimisation in hybrid video codecs [C] // Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2015: 671-675.
- [14] WANG X, SONG L, LUO Z, et al. Lagrangian method based rate-distortion optimization revisited for dependent video coding [C] // Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2017: 3021-3025.
- [15] LI C, WU D, XIONG H. Delay—power-rate-distortion model for wireless video communication under delay and energy constraints [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(7): 1170-1183.
- [16] LI C, XIONG H, WU D. Delay-rate-distortion optimized rate control for wireless video communication [C] // Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2014: 5996-6000.
- [17] HUANG B, CHEN Z, CAI Q, et al. Rate-distortion-complexity optimized coding mode decision for HEVC [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(3): 795-809.
- [18] CARBALLEIRA P, CABRERA J, ORTEGA A, et al. A graph-based approach for latency modeling and optimization in multiview video encoding [C] // Proceedings of the 2011 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON). Piscataway: IEEE Press, 2011: 1-4.
- [19] WENGER S. Temporal scalability using P-pictures for low-latency applications [C] // Proceedings of the 1998 IEEE Second Workshop on Multimedia Signal Processing. Piscataway: IEEE Press, 1998: 559-564.
- [20] PAN Z, JIN P, LEI J, et al. Fast reference frame selection based on content similarity for low complexity HEVC encoder [J]. Journal of Visual Communication and Image Representation, 2016, 40: 516-524.
- [21] PARK S H, DONG T, JANG E S. Low complexity reference frame selection in QTBT structure for JVET future video coding [C] // Proceedings of the 2018 International Workshop on Advanced Image Technology (IWAIT). Piscataway: IEEE Press, 2018: 1-4.
- [22] DESHPANDE S, HANNUKSELA M M, KAZUI K, et al. An improved hypothetical reference decoder for HEVC [C] // Interna-

- tional Society for Optics and Photonics. Bellingham: SPIE-INT SOC Optical Engineering, 2013: 866608.
- [23] SCHREIER R M, RAHMAN A M T I, KRISHNA-MURTHY G, et al. Architecture analysis for low-delay video coding [C] // Proceedings of the 2006 IEEE International Conference on Multimedia and Expo Workshops. Piscataway: IEEE Press, 2006: 2053-2056.
- [24] MEENDERINCK C, AZEVEDO A, JUURLINK B, et al. Parallel scalability of video decoders [J]. Journal of Signal Processing Systems, 2009, 57(2): 173-194.
- [25] CHI C C, ALVAREZ-MESA M, JUURLINK B, et al. Parallel scalability and efficiency of HEVC parallelization approaches [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1827-1838.
- [26] MOHAMED M, NEJMEDDINE B, NOUREDDINE B, et al. Performance evaluation of frame-level parallelization in HEVC intra coding using heterogeneous multicore platforms [C] // Proceedings of the 2018 International Conference on Applied Smart Systems (ICASS). Piscataway: IEEE Press, 2018: 1-6.
- [27] CHEN K, SUN J, DUAN Y, et al. A novel wavefront-based high parallel solution for HEVC encoding [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(1): 181-194.
- [28] WANG H, XIAO B, WU J, et al. A collaborative scheduling-based parallel solution for HEVC encoding on multicore platforms [J]. IEEE Transactions on Multimedia, 2018, 20(11): 2935-2948.
- [29] OHM J R, SULLIVAN G J, SCHWARZ H, et al. Comparison of the coding efficiency of video coding standards—Including high efficiency video coding (HEVC) [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1669-1684.
- [30] BOSSEN F, BROSS B, SUHRING K, et al. HEVC complexity and implementation analysis [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1685-1696.
- [31] TOPIWALA P, KRISHNAN M, DAI W. Performance comparison of VVC, AV1 and EVC [C] // Proceedings of Applications of Digital Image Processing XLII. Bellingham: SPIE-INT SOC Optical Engineering, 2019.
- [32] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(12): 1649-1668.
- [33] ZHANG T, SUN M T, ZHAO D, et al. Fast intra-mode and CU size decision for HEVC [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(8): 1714-1726.
- [34] MIN B, CHEUNG R C C. A fast CU size decision algorithm for the HEVC intra encoder [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(5): 892-896.
- [35] XIONG J, LI H, WU Q, et al. A fast HEVC inter CU selection method based on pyramid motion divergence [J]. IEEE Transactions on Multimedia, 2014, 16(2): 559-564.
- [36] PAN Z, KWONG S, SUN M T, et al. Early MERGE mode decision based on motion estimation and hierarchical depth correlation for HEVC [J]. IEEE Transactions on Broadcasting, 2014, 60(2): 405-412.
- [37] LIU Z, YU X, GAO Y, et al. CU partition mode decision for HEVC hardwired intra encoder using convolution neural network [J]. IEEE Transactions on Image Processing, 2016, 25(11): 5088-5103.
- [38] RYU S, KANG J. Machine learning-based fast angular prediction mode decision technique in video coding [J]. IEEE Transactions on Image Processing, 2018, 27(11): 5525-5538.
- [39] XU M, LI T, WANG Z, et al. Reducing complexity of HEVC: A deep learning approach [J]. IEEE Transactions on Image Processing, 2017, 27(10): 5044-5059.
- [40] AMESTOY T, MERCAT A, HAMIDOUCH W, et al. Tunable VVC frame partitioning based on lightweight machine learning [J]. IEEE Transactions on Image Processing, 2019, 29: 1313-1328.
- [41] TRAN T D, LIU L K, WESTERINK P H. Low-delay MPEG-2 video coding [C] // Proceedings of the SPIE-The International Society for Optical Engineering. Bellingham: SPIE-INT SOC Optical Engineering, 1997: 510-516.
- [42] TOURAPIS A M, LEONTARIS A, SUHRING K, et al. H. 264/14496-10 AVC reference software manual [EB/OL]. [2020-07-04]. [http://iphone.hhi.de/suehring/tml/JM%20Reference%20Software%20Manual%20\(JVT-AE010\).pdf](http://iphone.hhi.de/suehring/tml/JM%20Reference%20Software%20Manual%20(JVT-AE010).pdf).
- [43] DELA CRUZ A R, CAJOTE R D. Low complexity adaptive intra-refresh rate for real-time wireless video transmission [C] // Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA). Piscataway: IEEE Press, 2014: 1-5.
- [44] ZHOU Y R, LI G Q, NING S S. A new feedback-based intra refresh method for robust video coding [C] // Proceedings of the 2015 International Conference on Computer Science and Applications (CSA). Piscataway: IEEE Press, 2015: 218-221.
- [45] HENDRY, WANG Y K, CHEN J L, et al. CE11/AHG14: Test 3.1—An approach for support of GRA/GDR [EB/OL]. [2020-07-04]. http://phenix.it-sudparis.eu/jvet/doc_end_user/documents/15_Gothenburg/wg11/JVET-0124-v3.zip.
- [46] LEE Y G, SONG B C. An intra-frame rate control algorithm for ultralow delay H. 264/advanced video coding (AVC) [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2009, 19(5): 747-752.
- [47] CHOI H, NAM J, YOO J, et al. Rate control based on unified RQ model for HEVC [EB/OL]. [2020-07-04]. http://phenix.int-evry.fr/jct/doc_end_user/documents/8_San%20Jose/wg11/JCTVC-H0213-v3.zip.
- [48] CHOI H, YOO J, NAM J, et al. Pixel-wise unified rate-quantization model for multi-level rate control [J]. IEEE Journal of Selected Topics in Signal Processing, 2013, 7(6): 1112-1123.
- [49] WANG S, MA S, WANG S, et al. Quadratic p -domain based rate control algorithm for HEVC [C] // Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2013: 1695-1699.
- [50] LI B, LI H, LI L, et al. λ domain rate control algorithm for high efficiency video coding [J]. IEEE Transactions on Image Processing, 2014, 23(9): 3841-3854.
- [51] GAO W, KWONG S, JIA Y. Joint machine learning and game

theory for rate control in high efficiency video coding[J]. IEEE Transactions on Image Processing, 2017, 26(12): 6074-6089.

[52] KWONG S, ZHOU M, XUEKAI W E I, et al. Rate control method based on deep reinforcement learning for dynamic video sequences in HEVC [J]. IEEE Transactions on Multimedia, 2020, 99: 1.

[53] JIANG C, NOOSHABADI S. A scalable massively parallel motion and disparity estimation scheme for multiview video coding [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 26(2): 346-359.

[54] SHAHID M U, AHMED A, MARTINA M, et al. Parallel H. 264/AVC fast rate-distortion optimized motion estimation by using a graphics processing unit and dedicated hardware[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(4): 701-715.

[55] XIAO B, WANG H, WU J, et al. A multi-grained parallel solution for HEVC encoding on heterogeneous platforms[J]. IEEE Transactions on Multimedia, 2019, 21(12): 2997-3009.

[56] MOMCILOVIC S, ROMA N, SOUSA L. Exploiting task and data parallelism for advanced video coding on hybrid GPU + GPU platforms[J]. Journal of Real-Time Image Processing, 2016, 11(3): 571-587.

[57] INTEL. Scalable video technology for HEVC encoder (SVT-HEVC Encoder) [EB/OL]. (2020-08-20) [2020-08-26]. <https://github.com/OpenVisualCloud/SVT-HEVC>.

[58] SURUR. TPCAST's wireless Adapter coming to the Oculus Rift next week [EB/OL]. (2017-11-08) [2020-08-26]. <https://mspoweruser.com/tpcasts-wireless-adapter-coming-oculus-rift-next-week/>.

[59] BUYSSCHAERT C, DESCAMPE A, FÖBEL S, et al. Overview of JPEG XS [EB/OL]. 2018 [2020-08-26]. <https://jpeg.org/jpegxs/index.html>.

[60] AMIMON, Ltd. Wireless home digital interface™ specification v1.0 Revision 33 [EB/OL]. [2020-07-04]. <https://www.amimon.com/proavl>.

[61] NETINT. Codensity T408 video transcoder product brief [EB/OL]. [2020-08-26]. https://www.netint.ca/product/t400_transcoder/.

[62] INTEL. Scalable video technology for the visual cloud [EB/OL]. [2020-08-26]. https://01.org/sites/default/files/documentation/svt_aws_wp.pdf.

[63] CAST. H264-E-BPF ultra-fast AVC/H.264 baseline profile encoder [EB/OL]. [2020-08-26]. <https://www.cast-inc.com/compression/avc-hevc-video-compression/h264-e-bpf/>.

[64] RICHTER T, KEINERT J, FOESSEL S, et al. JPEG-XS-A high-quality mezzanine image codec for video over IP [J]. SMPTE Motion Imaging Journal, 2018, 127(9): 39-49.

[65] REZNIC Z. Successive refinement video compression; USA, W02018025211A1 [P]. 2018-02-08.

作者简介:

宋利 男, 博士, 教授, 博士生导师。主要研究方向: 新型视频编码、大数据压缩、移动计算视觉。

刘孝勇 男, 博士研究生。主要研究方向: 可伸缩视频编码、码率控制。

武国庆 男, 硕士研究生。主要研究方向: 视频编码。

朱辰 男, 博士研究生。主要研究方向: 视频编码。

黄琰 男, 博士研究生。主要研究方向: 视频编码。

解蓉 女, 博士, 副教授, 硕士生导师。主要研究方向: 视频编码与转码、图像/视频处理。

张文军 男, 博士, 教授, 博士生导师。主要研究方向: 图像通信与数字电视、宽带无线传输、系统芯片设计。

Low-latency video coding techniques

SONG Li^{1,2,*}, LIU Xiaoyong¹, WU Guoqing¹, ZHU Chen¹, HUANG Yan¹,

XIE Rong¹, ZHANG Wenjun^{1,2}

(1. Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China;

2. Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: With the widespread usage of video coding and transmission techniques, demands for video have increased dramatically. Real-time video communication has become the research focus of the video industry. Its core goal is to provide a better user experience and lower latency. Low-latency video coding is a key component for real-time video communication applications. The overall system latency can be effectively reduced by reducing the coding latency. First, this paper analyzes sources of latency in the video transmission system. Focusing on the general video coding framework, this paper introduces the generation mechanism of the coding latency. Then, mainstream video coding standards at home and abroad are outlined. The detailed description of the principle and models of rate-distortion optimization techniques provides a theoretical basis for the design of low-latency video encoders. Additionally, this paper summarizes how to optimize the coding latency in terms of the reference structures, pipeline design, encoding modes search, rate control, and hardware acceleration and generalizes industrial representative low-latency video coding schemes. Finally, this paper summarizes the limitations of the existing low-latency video coding techniques and presents future research directions.

Keywords: low latency; video coding; real-time video communication; rate-distortion optimization; encoding use cases

Received: 2020-08-26; **Accepted:** 2020-09-19; **Published online:** 2020-11-05 09:21

URL: kns.cnki.net/kcms/detail/11.2625.V.20201104.0859.001.html

Foundation items: National Key R & D Program of China (2019YFB1802701); National Natural Science Foundation of China (61671296)

* **Corresponding author.** E-mail: song_li@sjtu.edu.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0469

显著性引导的低光照人脸检测

李可夫¹, 钟汇才^{2,3}, 高兴宇^{2,*}, 翁超群², 陈振宇^{2,4},
李勇周², 王师峥^{2,3}

(1. 中国科学院大学 微电子学院, 北京 100049; 2. 中国科学院微电子研究所, 北京 100029;

3. 中国科学院物联网研究发展中心, 无锡 214000; 4. 国家电网有限公司大数据中心, 北京 100031)

摘

要: 针对卷积神经网络难以对低光照环境拍摄的图像进行人脸检测的问题。提出

了一种将图像显著性检测算法和深度学习相结合的算法,并应用于低光照人脸检测。所提算法将图像的显著性信息与图像原始 RGB 通道融合,用于神经网络训练。在低光照人脸数据集 DARK FACE 上进行了充分的实验,结果表明:所提方法在 DARK FACE 数据集上获得了比当前主流人脸检测算法更好的检测精度,进而验证了所提算法的有效性。

关键词: 人脸检测; 显著性引导; 深度神经网络; 低光照; 计算机视觉

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2021)03-0572-13

随着卷积神经网络的发展,作为通用目标检测中的一类任务,人脸检测到目前为止,已经取得了很大进步。除了 RCNN^[1]、Fast RCNN^[2]、Faster R-CNN^[3]和 YOLO^[4]等通用目标检测算法可以应用在人脸检测中,也有专为人脸检测设计的算法如 DSFD (Dual Shot Face Detector)^[5]、Pyramid-Box^[6]等,并取得了较好的效果。

由于目标检测任务中的背景环境各有差别,对检测结果也会产生相应影响和挑战,应对现实中各种真实视觉不利场景的挑战是一项热门研究课题。目前在人脸检测领域流行的基准数据集有 Fddb (Face Detection Data set and Benchmark) 数据集^[7]、MALF 数据集、WIDER FACE 数据集^[8]等,大部分人脸检测算法也是在以上数据集上进行评测的。2019 年北京大学发布的 DARK FACE 数据集提供了低光照图像和对应人脸标签,聚焦于低光照情况的人脸检测应用,这一条件下的人脸检测任务具有重要现实意义,在 CVPR 2019

UG2 + 国际竞赛中,设置了基于 DARK FACE 数据集的人脸检测项目。但是目前并没有专门为低光照条件设计的人脸检测算法,因此现有算法在 DARK FACE 数据集上的检测效果都不够理想^[9]。

本文在 CVPR 2019 UG2 + 国际竞赛 DARK FACE 检测项目获奖算法的基础上,提出了将深度神经网络和图像显著性方法结合的检测算法,通过将图像显著性信息加入到图像预处理和深度学习网络的训练过程中,帮助提升低光照人脸检测性能。具体而言,本文创新性地提出了将图像显著性检测与深度神经网络结合的人脸检测算法,该算法将显著性检测得到的显著图与图像原始信息融合,再进入神经网络一起进行训练。

作为对照,本文还将图像增强算法 MSRCR^[10]应用于数据集,对比验证图像显著性对人脸检测的帮助,此外,本文提出的检测算法还与其他人脸

收稿日期: 2020-08-27; 录用日期: 2020-10-21; 网络出版时间: 2020-11-05 13:29

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201104.1356.004.html

基金项目: 国家自然科学基金 (61702491, 61802390); 北京市自然科学基金 (4194095); 茂名市科技计划 (2020028)

* 通信作者。E-mail: gaopingyu@ime.ac.cn

引用格式: 李可夫, 钟汇才, 高兴宇, 等. 显著性引导的低光照人脸检测[J]. 北京航空航天大学学报, 2021, 47(3): 572-584.

LI K F, ZHONG H C, GAO X Y, et al. Saliency guided low-light face detection[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 572-584 (in Chinese).

检测算法进行了对比。实验结果表明,本文算法在 DARK FACE 数据集上获得了最好的检测精度。

1 相关工作

1.1 通用目标检测

目标检测是计算机视觉中的一大主要任务。近年来随着深度学习的发展,基于卷积神经网络的目标检测技术成为研究的热点。当前目标检测的算法主要可以分为两大类:

1) 一阶段目标检测算法。例如 YOLO、SSD^[11]等,这类算法没有使用候选区域(Region Proposal),直接通过主干网络同时检测出图像中物体的位置和类别信息。

2) 两阶段目标检测算法。包括 RCNN、Fast RCNN、Faster R-CNN 等,这类算法首先用相应的候选区域算法,类型有传统算法或者是神经网络,在输入图片上生成建议目标候选区域,然后再将所有的候选区域送入分类器中进行分类。

1.2 人脸检测

作为通用目标检测中的一类特例,人脸检测当今也倍受关注。目前有许多用于人脸检测算法评测的数据集,例如 FDDB,其重点用于研究表情、姿态、尺度、外观有较大可变性的无约束人脸检测,数据集图片来源于自然环境下拍摄的新闻报道图片,共计 2 845 张图片包含有 5 171 张人脸图像。2015 年发布的 WIDER FACE 数据集共有 32 203 张图片,包含 393 703 张人脸,而且在脸部的尺寸、姿势、表情、遮挡上都有很大变化,发布后被广泛应用于人脸检测算法的评估;还有 2015 年发布的 MAF 数据集,是为了更细粒度地评估野外环境中人脸检测的数据集。本文的研究内容是低光照条件下的人脸检测任务,选取了 2019 年发布的 DARK FACE 数据集进行研究和评价算法,该数据集提供了 6 000 张真实拍摄的低光照图像,这些图像拍摄于街道、教学楼、天桥、公园等处,其中人脸都进行了标注。DARK FACE 同时提供了 9 000 张未标注的低光照图像以及 100 张有标注的测试图像,同时 DARK FACE 官方提供了一个评分工具,通过计算检测结果与真实标签的差距得到精度用于衡量检测算法的性能。

人脸检测任务可以使用通用目标检测算法完成,也有专门为人脸检测任务设计的检测算法。

PyramidBox 就是一种专门为人脸检测任务设计的检测算法,提出了一种基于在图像上预设不

同大小、不同长宽比的候选锚点框 PyramidAnchors 的环境辅助方法,引入有监督的信息来为较小的、模糊的和部分遮挡的人脸学习环境特征,其设计的低层级特征金字塔网络(LFPN)能更好地融合环境特征和面部特征,同时可以较好地处理不同尺度的人脸。这种环境敏感的预测模型,能从融合的特征中学习准确的定位和分类,且该方法提出的关注尺度的 data-Anchor-sampling 策略,能改变训练样本的分布,重点关注较小的人脸。在通用人脸检测基准 FDDB 和 WIDER FACE 上,PyramidBox 达到了当时最佳水平。PyramidBox++^[12]在 PyramidBox 的基础上进行了改进,采用平衡数据锚点框采样(BDAS)策略,得到更均匀的不同大小人脸的采样。Dual-Pyramid Anchor 双金字塔锚点框通过引入渐进锚点框损失来促进特征学习,并通过密集上下文模块,将密集块合并到预测模块中,以有效地传递信息,并保持更多的多尺度上下文特性。

DSFD 也是一种专门为人脸检测任务设计的检测算法,在继承 SSD 检测框架的基础上,引入一种新颖的特征增强模块,结合 PyramidBox 低层的特征金字塔与 RFBNet^[13]中的感受野块(RFB)利用不同层级的特征信息,增强特征的可判别性和鲁棒性。渐进锚损失 PAL(Progressive Anchor Loss)将较小的锚拓展在高层特征图上以获得更多用于分类的语义信息和用于检测的高分辨率信息,并通过更小的锚去计算辅助监督损失。IAM(Improved Anchor Matching)策略提出了一种改进的锚点匹配方法,在 DSFD 中集成新的数据增强技术和锚设计策略提供给回归器更好的初始化,让锚和真实标签人脸尽可能匹配。

但是正如上文所言,这些经典的人脸检测算法均没有专门考虑低光照环境对人脸检测的影响,因此,直接将其应用于低光照环境人脸检测的性能仍然不理想。1.3 节将介绍一些经典的图像增强算法。

1.3 图像增强

在光照较低的环境下拍摄的照片往往存在亮度低、对比度差等问题。除了视觉上的主观效果,也会影响一些高层次的任务如检测与识别,因此低光照图像增强的研究具有很强现实意义。

图像增强的方法有很多,这里重点介绍基于 Retinex 理论^[14]的算法。Retinex 理论的基础是物体的颜色是由物体对光线的反射能力来决定的,而不是由反射光强度的绝对值来决定的,物体的

色彩不受光照非均匀性的影响,具有一致性。不同于传统的线性、非线性只能增强图像某一类特征的方法,Retinex 可以在动态范围压缩、边缘增强和颜色恒常 3 个方面达到平衡,因此可以对各种不同类型的图像进行自适应的增强。Retinex 原理如图 1 所示。图中: I 为人眼中看到的图像; R 为物体的反射分量; L 为环境光照射分量; (x, y) 为二维图像对应的像素位置。

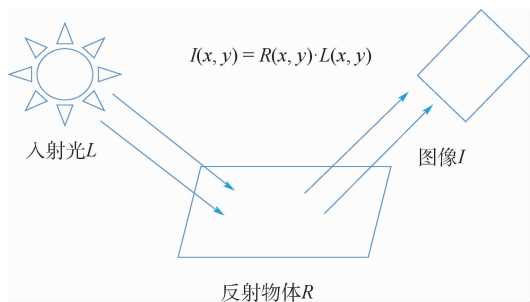


图 1 Retinex 原理示意图

Fig. 1 Schematic of Retinex principle

基于 Retinex 原理的代表算法有单尺度 Retinex SSR 算法^[15]。通过估算 L 来计算 R ,也就是图像增强最后的处理结果,具体来说, L 可以通过高斯模糊和 I 做卷积运算求得

$$\begin{cases} \log R = \log I - \log L \\ L = F * I \end{cases} \quad (1)$$

式中: F 为高斯模糊的滤波器;“ $*$ ”表示卷积运算。

在 SSR 增强算法的基础上,又发展出了多尺度 Retinex MSR 算法^[16],对原始图像进行不同尺度的高斯模糊,再将不同尺度的结果进行加权平均。以及带颜色恢复的多尺度 Retinex MSRCR 算法,在 MSR 的基础上,加入了色彩恢复因子来调节由于图像局部区域对比度增强而导致颜色失真的缺陷。

LIME^[17]算法也是基于 Retinex 理论的算法。该算法选取图像中每个像素通道中的最大值估计该图像各像素的光照度,然后通过强加入一种结构先验来细化光照度,最后根据 Retinex 理论合成增强结果。

本文选取 MSRCR 算法用于低光照人脸检测的图像增强预处理,与利用图像显著性做图像增强的方法对照。

1.4 图像显著性检测

图像显著性是图像中重要的视觉特征之一,广泛运用在压缩、编码、图像边缘加强、显著性目标分区和提取等方面。人眼在观察一副图像的时候,对图像各区域的重视程度不同,会把注意力聚

焦于图像中的某一部分区域,比如图中的一个建筑物或是一个人,这种机制使得观察者可以迅速处理看到的场景。

图像显著性检测的目的就是模拟人眼视觉注意力机制检测图像中的显著区域,也就是计算图像中不同区域的重要程度,从而将显著性高的前景和显著性低的背景分离。显著性检测与许多应用密切相关,包括目标检测、图像视频压缩、图像检索和图像重定向等。

显著性检测有以下多种算法:

Yun 和 Mubarak^[18]提出通过计算某个像素在整个图像上的全局对比度,即该像素与图像中其他所有像素在颜色上的距离之和,作为该像素的显著值。

HC (Histogram-based Contrast) 算法^[19]考虑了彩色信息。并且提出了降低颜色数量的方案,将 RGB 各分量分别映射成 12 等份,则映射后的图最多只有 $12 \times 12 \times 12$ 种颜色,这样就可以构造一个较小的直方图以提高计算效率。

Achanta 等^[20]基于局部对比度的 LAB 颜色空间计算距离,通过计算一个感知单元在不同邻域上的局部对比度来实现多尺度的显著性计算。

近年来随着深度学习的发展,也有许多基于深度学习的显著性检测算法,如 2019 年提出的基于深度卷积神经网络的特征图深度融合模型^[21]。

在人脸检测领域,此前也有过对人脸图像进行显著性检测的实验,如 Jian 等^[22]提出的模拟人类视觉系统进行显著目标检测的视觉注意感知模型。

本文的工作聚焦于低光照下的人脸检测应用,低光照环境带来的主要挑战是图像亮度不足、前后景难以区分,影响神经网络特征提取和学习。所以本文考虑的方案包括进行图像增强以及使用显著性检测分离前后景,帮助神经网络训练。具体实验包括将图像增强算法 MSRCR 与深度神经网络相结合进行人脸检测,以及将图像显著性检测算法应用于神经网络,实现了一个低光照人脸检测器。

在将图像显著性检测算法应用到人脸检测时,采用了 2 种方案:一种是将图像显著性应用于低光照图像增强,使用显著性引导增强后的数据集训练;一种是将图像原始信息结合图像显著性信息同时用神经网络进行训练的方法。

2 低光照人脸检测算法

本文所介绍的低光照人脸检测算法,既包括

了笔者在 CVPR 2019 UG2 + 国际竞赛 DARK FACE 检测项目获奖算法中使用的工程技巧,同时也融合了本文新提出的基于显著性引导的预处理方法和深度网络训练方法。本节将对这些工程技巧和创新算法进行详细介绍。

2.1 骨干网络

本文的检测算法选取 RetinaNet 神经网络^[23]作为骨干网络。RetinaNet 是一个用于目标检测的一阶段目标检测算法。

当前目标检测的 2 类算法中,一阶段目标检测算法的速度相对较快,而两阶段目标检测算法的准确度相对较高。导致一阶段目标检测算法的准确度不如两阶段目标检测算法的原因是正负样本的类别不均衡,在一张图片中,可能只有一小部分是包含目标的,而大部分是属于背景。由于负样本数量太多,且一般都是易分类的,这类易分类的负样本会对神经网络的训练起到主导作用,导致网络学不到有效的信息从而无法对目标准确分类。而两阶段目标检测算法会首先生成候选区域,这一步已经筛去了大量易分类的负样本,从而避免了正负样本的类别不均衡。为了解决正负样本类别不均衡的问题,RetinaNet 提出了一种新的损失函数:Focal loss,这个损失函数基于交叉熵损失(CE loss)函数修改得到:

$$CE(p, y) = \begin{cases} -\log p & y = 1 \\ -\log(1 - p) & \text{其他} \end{cases} \quad (2)$$

式中: y 为样本类别标签, $y = 1$ 为正样本; $p \in [0, 1]$ 为该类的估计概率。

定义 P_i 为

$$P_i = \begin{cases} p & y = 1 \\ 1 - p & \text{其他} \end{cases} \quad (3)$$

那么交叉熵可以表示为

$$CE(p, y) = CE(P_i) = -\log(P_i) \quad (4)$$

为了平衡正负样本的作用,赋予一个权重因子 $\alpha \in [0, 1]$ 给正样本, $1 - \alpha$ 给负样本,以类似 P_i 的方式定义 α_i ,损失函数可以重写为

$$CE(p, y) = -\alpha_i \log(P_i) \quad (5)$$

权重因子 α 可以平衡正负样本的重要性,但是不能帮助区分易分类样本和难分类样本。因此 Focal loss 还增加了调节因子 $(1 - P_i)^\gamma$ 再次修改损失函数,控制易分类样本的权重,以抑制容易分类的样本的梯度产生的负面作用:

$$FL(p, y) = FL(P_i) = -\alpha_i (1 - P_i)^\gamma \log(P_i) \quad (6)$$

当一个样本被错误分类, P_i 是很小的,对应的调节因子就接近于 1,进而这个样本权重就更大。如果 P_i 接近 1,那么调节因子就趋于 0,这个易分类样本的权重就被降低。参数 γ 可以控制分类样本被降低权重的比率。当 $\gamma = 0$, Focal loss 就等于 CE loss,在实际操作中, $\gamma = 2$ 效果最好。

Focal loss 通过调整正负样本的比例以及减少易分类样本的权重,使得神经网络在训练时更专注于难分类的样本。为了证明 Focal loss 的有效性,笔者设计了 1 个一阶段的目标检测网络 RetinaNet,其网络结构如图 2 所示。

RetinaNet 是由 1 个主网络和 2 个子网络组成的网络结构。主网络负责计算整个输入图像的卷积特征,这里主网络在残差网络(ResNet)^[24]结构上使用特征金字塔网络(FPN)^[25]来生成丰富的、多尺度的卷积特征。FPN 通过自上而下的通道和横向的连接来扩展了卷积神经网络,提升了卷积神经网络的多尺度预测能力。主网络输出进入 2 个子网络,第 1 个子网络用来对主网络的输出进行进一步的计算来完成边框分类;第 2 个子网络负责边框对真实标签的回归。其中分类子网络采用 Focal loss 作为损失函数,边框回归子网络采用 smooth L1 loss 作为损失函数。

Lin 等采用 COCO 数据集^[26]测试了 RetinaNet 的性能,并和其他目标检测算法进行比较,结果表明,RetinaNet 在 COCO 数据集上的检测精度不仅超过了 YOLOv2^[27]、SSD 和 DSSD^[28] 这些一阶段目标检测算法,还超过了 Faster R-CNN 等两阶段目标检测算法。

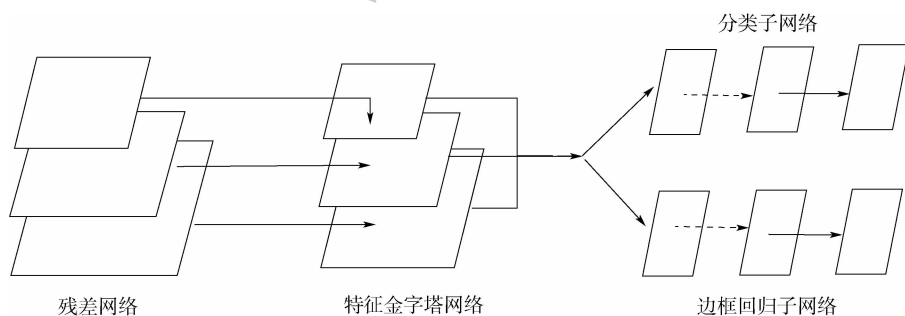


图 2 RetinaNet 网络结构

Fig. 2 Network structure of RetinaNet

RetinaNet 的检测效果较好,并且是结构较为简单,速度较快的一阶段目标检测算法。本文设计的低光照人脸检测算法采用 RetinaNet 作为核心神经网络。

2.2 基于 WIDER FACE 数据的预训练

本文方法通过训练神经网络 RetinaNet 识别低光照人脸图像。由于 DARK FACE 数据集提供的训练图片只有 6 000 张,样本数量较小,神经网络直接训练的效果并不理想。所以先采用样本量更大,人脸在尺寸、遮挡方面有丰富多样性的人脸数据集 WIDER FACE 做预训练,再将模型迁移到 DARK FACE 数据集进行训练,可以加快神经网络的收敛速度。WIDER FACE 数据集示例如图 3 所示。

进行预训练时选取 101 层的 ResNet 作为 RetinaNet 的主干网络。训练过程的主要细节包括:

1) 预定义的锚点框设置,一个长宽比:1:1.25, 2 个尺度:2S 和 $2\sqrt{2}S$ (S 为对应特征图的下采样尺度),最终特征图上每个位置有 2 个锚点框。本文的长宽比和尺度的选择参考了 2019 年刷新 WIDER FACE 多项记录的人脸检测算法 AFD-HP^[29] 的设置。

2) 对训练集使用了数据扩充的数据增强方法。通过水平翻转的方法处理训练集图像和标签,将翻转后的图像和标签加入训练。

3) 训练时 Batch size 设置为 4,这和本文实验时硬件环境的 GPU 数量相同。

4) 初始学习率设置为 0.01,迭代 10 万步后降低为 0.001,训练过程共进行了 20 万步迭代。前 1 000 步采用预热 (warm up) 的方式设置学习率,原理是在训练初期需要选择比较大的学习率,但是较大的学习率容易使训练过程具有较大的不稳定,所以把前 1 000 步作为学习率的热身阶段,在该阶段将学习率从 0 逐渐增加到指定的学习率。

5) 使用随机梯度下降 (SGD) 算法更新梯度,

权重衰减 (weight_decay) 因子设置为 0.000 1,动量 (momentum) 因子设置为 0.9。

6) 使用 python 库 torchvision 中提供的 ResNet101 预训练模型。

经过对 WIDER FACE 数据集的训练,得到 1 个用于 DARK FACE 数据集的预训练模型,该模型在 WIDER FACE 验证集上测试得到的精度为 0.697。

将得到的 WIDER FACE 模型直接用于检测 DARK FACE 的测试集,使用 DARK FACE 官方提供的评价工具,得到精度为 0.194,和在 WIDER FACE 测试集上的结果相差很远,意味着尽管预训练使用的数据集覆盖了姿势、外观、比例等方面的变化,但仍然不足以捕捉低光照条件下的面部特征。

由于模型在 WIDER FACE 测试集已经得到较高准确率,可以认为该模型对人脸已经有较好的识别能力,但是对于拍摄于低光照环境的图像的识别能力较差。接下来便是用预训练的 WIDER FACE 人脸检测模型为基础,迁移到 DARK FACE 数据集继续训练。

2.3 基于预训练模型的迁移学习

使用 DARK FACE 的训练集图片在 WIDER FACE 的模型上继续训练进行迁移学习。训练使用了 DARK FACE 提供的全部 6 000 张带标注图片,测试使用 DARK FACE 公布的 100 张图片测试集。为了验证图像增强和显著性对神经网络训练的作用,分别进行了以下实验。

2.3.1 使用 DARK FACE 原始图像训练

加载基于 WIDER FACE 数据集预训练的模型,使用 DARK FACE 数据集的原始图像和标签进行迁移学习 (见图 4)。

训练时 Batch size 设置为 4,学习率固定为 0.01,前 1 000 步采用 warm up 启动。训练完的模型在 DARK FACE 测试集上检测得到的精度为 0.504。

2.3.2 MSRCR 算法增强 DARK FACE 数据集再训练

图像增强算法可以让低光照图像亮度提高,首先尝试使用 MSRCR 算法增强测试集图片提高图片亮度,从图 5 中可以看到增强后的图片有很好的主观视觉效果。之后首先直接使用 WIDER FACE 预训练模型对增强后的测试集检测,得到的精度为 0.300。这一步验证了将 MSRCR 图像增强算法直接应用于预训练的人脸检测模型上,就可以帮助 RetinaNet 进行人脸检测。

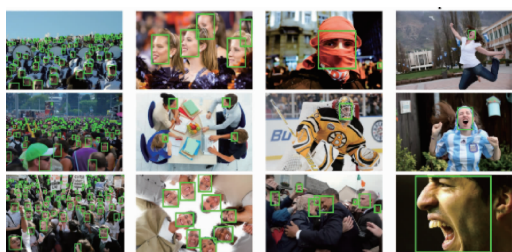


图3 WIDER FACE 数据集图像示例

Fig.3 Image examples of WIDER FACE dataset



图 4 DARK FACE 数据集原图

Fig. 4 Original images of DARK FACE dataset

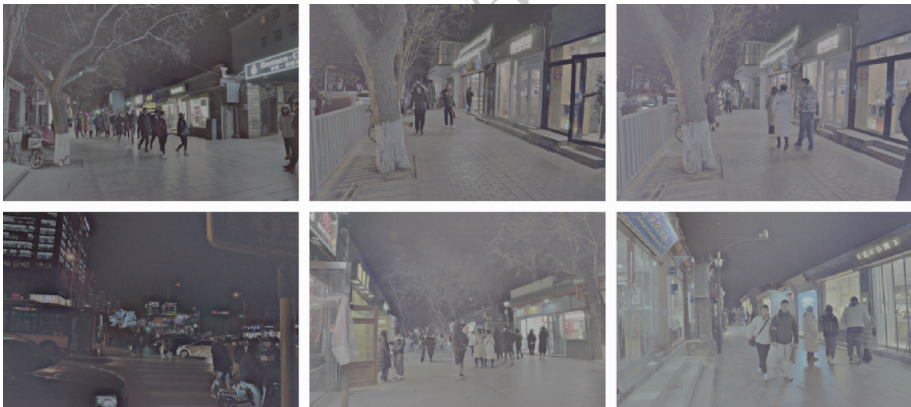


图 5 采用 MSRCR 算法的增强图像

Fig. 5 Images enhanced using MSRCR algorithm

接下来再使用 MSRCR 图像增强算法处理 DARK FACE 训练集,再将处理后的训练集用于 RetinaNet 训练。使用训练好的 WIDER FACE 模型作为预训练模型进行迁移学习。为了验证图像增强算法的作用,训练过程与 2.3.1 节中使用的超参数完全一致,只是训练集使用了 MSRCR 算法增强。最终得到的检测模型在 DARK FACE 测试集上检测得到的精度为 0.522。

2.3.3 显著性引导增强 DARK FACE 训练集再训练

本文使用的图像显著性检测算法由侯晓迪博士等在 2012 年提出^[30]。该算法对图片处理的结果是计算生成了图像不同区域显著性的差距,并以不同大小的权重数值加以区分。对于前景给予更大的权重,背景则给予更小的权重,这一结果以一个显著图的形式表现。

以 DARK FACE 的图像为例,每张图的尺寸是 $720 \times 1\,080$ 像素,每个像素包括 RGB 3 个通道。显著性检测算法的结果是一个 $720 \times 1\,080$ 大小的矩阵,矩阵中每个值对应图像上相应像素位置的权重,这里将这一矩阵称为显著图。

显著图的实现过程:显著图生成的基础在于

将图像中的前景部分分离出来。这一点可以通过离散余弦变换 (DCT) 将图片从空间域转换到频域,再通过 sign 操作实现近似的分离前景的作用。然后进行逆变换,使用逆离散余弦变换 (IDCT) 操作将其带回空间域。过程可表示为 $\text{IDCT}(\text{sign}(\text{DCT}(x)))$ (7) 式中: x 为图像的空间域信息。

得到重构的空间域信息后,再通过平均图像 3 个通道中的像素值,将 3 个通道的信息集成到单通道中。

再对每个通道 c 的信息进行平滑抑制噪声,这里使用了高斯核 K_g 以及矩阵的哈达玛积 $\circ^2$,最终可以得到显著图的计算方法如下:

$$\sum_c K_g * \left[\frac{1}{3} \text{IDCT}(\text{sign}(\text{DCT}(x^c))) \right]^{\circ^2} \quad (8)$$

式中: “ $*$ ” 表示卷积。

对 DARK FACE 数据集中的每幅图像使用显著性检测算法计算出显著图,将显著图中每个像素点的值映射到 0 ~ 255 的范围,显示为灰度图后可可视化的结果如图 6 所示。

显著图中的信息可以应用于低光照图像增

强,将待处理图片的像素乘以该矩阵对应位置的值,便能得到根据图像不同位置显著性差异进行处理的结果,再将该结果乘上一个固定系数 r 用来调节图像亮度,便可以得到显著性算法对低光照图像的最终增强结果。可表示为:图像增强结果 = 图像 $\times$ 显著图 $\times$ 亮度调节系数 r ,显著图中的权值范围映射为 $0 \sim 255$ 。实验尝试过不同的系数调节图像亮度,分别取亮度调节系数 $r = 0.1$

以及 $r = 0.2$ 时图像增强结果如图 7 和图 8 所示。可以看出,显著性引导的图像增强算法对图像的增强效果不是均匀的而是有局部突出的,并且由主观效果可发现其中突出部分可以显示出在图像中的光源亮度变化或是物体与周围场景之间的较大区别,如行人的所在位置和周围地面之间的权重有区别,也就是起到了图像分离前景与背景的效果。

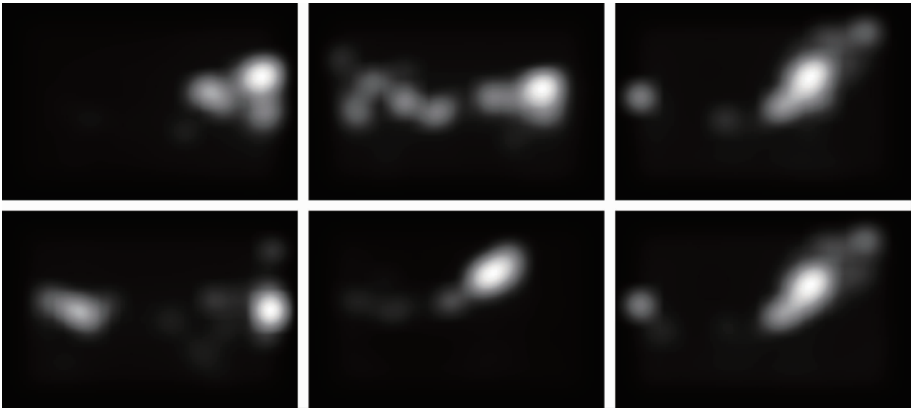


图 6 DARK FACE 数据集图像的显著图
Fig. 6 Saliency maps of DARK FACE dataset images



图 7 显著性引导增强的 DARK FACE 图像 ($r = 0.1$)
Fig. 7 DARK FACE images enhanced using saliency guidance ($r = 0.1$)



图 8 显著性引导增强的 DARK FACE 图像 ($r = 0.2$)
Fig. 8 DARK FACE images enhanced using saliency guidance ($r = 0.2$)

本文尝试使用了不同取值的 r 值调节图像亮度。将 r 设为 0.1 时使用 WIDER FACE 预训练模型在 DARK FACE 测试集上检测得到的精度为 0.299, r 设为 0.2 时得到的精度为 0.304。

使用显著性引导增强的 DARK FACE 图片用于训练,以及训练好的 WIDER FACE 作为预训练模型进行迁移学习。为了验证图像增强算法的作用,训练过程与 2.3.1 节中使用的超参数完全一致,只是训练集使用了显著性引导增强。最终得到的检测模型在 DARK FACE 测试集上检测得到的结果是 r 设 0.1 时精度为 0.540, r 设 0.2 时为 0.533(见图 8)。

对比可以发现,MSRCR 算法处理数据再训练得到的效果并没有优于直接使用原图进行训练的效果。而使用显著性算法进行图像增强再训练,得到的效果可以优于 MSRCR 算法处理的结果。验证显著性引导的图像增强方法对于 RetinaNet 神经网络可以起到比 MSRCR 算法更大帮助。并且可以发现,虽然设置不同的系数调节图片亮度后主观效果差异很大,但是对使用增强算法后的图片训练得到的模型的检测精度影响不大。

2.4 显著图融合训练

显著性引导增强图像的方法根据显著图来对图像中不同位置进行不同程度的增强。使用图像增强结果训练后的效果比原图训练更好,验证了图像显著性信息可以帮助神经网络学习。而这种方法与使用 MSRCR 算法增强图像以及用原始图像训练的方法都不需要修改核心网络结构。下文提出方法的不同之处在于同时将原始图像和显著图加入训练,网络结构也有相应变化。

考虑到对图片进行基于显著图的增强的过程

中,原始图像乘上对应矩阵再调节最终亮度系数的过程,可能会使得图片的像素值因为超过值域而丢失一部分信息。以及直接使用显著性引导增强的最终结果来作为训练样本,虽然效果总体比不做处理更好,但也可能原图和显著性引导增强后的结果相比也有对于神经网络更有效的信息,所以这里提出一种在完全保留原始图片信息的基础上,又同时使用图片显著性信息来帮助 RetinaNet 网络进行训练的方法。

2.3.1 节和 2.3.2 节中,神经网络仅使用 DARK FACE 训练集图片或是进行过图像增强后的训练集图片作为训练样本。新的做法是将图像的显著图 and 原图一起加入训练,修改 RetinaNet 的网络入口,原始的网络输入是一副图片,也就是 RGB 三通道,由于显著图和原始图像的尺寸是一样的,在 DARK FACE 数据集中都是表现为 720×1080 的尺寸。这里将显著图作为第 4 个通道,修改网络的输入,输入信息由图片尺寸 $\times 3$ 的原图像素信息改为图片尺寸 $\times 4$,前 3 个通道为原图片的 RGB 三通道像素,第 4 个通道为显著图表示的图像各位置显著性权值,将图像原始信息和显著性信息融合送入神经网络训练如图 9 所示。

本文所使用的 RetinaNet 的主干网络是 ResNet101,网络结构的入口接受三通道的输入数据,然后进行卷积操作,改为四通道。修改好 ResNet 的入口后,继续加载预训练的 WIDER FACE 模型再继续训练。由于预训练的 WIDER FACE 模型的入口是原始的三通道 ResNet,所以加载模型时需要忽略网络的第一层,只将网络第一层之后保存的权重读入。只忽略网络的一层,保存了网络其他层的权重信息,模型依然具有之前训练好的人脸检测能力。

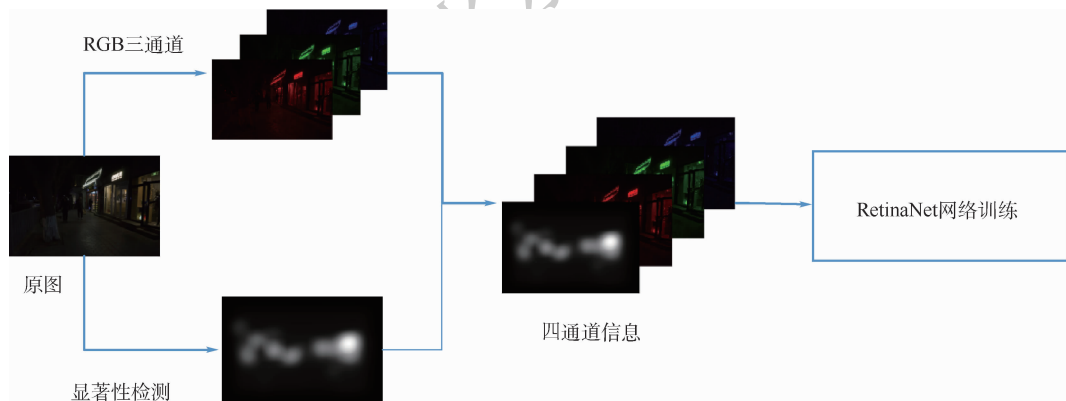


图 9 原图融合显著图四通道训练示意图

Fig.9 Schematic of training using 4 channel input fused by original images and saliency map

2.5 多尺度检测结果集成

RetinaNet 模型训练完成后,在检测人脸时采用多个尺度的图像分别测试。对 DARK FACE 测试集进行测试时,将测试图片进行多个尺度的缩放,本文实验时选取的缩放尺度包括:1,1.5,2,2.5。

同时对每个尺度的图片都进行左右翻转,分别进行测试,最后将得到的所有结果集成在一起作为最终结果。

集成多尺度检测结果时,首先将各尺度检测结果中的选择框坐标根据缩放比例以及是否左右翻转,重新计算归一到尺度为 1 且未翻转的标准。然后采用非极大值抑制(NMS)方法,在多个选择框中筛选出最终结果。检测结果中的多个选择框中有位置存在重叠的,需要去掉多余的选择框,NMS 方法分别在图像局部选出置信度最大的框,然后去掉和这个框的面积交并比(IOU)大于一定阈值的框,实验中的交并比阈值设置为 0.5。

3 实验评估

3.1 实验平台

本文实验所用的硬件平台为:4 块 GeForce RTX 2080 Ti GPU, Intel (R) Xeon (R) CPU E5-2680 v4 @ 2.40 GHz。软件环境为:操作系统为 ubuntu 18.04, NVIDIA 驱动版本 440.4, Cuda 版本 10.1.243, Python 版本 3.6.9, Pytorch 版本 1.3.1。

3.2 性能分析

3.2.1 评价标准

DARK FACE 数据集提供了一个用于评价检测结果与真实标签接近程度的工具。实现将检测结果中的 box 位置和标签对比,将交并比大于 0.5 的结果判定为正类,反之判为负类。

衡量目标检测结果的评价指标一般包括召回率(Recall)、查准率(Precision)、精度(Average Precision, AP)。这几个指标的计算方法如下:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

TP 为将正类预测为正类数, FN 为将正类预测为负类数, FP 为将负类预测为正类数, TN 为将负类预测为负类数。

对于同一份检测结果,设置不同的置信度阈值,其中判定为正类的样本数量不同,会得到不同的查准率和召回率,查准率越高,召回率越低。将不同查准率和对应召回率以二维坐标的形式表示,可以得到 AP,它是查准率-召回率曲线所围成

区域的面积,图 10 为本文不同实验结果的查准率-召回率曲线。

DARK FACE 数据集提供的测试工具采用 AP 来评价结果,综合考虑了召回率和查准率的影响,具体的计算方法为:假设检测结果中有 N 个 box,每个 box 都有一个置信度 score,其中有 M 个是和真实标签的交并比大于 0.5 的正例,将所有 N 个 box 根据 score 大小降序排列。取包括前 m 个正例(m 从 1 递加到 M)的 n 个 box,会得到 M 个召回率值($1/M, 2/M, \dots, M/M$),对于每个召回率值,可以计算出对应的查准率(m/n),然后对这 M 个查准率值取平均,即得到最后的 AP 值。

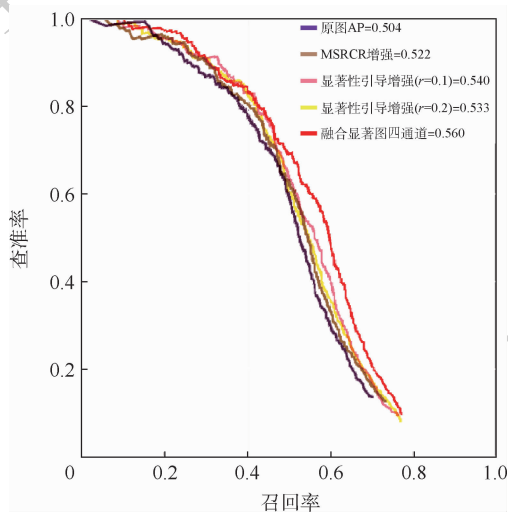


图 10 查准率-召回率曲线

Fig. 10 Precision-recall curves

3.2.2 实验结果

为了全面地对本文的人脸检测算法进行性能评估,对比了另外 2 组基于 DARK FACE 数据集的基准算法检测结果。

第 1 组检测结果是在 CVPR 2019 UG2 + 比赛前,官方发布数据库的同时,一起发布的一个 DARK FACE 基准检测精度^[9]。包括了 4 种人脸检测算法在 DARK FACE 上的检测精度: DS-FD^[5], PyramidBox^[6], SSH^[31], Faster R-CNN^[3]。

分别用这 4 种算法在 WIDER FACE 上训练神经网络,然后对未经图像增强的原始图像镜像检测得到精度。

同时, DARK FACE 发布的基准还使用了 7 种用于图像增强的算法: BIMEF^[32], Dehazing^[33], LIME^[17], MF^[34], MSRCR, JED^[35], RetinexNet^[36]。

分别使用这 7 种图像增强算法对数据集进行预处理,然后将 DSFD 和 PyramidBox 2 种人脸检测算法分别应用于处理后的数据,得到的结果如表 1 和表 2 所示。

表 1 DARK FACE 数据集发布的检测精度
(使用图像增强算法)

Table 1 Detection accuracies published from DARK FACE dataset (with image enhancement)

检测算法	精度
DSFD + MF	0.414
DSFD + MSRCR	0.413
DSFD + LIME	0.403
DSFD + BIMEF	0.402
DSFD + Dehazing	0.365
DSFD + RetinexNet	0.332
DSFD + JED	0.179
PyramidBox + MF	0.263
PyramidBox + Dehazing	0.249
PyramidBox + LIME	0.248
PyramidBox + MSRCR	0.246
PyramidBox + BIMEF	0.245
PyramidBox + RetinexNet	0.207
PyramidBox + JED	0.146

表 2 DARK FACE 数据集发布的检测精度
(未使用图像增强算法)

Table 2 Detection accuracies published from DARK FACE dataset (without image enhancement)

检测算法	精度
DSFD	0.153
Faster R-CNN	0.017
PyramidBox	0.132
SSH	0.076

第 2 组检测结果是在 CVPR 2019 UG2 + 比赛后,官方公布的一组基准检测精度。其中包括了 4 种基准算法的检测精度和 DSFD 以及 PyramidBox 2 种算法在分别使用了 7 种图像增强算法后的检测精度^[37],如表 3 和表 4 所示。

以上 2 组结果说明在大多数情况下,与未经图像增强算法的数据相比,使用了图像增强算法再训练得到的检测精度显著提高。比赛方总结指出与普通光照条件的数据集相比,检测器的整体性能仍然下降了很多,说明图像增强算法和人脸检测算法级联有很大的改善空间^[34]。本文分析以上图像增强算法可以帮助神经网络应对低光照环境下亮度不足、前后景难以区分、神经网络难以提取特征的问题。但以上图像增强算法对于前后景的区分度帮助还不足,并且可能损失原始图像的部分信息,所以效果不如本文提出的使用显著性引导的检测方式。

表 5 为不同模型在测试集上的相应检测精度,可以看到本文提出的显著性引导的人脸检测算法和目前其他人脸检测算法相比有更好的效果。其中 DARK FACE 原图融合显著图四通道训练的模型检测精度达到了 0.560,图 11 为该模型在测试集上检测得到的边框在原图上的可视化效果。

表 3 CVPR 2019 UG2 + 国际竞赛发布的基准
检测精度 (使用图像增强算法)

Table 3 Benchmark accuracies published from CVPR 2019 UG2 + competition (with image enhancement)

检测算法	精度
DSFD + MF	0.393
DSFD + MSRCR	0.393
DSFD + BIMEF	0.383
DSFD + LIME	0.383
DSFD + Dehazing	0.348
DSFD + RetinexNet	0.316
DSFD + JED	0.170
PyramidBox + MF	0.251
PyramidBox + Dehazing	0.237
PyramidBox + LIME	0.237
PyramidBox + MSRCR	0.235
PyramidBox + BIMEF	0.234
PyramidBox + RetinexNet	0.199
PyramidBox + JED	0.138

表 4 CVPR 2019 UG2 + 国际竞赛发布的基准
检测精度 (未使用图像增强算法)

Table 4 Benchmark accuracies published from CVPR 2019 UG2 + competition (without image enhancement)

检测算法	精度
DSFD	0.136
Faster R-CNN	0.125
PyramidBox	0.069
SSH	0.017

表 5 本文实验模型检测精度
Table 5 Detection accuracy of experimental models from this paper

训练方法	精度
DARK FACE 数据集原图训练	0.504
MSRCR 增强 DARK FACE 训练	0.522
显著性引导增强 DARK FACE 训练 ($r=0.1$)	0.540
显著性引导增强 DARK FACE 训练 ($r=0.2$)	0.533
DARK FACE 原图融合显著图四通道训练	0.560



图 11 检测结果可视化

Fig. 11 Visualization of detection results

4 结 论

根据实验结果,对比不同方案在 DARK FACE 数据集上得到的检测精度可知:

1) 将原始图像乘上图像的显著图和亮度调节系数,对图像进行增强,再由 RetinaNet 训练的模型,比使用原始图像训练出的模型检测精度更高,也比使用 MSRCR 算法增强的图像训练出的模型检测精度更高。

2) 使用图像的显著图作为神经网络输入的第 4 个通道,与图像的 RGB 三通道融合一起训练,得到的模型比以上模型检测精度都更高。并且超过了目前 DARK FACE 官方发布的基准检测精度和 CVPR 2019 UG2 + 比赛发布的基准检测精度。以上实验结果验证了图像的显著性特征对于神经网络检测低光照下的人脸能起到帮助作用。根据实验结果分析,低光照下的人脸检测主要挑战在于:由于亮度过低导致的前后景难以区分,进而影响神经网络提取和学习图像特征,通过图像显著性引导,给予图像前后景不同的权重,将显著性作为输入神经网络的特征可以帮助神经网络更好地检测人脸。另外本文采用了通用人脸数据集 WIDER FACE 做预训练,学习到的人脸特征对低光照人脸检测有帮助作用。

基于本文工作可继续深入的点包括:①尝试其他图像显著性检测算法,尤其是深度学习的图像检测性算法与人脸检测算法的结合;②将该方案扩展到通用目标检测领域,在通用目标检测上发挥图像显著性的作用。

参考文献 (References)

[1] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the 2014 IEEE Conference on Com-

puter Vision and Pattern Recognition. Piscataway: IEEE Press, 2014:580-587.

[2] GIRSHICK R. Fast RCNN[C]// Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press,2015:10-15.

[3] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:Towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(6):1137-1147.

[4] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:Unified,real-time object detection[C]. In Proceedings of the IEEE Computer Vision & Pattern Recognition. Piscataway:IEEE Press,2016:779-788.

[5] LI J,WANG Y,WANG C,et al. DSFD: Dual shot face detector [C] // Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2020:5060-5069.

[6] TANG X,DU D K,HE Z Q,et al. PyramidBox: A context-assisted single shot face detector [EB/OL]. (2018-08-17) [2020-09-12]. <https://arxiv.org/abs/1803.07737?context=cs>.

[7] JAIN V,LEARNED-MILLER E G. Fddb: A benchmark for face detection in unconstrained settings [EB/OL]. [2020-09-12]. http://works.bepress.com/erik_learned_miller/55/.

[8] YANG S,LUO P,LOY C C,et al. WIDER FACE: A face detection benchmark [C] // Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2016:5525-5533.

[9] YE Y,YANG W H,REN W Q,et al. UG²⁺ Track 2: A collective benchmark effort for evaluating and advancing [EB/OL]. [2020-03-31]. <http://arxiv.org/abs/1904.04474v4>.

[10] JOBSON D J,RAHMAN Z,WOODELL G. 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.

[11] LIU W,ANGUELOV D,ERHAN D,et al. SSD: Single shot multibox detector [C] // European Conference on Computer Vision. Berlin: Springer,2016:21-37.

[12] LI Z H,TANG X,HAN J Y,et al. PyramidBox + + : High performance detector for finding tiny face [EB/OL]. [2020-09-12]. <https://arxiv.org/abs/1904.00386>.

- [13] LIU S, HUANG D, WANG Y H. Receptive field block net for accurate and fast object detection [C] // European Conference on Computer Vision. Berlin: Springer, 2018: 404-419.
- [14] LAND E H. The Retinex theory of color vision [J]. Scientific American, 1977, 237(6): 108.
- [15] JOBSON D J, RAHMAN Z, WOODDELL G A. Properties and performance of a center/surround Retinex [J]. IEEE Transactions on Image Processing, 1997, 6(3): 451-462.
- [16] RAHMAN Z U, JOBSON D J, WOODDELL G A. Multi-scale Retinex for color image enhancement [C] // Proceedings of 3rd International Conference on Image Processing. Piscataway: IEEE Press, 1996: 1003-1006.
- [17] 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.
- [18] YUN Z, MUBARAK S. Visual attention detection in video sequences using spatiotemporal cues [C] // Proceedings of the 14th ACM International Conference on Multimedia. New York: ACM Press, 2006: 815-824.
- [19] CHENG M M, ZHANG G X, MITRA N J. Global contrast based salient region detection [C] // Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2011: 569-582.
- [20] ACHANTA R, ESTRADA F, WILSON P, et al. Salient region detection and segmentation [C] // Proceedings of the 6th International Conference on Computer Vision Systems. Berlin: Springer, 2008.
- [21] 张冬明, 靳国庆, 代锋, 等. 基于深度融合的显著性目标检测算法 [J]. 计算机学报, 2019, 42(9): 2076-2086.
- ZHANG D M, JIN G Q, DAI F, et al. Salient object detection based on deep fusion of hand-crafted features [J]. Chinese Journal of Computers, 2019, 42(9): 2076-2086 (in Chinese).
- [22] JIAN M W, LAM K M, DONG J Y, et al. Visual-patch-attention-aware saliency detection [J]. IEEE Transactions on Cybernetics, 2015, 45(8): 1575-1586.
- [23] TSUNG-YI L, PRIYA G, ROSS B G, et al. Focal loss for dense object detection [C] // Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 2999-3007.
- [24] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition [C] // Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
- [25] LIN T Y, DOLLA P, CIRSHICK R, et al. Feature pyramid networks for object detection [C] // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 936-944.
- [26] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context [C] // European Conference on Computer Vision. Berlin: Springer, 2014: 740-755.
- [27] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger [C] // Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 6517-6525.
- [28] FU C Y, LIU W A, RANGA A, et al. DSSD: Deconvolutional single shot detector [EB/OL]. [2019-09-12]. <http://arxiv.org/abs/170.06659>.
- [29] FAEN Z, FAN X Y, AI G, et al. Accurate face detection for high performance [EB/OL]. (2019-05-24) [2020-09-12]. <http://arxiv.org/abs/1905.01585>.
- [30] HOU X D, HAREL J, KOCH C. Image signature: Highlighting sparse salient regions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(1): 194-201.
- [31] NAJIBI M, SAMANGUEI P, CHELLAPPA R, et al. SSH: Single stage headless face detector [C] // Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 4875-4884.
- [32] YING Z Q, LI G, GAO W. A bio-inspired multi-exposure fusion framework for low-light image enhancement [EB/OL]. [2020-09-12]. <https://arxiv.org/abs/1711.00591>.
- [33] DONG X, WANG G, PANG Y, et al. Fast efficient algorithm for enhancement of low lighting video [C] // Proceedings of the 2011 IEEE International Conference on Multimedia and Expo. Piscataway: IEEE Press, 2011: 1-6.
- [34] FU X Y, ZENG D, HUANG Y H, et al. A fusion-based enhancing method for weakly illuminated images [J]. Signal Processing, 2016, 129: 82-96.
- [35] REN X D, LI M D, CHENG W H, et al. Joint enhancement and denoising method via sequential decomposition [C] // 2018 IEEE International Symposium on Circuits and Systems (ISCAS). Piscataway: IEEE Press, 2018: 1-5.
- [36] WEI C, WANG W, YANG W, et al. Deep Retinex decomposition for low-light enhancement [C] // Proceedings of British Machine Vision Conference, 2018: 155.
- [37] YANG W H, YUAN Y, REN W Q, et al. Advancing image understanding in poor visibility environments: A collective benchmark study [J]. IEEE Transactions on Processing, 2020, 29: 5737-5752.

作者简介:

李可夫 男, 硕士研究生。主要研究方向: 人工智能、计算机视觉。

钟汇才 男, 博士, 研究员。主要研究方向: 人工智能。

高兴宇 男, 博士, 研究员。主要研究方向: 人工智能。

Saliency guided low-light face detection

LI Kefu¹, ZHONG Huicai^{2,3}, GAO Xingyu^{2,*}, WENG Chaoqun², CHEN Zhenyu^{2,4},

LI Yongzhou², WANG Shizheng^{2,3}

- (1. School of Microelectronics, University of Chinese Academy of Sciences, Beijing 100049, China;
- 2. Institute of Microelectronics of Chinese Academy of Sciences, Beijing 100029, China;
- 3. Chinese Academy of Sciences R & D Center for Internet of Things, Wuxi 214000, China;
- 4. Big Data Center of State Grid Corporation of China, Beijing 100031, China)

Abstract: To deal with the problem that it is hard for convolution neural network to do face detection in low light environment, we propose a method combining image saliency and deep learning and apply it to low-light face detection, which integrates saliency and the original RGB channels of the image into neural network training. Sufficient experiments are implemented on DARK FACE, a low-light face dataset, and the results show that the proposed low-light face detection method achieves better detection accuracy than the existing mainstream face detection algorithms on DARK FACE, thus confirming the validity of the proposed method.

Keywords: face detection; saliency guided; deep neural network; low light; computer vision

Received: 2020-08-27; **Accepted:** 2020-10-21; **Published online:** 2020-11-05 13:29
URL: kns.cnki.net/kcms/detail/11.2625.V.20201104.1356.004.html
Foundation items: National Natural Science Foundation of China (61702491,61802390); Beijing Natural Science Foundation (4194095); Maoming Science and Technology Plan (2020028)
*** Corresponding author.** E-mail: gaoxingyu@ime.ac.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0459

基于对抗和迁移学习的灾害天气卫星云图分类

张敏靖¹, 白琮^{1,2,*}, 张敬林³, 郑建炜^{1,2}

1. 工业大学 计算机科学与技术学院, 杭州 310023; 2. 浙江省可视媒体智能处理技术研究重点实验室, 杭州 310023;

3. 南京信息工程大学 大气科学学院, 南京 210044)

摘 要: 针对卫星云图中的灾害天气数据存在严重不平衡问题, 提出一个结合生成对抗学习(GAN)和迁移学习(TL)的卷积神经网络(CNN)框架以解决上述问题进而提高基于卫星云图的灾害天气分类精度。该框架主要包含基于GAN的数据均衡化模块和基于迁移学习的CNN分类模块。上述2个模块分别从数据和算法层面解决数据的类间不平衡问题, 分别得到一个相对均衡的数据集和一个可在不同类别数据上提取相对均衡特征的分类模型, 最终实现对卫星云图的分类, 提高其中灾害天气的卫星云图类别分类准确率。与此同时所提方法在自建的大规模卫星云图数据上进行了测试, 消融性和综合实验结果证明了所提数据均衡方法和迁移学习方法是有效的, 且所提框架模型对各个灾害天气类别的分类精度都有显著提升。

关键词: 灾害天气; 卫星影像; 图片分类; 生成对抗网络; 类间不平衡

中图分类号: TB553

文献标志码: A **文章编号:** 1001-5965(2021)03-0585-11

全球75%经济损失源于灾害天气, 每年约1万多人因恶劣天气而死亡^[1-2]。灾害天气, 包括台风、强对流和沙暴, 严重威胁人民生命财产安全, 监测灾害天气的形成发展过程是气象灾害预测预报的基础。通过观测卫星云图进行监测是重要的手段之一, 因为地球的大部分地区被云覆盖, 各种天气现象总是和云有着密不可分的联系。卫星云图是由气象卫星自顶而下观测云层覆盖和地球表面的图像, 可以用来识别不同的天气状态, 评估其强度和未来发展趋势等, 为天气预报和灾害天气预测提供全天候的依据。本文聚焦于卫星云图中的灾害天气分类问题, 即在卫星云图数据中分类出带有热带气旋、温带气旋等可能带有灾害天气的云图。但是在实际的卫星云图中, 往往是非灾害天气类别的图片占据了原始数据的大多

数, 而各个灾害天气的数量相对较少, 数据呈现了不平衡的分布形态。这样的数据分布使得分类器在进行训练的时候, 会比较关注占据数据大多数的非灾害天气样本, 故而虽然总体的分类精度高, 但是如热带气旋、温带气旋等这些对于实际研究非常具有指导意义的类别, 并没有从分类器中得到很好的区分。因此需解决卫星云图灾害天气分类中类间不均衡的问题, 才能较好的将各个灾害天气从非灾害天气中区分出来。

图片数据的类间不平衡问题, 近年来一直是一个研究的热点^[3]。图片数据的类间不平衡是指在分类问题中不同类别的训练样例数目差别很大的情况。这一情况与实际生产生活中的数据分布情况相似, 非常具有研究的意义和必要性。2012年Krizhevsky等^[4]在ILSVRC-2012^[5]比赛中

收稿日期: 2020-08-25; 录用日期: 2020-09-04; 网络出版时间: 2020-10-12 16:04

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201012.1409.001.html

基金项目: 国家重点研发计划(2018YFE0126100); 国家自然科学基金(41775008, 61976192, 61702275); 浙江新苗人才计划(2020R403059)

* 通信作者: E-mail: congbaizj@zjut.edu.cn

引用格式: 张敏靖, 白琮, 张敬林, 等. 基于对抗和迁移学习的灾害天气卫星云图分类[J]. 北京航空航天大学学报, 2021, 47(3): 585-595. ZHANG M J, BAI C, ZHANG J L, et al. Classification of satellite cloud images of disaster weather based on adversarial and transfer learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 585-595 (in Chinese).

获得了冠军,成功的将深度学习应用于图片分类的问题上^[3,6-7],至此之后各类深度学习的框架模型开始涌现。但是研究者主要关注平衡数据分布的数据集,关于长尾分布数据的研究并未深入。尤其是在卫星云图的灾害天气分类问题的研究上,由于原始数据获取和处理的成本较大,相关的分类研究还较少。灾害天气分类问题中类间不平衡问题较为突出,故而本文对不平衡的卫星云图灾害天气分类问题展开研究。

本文针对不平衡卫星云图灾害天气分类,聚焦数据和算法混合的思路,提出了一种结合生成对抗学习(GAN)和迁移学习(TL)的分类训练模型框架,对云图数据分别进行过采样和欠采样处理,并采用迁移学习进行灾害天气云图的分类。通过在自建的 LSCIDWS-S 大尺度卫星云图数据集进行实验,证明了所提框架的有效性。本文的主要贡献如下:

1) 提出了一个 GAN + TL 的训练模型框架。该框架是针对不平衡卫星云图灾害天气所设计的分类框架,主要由数据平衡化处理模块和图片分类 2 个模块组成。

2) 在该框架中,GAN 用于高质量的图片生成,代替传统的简单复制的过采样方法,同时结合了欠采样,对原始不平衡的数据分布进行了均衡化的处理。在卷积神经网络(CNN)分类训练过程中引入了迁移学习的方法,使得整体的分类性能在原有的基础上得到了进一步的提升。

3) 实验结果表明,分类器整体的分类性能得到了一定的均衡。这对于实际研究有一定的借鉴意义,即更受到实际应用所关注但容易被分类器忽略的灾害天气样本的正判率得到了一定的提升。

1 相关工作

1.1 生成对抗网络

生成对抗网络(Generative Adversarial Networks, GANs)是在 2014 年,由 Goodfellow 等^[8]提出。GANs 的基本思想是源于博弈论中的零和游戏。它的网络结构由一个生成器(Generator)和一个判别器(Discriminator)组成,生成器的作用是为了尽可能地去学习数据的真实分布情况从而生成数据,而判别器的作用是判断输入的数据是来源于真实的数据还是由生成器生成的,二者之间不断的进行优化从而达到相对平衡。根据生成对抗网络有生成样本的这一特点,DCGAN(Deep

Convolutional GAN)^[9]为首个将 CNN 与 GAN 相结合以生成相应的图片样本,但是生成的图片质量不高并且不稳定。2020 年,NVIDIA 研究人员发布了 StyleGAN2^[10],该网络设计了具有非常规的生成器架构,从而可以生成高质量图片,并且训练过程较为稳定。本文所提方法中的 GAN 的设计就引用了该网络结构,从而可以生成相对质量较高的过采样样本。

1.2 图片分类中的类间不平衡问题

关于图片数据类间不平衡的研究主要可以分为 3 层次:数据、算法和数据算法的两相结合^[3,11]。关于数据层面的研究主要是对原始不平衡的数据进行均衡化的处理,把不平衡的数据转化为相对平衡的数据再加入模型中展开之后的训练。Hensman 和 Masko^[11]提出了提升样本的解决思路,主要是对样本中数量较少的类别,对其进行简单的复制从而达到扩充样本数量的效果,该方法虽然简单但是性能提升有限。基于算法层面的改进,主要是对损失函数的重新设计以及学习方式的改进。Wang^[12]等提出了 MSFE(Mean Squared False Error Loss)函数,该损失函数可以很好的平衡大样本和小样本之间的关系,从而也可以达到较好的分类性能。数据与算法的结合则是一种数据和算法混合的方法,如 He^[13]等提出的 LMLE(Large Margin Local Embedding),该方法采用了 5 倍抽样法和 THL(Tripleheader Hinge Loss)这一损失函数。本文所提框架是结合了数据和算法混合的方法,在数据层面上进行了包括对原始数据欠采样和引入了 StyleGAN 过采样的数据均衡化处理,而在算法层面则引入了迁移学习的思想。

1.3 迁移学习

迁移学习是把源域的知识迁移到目标域的学习方法,可使得目标域能够取得更好的学习效果。在深度学习中,神经网络从一个任务中学习到的知识可以应用到另一个相关的独立任务当中。在类间不平衡问题的处理上,迁移学习可以对相对平衡的数据集中训练出的模型进行迁移学习,该模型较好的学习到少数样本的类别特征,因而取得了不错的效果,如 Lee 等^[14]提出了二阶段的训练方法;Kang 等^[15]提出了 CRT(Classifier Retraining)方法,该方法是使用类平衡采样的数据对分类器进行重新训练,故而本文中的分类模块也会引入迁移学习这一思想。结合卫星云图灾害数据的实际情况,本文中所采用的迁移学习的思路是将原始数据集训练处的结果迁移到均

衡化处理后的数据,该过程主要是为了在提升各个少样本的天气类别分类精度的基础上,尽可能的保留原始数据中非灾害天气这一类别的精度。

2 本文所提方法

本文提出了一个基于 GAN + TL 的卫星云图灾害天气分类的框架,如图 1 所示。该框架主要分为 2 部分,一个是数据均衡化处理模块,另外一个为图片分类模块。在数据平衡化处理模块中首先是对原始不均衡的数据分布进行处理,处理之后得到一个相对较为均衡的数据分布情况。数据处理的过程采用不同的手段,对多数据样本的类别进行欠采样,对少样本的数据进行过采样。具体来讲,过采样的方法是采用生成对抗网络,对数据样本进行扩充。而欠采样是将样本根据设定的阈值进行缩减。在图片分类模块,首先在原始数据分布的数据集进行训练,之后将训练出来的模型迁移学习到类别较为均衡分布的数据集上要

分类中所研究的大多数数据分布的情况,各个类别的数据量基本相近,而且也取得相对较好的分类性能。本文数据处理模块的主要目的是将原始数据集的分布趋向理想数据分布的方向进行改进,这样可以把数据不均衡的问题转化为数据均衡化处理的问题。数据均衡化的过程分为基于阈值 N 的欠采样和基于 GAN 的过采样,具体细节如下:

步骤 1 基于阈值欠采样。在均衡化处理的过程中,首先根据各个类别之间的数量关系,设置一个较为合理的阈值 N ,然后根据这个阈值,对样本数据量大于这个阈值的类别进行数据随机丢弃的处理。本文采用的是去掉样本数量最多类别的数量和样本数量最少类别的数量,然后取剩下类别的样本数量计算平均数的方法确定阈值。具体实现如式(1)所示, X_{total} 为数据集中包含的总数量; X_{max} 和 X_{min} 分别为类别数量最大和类别数量最小的数量, n 为数据集中的类别数量。确定好阈值之后,对于样本数量大于该阈值的类别,进行随机欠采样,使得类别的数量达到阈值为止,此时的数据分布如图 2(c)所示。

$$N = \frac{X_{total} - X_{max} - X_{min}}{n}$$
 (1)

步骤 2 基于 GAN 的过采样。在完成数据欠采样的操作之后,对相应的样本数量少的类别进行过采样。本文提出以 StyleGAN2 为基础网络基于 GAN 的过采样方法。即首先把 StyleGAN2 在进行过采样的类别数据上进行训练,之后用训练好的生成器生成相应类别的数据并加入到已经完成欠采样操作的数据集中,此时的数据分布如图 2(d)所示。图 3 为 GAN 设计的核心思想流程, $G(z)_{fake}$ 为随机化的初始噪音, G 为生成器,用来生成图片; D 为判别器,用于判别生成图片的真假; $Data_{real}$ 为本框架结构中要进行数据增强的部分; $real$ 和 $fake$ 表示经过判别器判断生成的数据为真还是假。

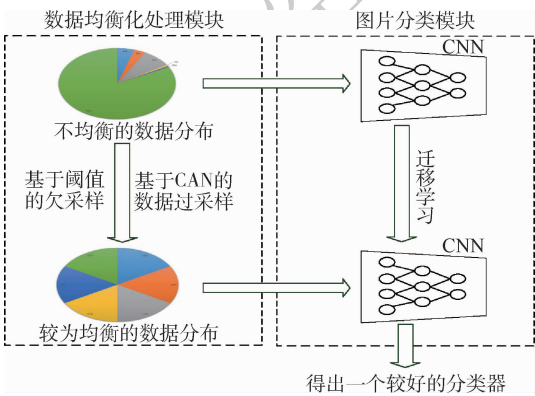


图 1 处理气象卫星中类间不平衡的模型框架图
Fig.1 Model framework for handling imbalance between classes in meteorological satellites

2.1 数据均衡化处理模块

数据均衡化处理过程的具体细节如图 2 所示,图 2(b)为理想数据分布,是现如今深度学习

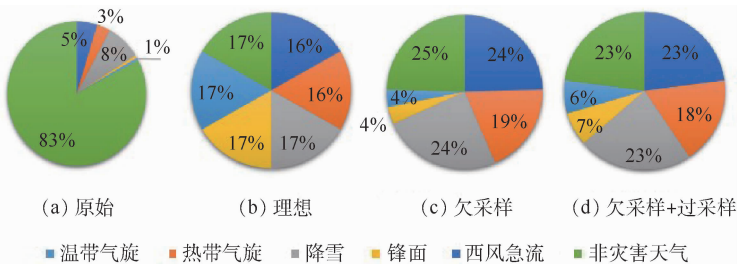


图 2 不同方法对应数据分布情况的百分比占比示意图

Fig.2 Schematic diagram of data percentage proportion of data distribution corresponding to different methods

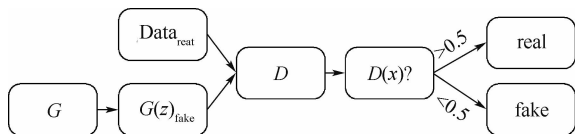


图 3 GAN 的核心思想

Fig. 3 Core idea of GAN

StyleGAN2 也是在该思想流程上的改进,可以生成更加高质量的图片。因为本文中所用的 LSCIDWS-S 数据集,原本就是高质量的卫星云图,故而本文选择了 StyleGAN2 作为 GAN 数据均衡化处理模块中的基础网络。StyleGAN2 主要在消除图片伪像上进行了进一步的改进,图片的伪像就是生成图片中图像上呈现出的类似于水滴的特征,该算法将改进的方向定位到了 AdaiN 的运算中,该算法的特点可分别归一化到每个特征图的均值和方差。具体的修改细节如下:首先对每个输入特征图的尺度根据调制卷积操作进行相应的调整,如式(2)所示, w 和 w' 分别为原始权重和调制权重; s_i 为与第 i 个输入特征图对应的比例; j 和 k 分别为特征图和卷积的空间下标。式(3)为完成相应调制卷积操作之后的输出权重(调制权重)的标准差。式(4)表示式(2)中 σ_j 固化到卷积权重中去, ε 为很小的数值,是为了确保被除数不为 0。

$$w'_{ijk} = s_i w_{ijk} \quad (2)$$

$$\sigma_j = \sqrt{\sum_{i,k} (w'_{ijk})^2} \quad (3)$$

$$w''_{ijk} = \frac{w'_{ijk}}{\sqrt{\sum_{i,k} (w'_{ijk})^2 + \varepsilon}} \quad (4)$$

3.2 图片分类模块

在完成数据均衡化处理之后,进入分类训练的模块,流程如图 4 所示。先根据原始不均衡的数据集训练出一个模型,再将不均衡分布数据训练出来的模型进行迁移学习,即把上述训练得到的模型权重初始化到较为均衡分布的数据集的分类模型上。采取这样二阶段训练的目的,主要是为了解决,均衡化后的数据分布所训练出的模型会丢失较多关于原样本数量较多类别特征信息的问题。故而能在牺牲样本数量较多类别的分类精度的前提下,提升其他各个类别的分类性能。同时二阶段训练的处理方法,也使得原始不均衡的数据分布和后处理的较为均衡的数据分布之间建立相应的关联。2 个模块之间的相互关联,使得整个数据处理和之后的图片分类过程形成一个闭环,也使得分类器的性能达到相应的稳定 and 平衡。对应算法步骤如下所示:

算法 1 图片分类框架算法。

输入:原始训练集 X_{train} ,均衡化处理后的训练集 X'_{train} ,模型训练的次数 m 。

输出:对测试集 X_{test} 的分类结果。

1. 随机初始化用于迁移学习的网络参数 M_i

2. Repeat

for $i = 1$ to m do

根据网络预测结果与真实的标签进行损失计算

反向传播更新网络的参数 W_i

3. 获得迁移学习的模型 M_i

4. 用模型 M_i 初始化分类模型 M_c 的参数

5. Repeat

for $i = 1$ to m do

根据网络预测结果与真实的标签进行损失计算

反向传播更新网络的参数 W_i

6. 完成最终分类模型 M_c 的训练

7. 将待预测的样本输入 M_c 获得最终分类结果

本文在图片分类模块中采用 ResNet101 作为训练过程中的基础模型。这主要是因为本文的数据集原始数量大,希望可以用深层次的网络取得较好的性能,但大量研究表明,随着网络深度的增

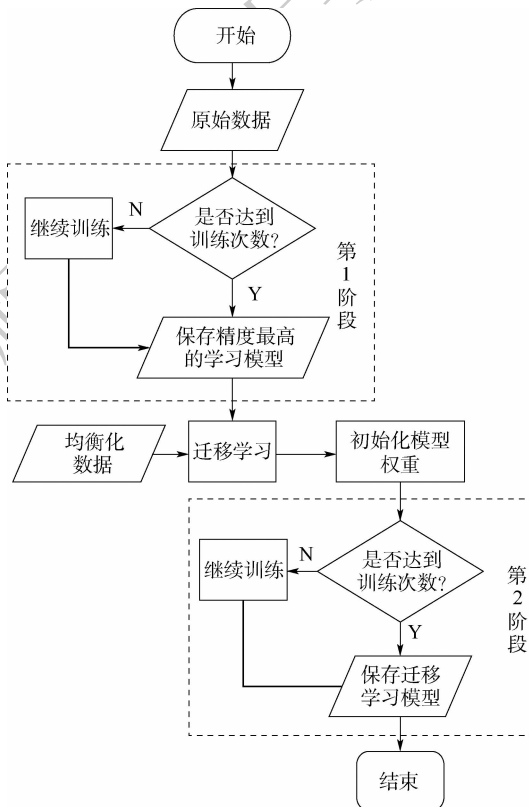


图 4 分类模块训练的流程图

Fig. 4 Flowchart of classification module training

加,会出现梯度爆炸,导致无法收敛这一问题。而残差思想的提出^[13]可以使得网络的性能不随网络深度的增加而退化,因此本文选择 ResNet101 作为分类模块中的基础模型。

以上 2 个模块的结合,构成了灾害天气卫星云图的分类框架,并充分考虑了数据中的不平衡问题。

3 实验结果及分析

3.1 数据集

因目前没有公开可用的云图数据库,本论文实验数据集采用自建的数据集,称之为 LSCIDMR-S (Large-scale Satellite Cloud Image Database for Meteorological Research System)。LSCIDMR-S 是以葵花-8 号气象卫星为数据来源建立的一个大尺度静止气象云图的单标签数据集。该数据集的数据采集时间跨度为 1 年,包含了温带气旋、热带气旋、锋面、西风急流、降雪、高冰云、低水云、海洋、沙漠、植被和其他总共 11 个类别总计 104 390 张图片,图片的原始大小为 1 000 × 1 000 像素。在本文中,因主要聚焦于灾害天气的分类识别,故将高冰云、低水云、海洋、沙漠、植被和其他合并为非灾害天气类别,图 5 为数据集中的部分云图示例,表 1 为数据集中各个类别的分布情况。重新划分后的数据不平衡比率 (Imbalanced Ratio, IR) 为 137.25。IR 为衡量数据集不均衡程度的一个指标,其具体计算如式(5)所示,是数据量最多的类别的数量和数据量最少的类别的数量比例,一般大于 10 认定为类间不平衡的数据集^[3,7]。 C_i 为

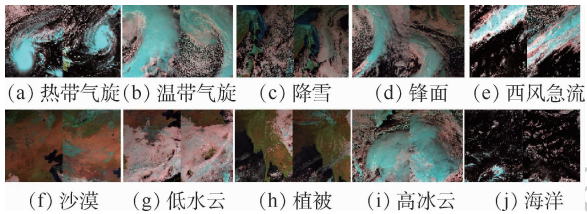


图 5 LSCIDMR-S 数据集的部分示意图

Fig. 5 Partial schematic of LSCIDMR-S dataset

表 1 LSCIDMR-S 处理之后的数据分布情况表

Table 1 Data distribution of LSCIDMR-S after processing

类别	数量	比例/%
温带气旋	4 985	4.78
热带气旋	3 305	3.17
降雪	8 522	8.16
锋面	634	0.61
西风急流	628	0.60
非灾害天气	86 315	82.69
总计	104 390	100

第 i 个类别对应的类别数量。

$$IR = \frac{\max_i \{ |C_i| \}}{\min_i \{ |C_i| \}} \quad (5)$$

3.2 评估方法

本文中的实验评估方法采用分类中通常使用的总体精度 (Overall Accuracy) 和各个类别的分类精度 (Category Accuracy) 进行评估。总体精度是指预测正确的标签数量和待预测的总标签数量的比例,这一指标只能笼统的评价模型的整体性能。对于长尾数据集的分布,单一的总体精度还不足以充分的体现这一模型与实际问题的贴合程度。单一的总体精度的虚高并不能很好地表示模型的性能很好,很有可能是因为数据集中占据绝大多数类别的单个类别的性能好。比如本文中的非灾害天气类别这一类别,占总数据集的 82.69%,如果总体精度达到了 80%,也很有可能只是单一的非灾害天气类别这个类别的精度高而已。而在实际的长尾分布数据集当中,占数据量少的类别往往更是应该关注的对象。故而占样本数量较少的温带气旋、热带气旋、西风急流、锋面和降雪,它们单个类别的分类精度对于实际问题的研究更加有意义,故而本文还采用各个类别的分类精度。

总体精度和单个类别的精度能从数值上说明一个模型的整体性能。于此同时本文还采用了 ROC 曲线作为评估指标^[9],该曲线可视化了正确分类的阳性样本与阴性样本之间的关系,故而 ROC 曲线是衡量模型在不均衡数据集中性能的一个重要指标。ROC 曲线通常用于二分类的研究,横坐标为假阳性 (特异度),纵坐标为真阳性 (灵敏度)。本文为将其扩展到多分类问题上,首先对输出进行二值化,然后分别进行如下操作:①对每个类别绘制了一个对应的 ROC 曲线;②Micro-average 通过把多分类问题转化为二元预测来绘制 ROC 曲线;③Macro-average 用于多分类的评估方法是对每个标签给予相同的权重,实现宏观的平均,最后将同一个类别的数据汇总到 1 张 ROC 曲线上。ROC 曲线下方与坐标轴围成的面积被定义为 AUC (Area Under Curve),表示预测的正例样本排在负例样本前面的概率,这个面积的数值通常介于 0.5 ~ 1 之间,数值越大,表明分类方法的性能越好。

3.3 参数设定

本文实验均在一个配备了 32 GB 内存和 3.6-GHz Inter (R) Core i9-9900K CPU 处理器及 GeForce RTX 2080Ti 显卡的工作站上进行。

对于数据集的训练集和测试集按照 9:1 的比

例进行划分。对于数据集欠采样的部分,按照数据均衡的方向调整,设置了阈值 $N = 3\ 826$,该阈值根据训练集中去掉了类别中数量的最大值和最小值取均值。然后对数据量超过这个数值的类别的数据进行随机丢弃直到数据数量达到 $3\ 826$ 。对于数据过采样部分,是基于数据欠采样的基础上对数据再进行进一步的处理,对于类别数量较少的锋面和西风急流扩充 k 倍,默认为 1 ,相应的数据数量分别为 571 和 766 。对于参数,按照 0.5 的步长设置进行了相应的参数实验。

使用 StyleGAN2 生成的图片大小为 256×256 像素。分类模型中的各个超参数分别设置为 $\text{learning rate} = 0.001$, $\text{momentum} = 0.9$, $\text{batch_size} = 64$,图片统一为 256 像素 $\times 256$ 像素。每个模型都训练 20 次,保留总精度最高的模型,进行指标计算。

3.4 实验结果及分析

针对本文所提 GAN + TL 框架,设计对比实验证明所提模型方法有效。分别为:采用未经过任何处理的原始数据进行训练与分类的 Base 方法;对原始数据按照对超过阈值的类别进行随机欠采样处理后的 Base_under 方法;在 Base_under 的基础上对 Base 进行迁移学习的 Base_under_t 方法;Base_under_over 是对原始数据集进行按阈值随机欠采样之后的基础上,再对原始数据集中样本数量较少的类别进行机械复制的过采样方法;Base_under_over_t 是在 Base_under_over 方法的基础上对 Base 进行迁移学习的结果;之后的 Base_under_gan 相比于 Base_under_over 是用生成对抗网络来代替传统的复制对数据进行过采样,从而使得数据分布达到一个较为均衡的状态;最后的 Base_under_gan_t 也是本文所提的 GAN +

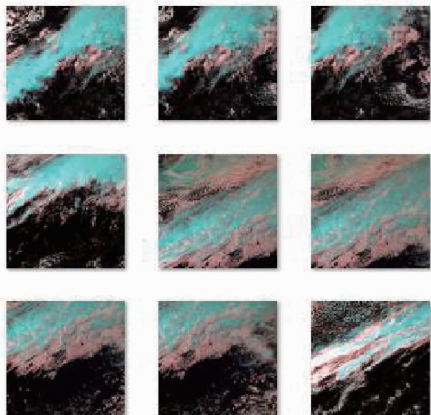
TL 框架,即在 Base_under_gan 的数据处理基础上对 Base 训练出的模型进行迁移学习。接下来的实验分析中,也将主要从数据平衡化模块和图片分类模块分别展开分析。

首先是数据均衡化,表 2 为不同方法所对应的数据分布情况和相应的数据不平衡系数。可知,原始数据集的数据不平衡系数达到 137.25 。而经过本文所提数据均衡化处理之后,数据不平衡系数降到了 3.35 。图 2 为不同方法对应的数据分布情况的百分比占比示意图,图 2(a)为数据的原始分布示意图,图 2(b)为大多数研究中数据理想的均衡分布图,图 2(c)和图 2(d)分别对应了不同数据均衡化处理的过程。图 6 为基于 GAN 的数据过采样的部分结果示意图,以西风急流为例,图 6(a)为原始数据集中西风急流的部分示意图,图 6(b)为基于 GAN 生成的数据样例,从图片中可以观测到 StyleGAN2,可以较好地学习到图像的轮廓、纹理、颜色等特征。虽然伴随着一定的噪音,但是从表 3 的实验结果中可以观察到,StyleGAN2 生成的图片信息能够较好地学习原图像的特征。

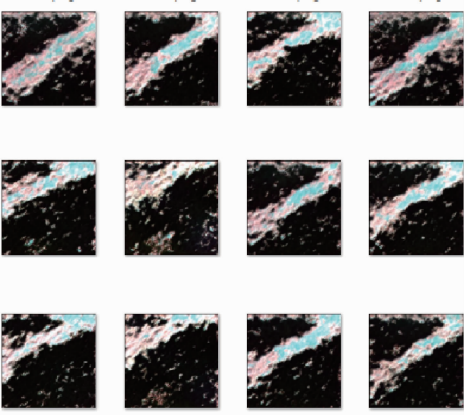
表 2 各方法对应的数据分布及数据不平衡系数

Table 2 Data distribution and data imbalance degree corresponding to each method

方法	数据集的分布(对应图 2)	IR
Base	1	137.25
Base_under	2	6.759
Base_under_t	2	6.759
Base_under_over	3	3.35
Base_under_over_t	3	3.35
Base_under_gan	3	3.35
Base_under_gan_t	3	3.35



(a) 原始数据集中西风急流的部分示意图



(b) 基于GAN生成的图片示例

图 6 基于 GAN 的数据过采样生成的图片

Fig.6 Schematic diagram of oversampling data image generated by GAN

表 3 为分类实验的总体精度和各个类别的分类精度。可知,本文所提的数据处理:欠采样、基于 StyleGAN2 的过采样方法和基于不平衡样本的迁移学习(Base_Under_Over_Gant)的方法对解决类间不平衡问题有效。由表 3 可知,虽然 Base 方法的整体精度和非灾害天气这一类别的分类性能达到了最优,但是对于西风急流、热带气旋、锋面和温带气旋这 4 个类别的数据,他们的分类精度还非常低。这 4 个类别的原始数据量较少,但是能准确地识别它们对于实际应用场景非常有意义。上述实验结果进一步说明了长尾数据的分布对于 CNN 的特征提取有一定的影响,在分类的时候会更加关注数量多的类别,因而数量多的类别(非灾害天气)能取得较好的提取特征,进而忽略了其他数量较少类别的特征的学习,由此对数据量较大的类别(非灾害天气)进行处理就非常有必要。对于阈值大于 $N = 3\ 826$ 的类别进行了随机丢弃的欠采样处理,从表 3 中可以看出,Base_under 与最开始的 Base 相比,虽然损失了非灾害天气这一类别的精度,但是其他各个类别的精度都得到了一定的提升,与此同时,Base 方法中对数量较多的非灾害天气这一类别的特征有较好的学习,故而把 Base 方法训练的结果迁移学习到进行调整的 Base_under 模型,得到了新的模型 Base_under_t,发现部分类别的精度会得到一定的提升。故而提升少数类别的数量,可以帮助 CNN 均衡的提取各个类别的特征。将欠采样和过采样相结合的同时,再加上迁移学习,这给训练一个更好的分类器提供了思路。用 GAN 对于少数样本的数据进行过采样,从表 3 中可以看出,Base_under_gan 和 Base_under_over 2 个模型相比,Base_under_gan 的总精度和绝大部分类别的分类精度基本高于 Base_under_over。这说明用 GAN 生成图像的过采样方法比简单的复制粘贴过采样的方法能取得更好的分类效果。最后,本文所提出的 Base_

over_gan_t 模型基本在所有类别都取得了相对较高的精度。在降雪类别的数据上虽然没有取得最高的分类精度,但是也取得了相对不错的精度。究其原因是在对数据进行了欠采样和过采样处理之后的数据各个类别的分布比例虽然达到了一定的均衡,但是降雪(如图 2(d)所示)这一类别相较于其他类别的数量占比较大,故而分类器在该类别的特征提取上能够取得较优的性能。表 3 中数据部分加粗的是各个类别取得的最高精度。

图 7(a)~图 7(h)分别对应非灾害天气、西风急流、热带气旋、降雪、锋面、温带气旋、Micro-average 和 Macro-average 在各个不同模型下的 ROC 曲线图,图中的 Model 1~7 分别对应表 2 中的各个方法。表 3 中本文所提 Base_under_gan_t 方法,相比于其他的方法,整体的性能达到了最佳,对应的 ROC 曲线的顶角靠近左上角,对应的 AUC 与其他方法相比达到了最高:非灾害天气(0.83)、西风急流(0.85)、热带气旋(0.85)、降雪(0.94)、锋面(0.73)和温带气旋(0.83)。

之后对生成对抗网络进行数据扩充的倍数 k 对各个类别分类精度的影响进行了进一步的探究,实验结果如图 8 所示。图 8(a)为基于 GAN 的数据过采样之外不采用迁移学习进行模型训练的分类性能情况;图 8(b)为同时采用基于 GAN 的数据过采样和迁移学习进行模型训练之后的分类性能情况。 $k = 0.5、1、1.5、2、2.5、3、3.5、4、4.5、5、5.5$,训练数据集对应的 IR 分别为 4.5、3.35、2.70、2.25、1.93、1.69、1.50、1.35、1.23、1.13、1.04,相应的各类别的数据分布情况如图 9 所示。从图 8 中可以看出,无论是否采用迁移学习,随着值的增大,总精度基本稳定在 0.75。而其他各个数据类别的分类精度会呈现先升高,然后稳定在一定数值之后再下降的趋势,当值介于 1~2 之间的时候整体的分类器的性能都取得较为均衡

表 3 各个模型的总精度和分类精度的统计

Tabel 3 Statistics of total accuracy of each model and accuracy of each category (Accuracy)

序号	方法	总精度	分类精度					
			非灾害天气	西风急流	热带气旋	降雪	锋面	温带气旋
1	Base	0.889 3	0.966 6	0.209 7	0.215 2	0.821 6	0.095 2	0.317 3
2	Base_under	0.733 1	0.743 1	0.209 7	0.454 5	0.841 5	0.127 0	0.759 0
3	Base_under_t	0.761 8	0.765 4	0.338 7	0.648 5	0.940 1	0.190 5	0.594 4
4	Base_under_over	0.703 8	0.708 5	0.387 1	0.318 2	0.843 9	0.079 4	0.757 0
5	Base_under_over_t	0.765 1	0.771 3	0.612 9	0.336 4	0.882 6	0.396 8	0.722 9
6	Base_under_gan	0.732 0	0.716 4	0.629 0	0.687 9	0.946 0	0.333 0	0.728 9
7	本文	0.774 4	0.763 6	0.709 7	0.736 4	0.929 6	0.460 3	0.769 1

的结果,称之为最佳取值范围。总体上当处于最佳取值范围时,采用迁移学习之后的总精度和各个类别的分类精度都有一定的提升。而随着 k 的增加,IR 趋近于 1 的时候,对不均衡分布数据训练出来的模型进行迁移学习的分类器的性能提

升并没有明显效果。这主要是因为样本增加的数量远大于该类别原始数据的数据,通过 GAN 进行数据的过采样会导致分类器在特征学习过程中受到生成数据中噪声的影响,进而影响特征学习效果。

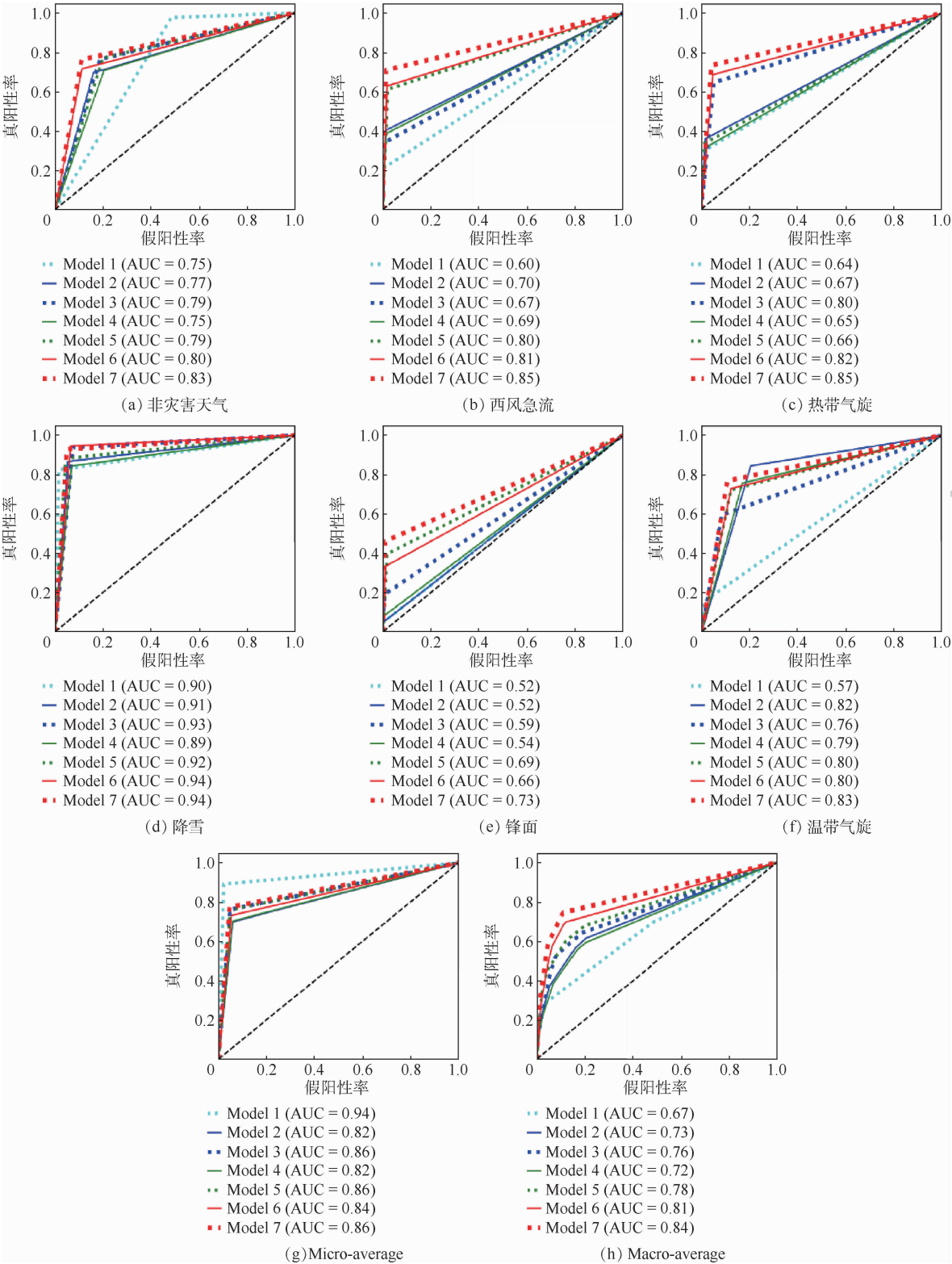


图 7 各个模型中各个类别对应的 ROC 曲线
Fig. 7 ROC curve corresponding to each category in each model

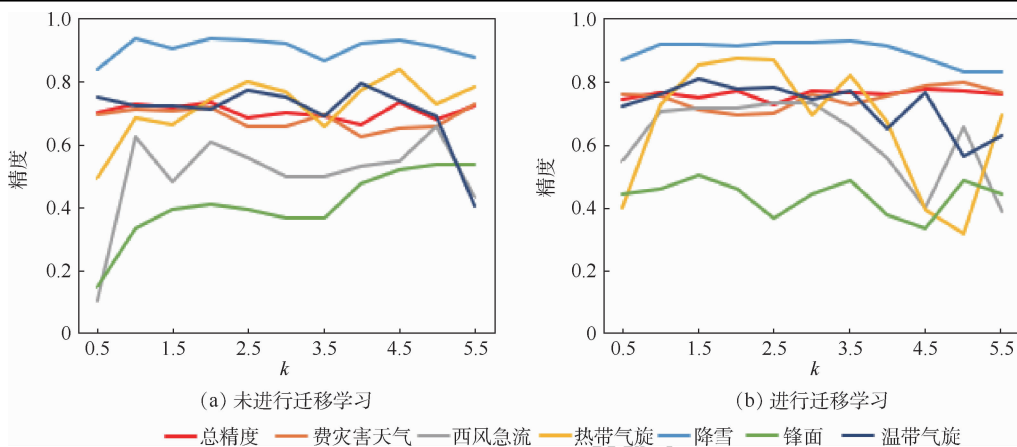


图 8 参数 k 对分类性能的影响

Fig. 8 Influence of parameter k on classification performance

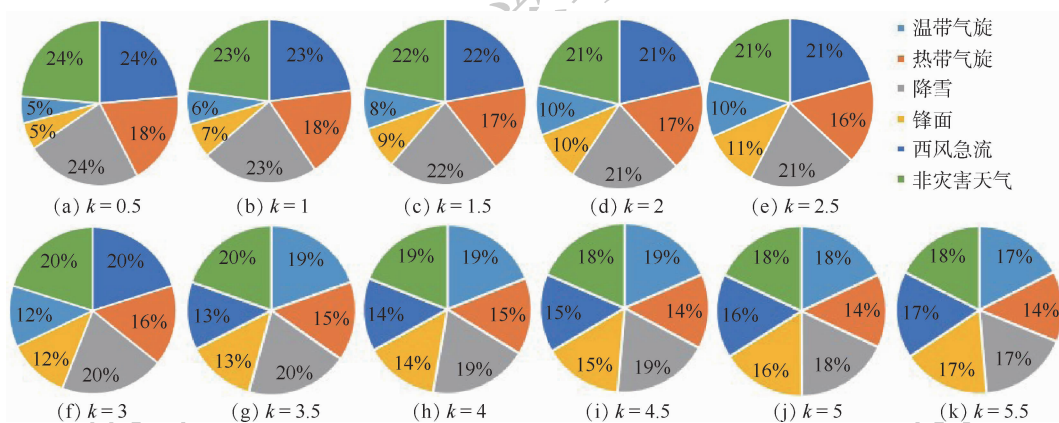


图 9 不同 k 对应的数据分布

Fig. 9 Data points corresponding to different k

4 结 论

1) 本文提出了一个结合生成对抗网络和迁移学习处理灾害天气气象卫星云图中的长尾数据分类的框架。该框架分为数据均衡化处理模块和数据分类 2 个模块。数据均衡化处理模块采用 GAN 对少样本的数据类别进行过采样,然后将过采样和欠采样相结合实现原始数据均衡化处理。

2) 通过上述过程,GAN 可根据数据的分布情况生成新的过采样数据,进而能够给 CNN 中的特征提取提供更加优质的图片信息;在图片分类模块中,采用对原始不均衡数据集训练得到的模型进行迁移学习,用所得的模型对图片进行分类的方法。

3) 在自建的大规模卫星云图数据上的多方面实验证明,所提框架中的基于 GAN 的数据过采样和基于迁移学习的模型训练方法可以较好地解决卫星云图中的数据不平衡问题。所提框架在传递数量较多的数据类别特征信息的同时又可对数量较少的数据类别提取较好的特征,故而在提升

少量样本类别分类精度的同时,也尽可能地保证大量样本的分类精度。为之后解决类间不平衡的长尾数据分布提供了一个可以借鉴的解决思路。与此同时,虽然其他各个类别分类的精度都得到了可观的提升,但是数据分类的总体精度和非灾害天气的分类精度有了一定的下降,其中非灾害天气精度下降可能是因为随机欠采样不能充分保留原始数据的多样性(原始灾害天气中并没有进行更加细致的类别划分,从而不能有计划地从各个非灾害天气类别中进行随机欠采样)。这也是之后研究中所需要进一步改进和研究的方向:即在保证各个少量样本类别分类精度得到提升的同时,分类的总体精度也要保证一定的提升。

参考文献 (References)

- [1] PARMESAN C, ROOT T L, WILLIG M R. Impacts of extreme weather and climate on terrestrial biota [J]. Bulletin of the American Meteorological Society, 2000, 81 (3): 443-450.
 - [2] 王会军. 把脉极端天气气候事件, 构建预测方法 [J]. 大气科学学报, 2019, 42 (1): 1.
- WANG H J. Take the pulse of extreme weather and climate

events and construct prediction methods[J]. Journal of Atmospheric Sciences,2019,42(1):1 (in Chinese).

[3] JOHNSON J M,KHOSHGOFTAAR T M. Survey on deep learning with class imbalance [J]. Journal of Big Data, 2019, 6 (1):27.

[4] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks [C] // Advances in Neural Information Processing Systems. 2012: 1097-1105.

[5] DENG J,DONG W,SOCHER R, et al. Imagenet: A large-scale hierarchical image database [C] // 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press,2009:248-255.

[6] SAIYEDA A,MIR M A. Cloud computing for deep learning analytics:A survey of current trends and challenges[J]. International Journal of Advanced Research in Computer Science, 2017,8(2):68-72.

[7] LEEVY J L,KHOSHGOFTAAR T M,BAUDER R A, et al. A survey on addressing high-class imbalance in big data[J]. Journal of Big Data,2018,5(1):42.

[8] GOODFELLOW I,POUGET-ABADIE J,MIRZA M, et al. Generative adversarial nets [C] // Advances in Neural Information Processing Systems. 2014:2672-2680.

[9] RADFORD A,METZ L,CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks [EB/OL]. (2016-01-07) [2020-07-23]. <https://arxiv.org/abs/151/06434v2>.

[10] KARRAS T,LAINE S,AITTALA M, et al. Analyzing and improving the image quality of stylegan [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2020:8110-8119.

[11] HENSMAN P,MASKO D. The impact of imbalanced training data for convolutional neural networks [EB/OL]. [2020-07-23]. <http://diva-portal.org/smash/get/diva2:811111/FULLTEXT01.pdf>.

[12] WANG S,LIU W,WU J, et al. Training deep neural networks on imbalanced data sets [C] // 2016 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE Press, 2016:4368-4374.

[13] HE K,ZHANG X,REN S, et al, Deep Residual Learning for Image Recognition [C] // Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway:IEEE Press,2016:770-778.

[14] LEE H,PARK M,KIM J. Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning [C] // 2016 IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE Press, 2016: 3713-3717.

[15] KANG B,XIE S,ROHRBACH M, et al. Decoupling representation and classifier for long-tailed recognition [EB/OL]. (2020-02-19) [2020-07-23]. <https://arxiv.org/abs/1910.09217?context=cs.CV>.

作者简介:

张敏靖 女,硕士研究生。主要研究方向:计算机视觉与多媒体信息处理。

白琮 男,博士,副教授,博士生导师。主要研究方向:计算机视觉与多媒体信息处理。

张敬林 男,博士,教授,博士生导师。主要研究方向:计算机视觉与遥感信息处理。

郑建炜 男,博士,副教授,博士生导师。主要研究方向:计算机视觉与遥感信息处理。

Classification of satellite cloud images of disaster weather based on adversarial and transfer learning

ZHANG Minjing¹, BAI Cong^{1,2,*}, ZHANG Jinglin³, ZHENG Jianwei^{1,2}

- (1. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China;
- 2. Key Laboratory of Visual Media Intelligent Processing Technology of Zhejiang Province, Hangzhou 310023, China;
- 3. School of Atmospheric Science, Nanjing University of Information Science and Technology, Nanjing 210044, China)

Abstract: Weather can be forecasted based on clouds. However, how to use deep learning technology to achieve automatic weather forecasting, especially the automatic recognition of disaster weather, is still an unexplored field. Hence, it is necessary to carry out research on the basic problem in the field of automatic identification: the classification of satellite cloud images. Satellite cloud images have serious data imbalance problems. That is, cloud image data related to severe weather accounts for a very small proportion of all cloud image data. Therefore, this paper proposes a framework combining Generative Adversarial Network (GAN) and Transfer Learning (TL) based Convolutional Neural Network (CNN) to solve the problem of low accuracy of disaster weather classification based on satellite cloud images. The framework is mainly divided into a data balancing module based on GAN and a CNN classification module based on transfer learning. The above two modules solve the data imbalance problem from the data and algorithm level respectively, and obtain a relatively balanced dataset and a classification model that can extract relatively balanced features on different types of data. Eventually, the classification of satellite cloud images is achieved and the accuracy of the classification of satellite cloud images in disaster weather is improved. The method proposed in this paper has been tested on self-built large-scale satellite cloud image data. The ablative properties and comprehensive experimental results prove that the proposed data balancing method and transfer learning method are effective, and the proposed framework model has significantly improved the classification accuracy of various disaster weather categories.

Keywords: disaster weather; satellite image; image classification; generative adversarial networks; data imbalance

Received: 2020-08-25; **Accepted:** 2020-09-04; **Published online:** 2020-10-12 16:04
URL: kns.cnki.net/kcms/detail/11.2625.V.20201012.1409.001.html
Foundation items: National Key R & D Program of China (2018YFE0126100); National Natural Science Foundation of China (41775008, 61976192, 61702275); Zhejiang Xinmiao Talent Projects (2020R403059)
*** Corresponding author.** E-mail: conghai@zjut.edu.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0470

基于多语义线索的跨模态视频检索算法



丁洛¹, 李逸凡¹, 于成龙², 刘洋¹, 王轩^{1,3}, 漆舒汉^{1,3,*}

(1. 哈尔滨工业大学(深圳) 计算机科学与技术学院, 深圳 518055;

2. 深圳信息职业技术学院 数字媒体学院, 深圳 518172; 3. 鹏城实验室, 深圳 518055)

摘 要: 针对现有的大多数跨模态视频检索算法忽略了数据中丰富的语义线索, 使得生成特征的表现能力较差的问题, 设计了一种基于多语义线索的跨模态视频检索模型, 该模型通过多头目自注意力机制捕捉视频模态内部对语义起到重要作用的数据帧, 有选择性地关注视频数据的重要信息, 获取数据的全局特征; 采用双向门控循环单元 (GRU) 捕捉多模态数据内部上下文之间的交互特征; 通过对局部数据之间的细微差别进行联合编码挖掘出视频和文本数据中的局部信息。通过数据的全局特征、上下文交互特征和局部特征构成多模态数据的多语义线索, 更好地挖掘数据中的语义信息, 进而提高检索效果。在此基础上, 提出了一种改进的三元组距离度量损失函数, 采用了基于相似性排序的困难负样本挖掘方法, 提升了跨模态特征的学习效果。在 MSR-VTT 数据集上的实验表明: 与当前最先进的方法比较, 所提算法在文本检索视频任务上提高了 11.1%; 在 MSVD 数据集上的实验表明: 与当前先进的方法比较, 所提算法在文本检索视频任务上总召回率提高了 5.0%。

关键词: 跨模态视频检索; 多语义线索; 多头目注意力机制; 距离度量损失函数; 多模态

中图分类号: TP391.4

文献标志码: A

文章编号: 1001-5965(2021)03-0596-09

随着多媒体数据的爆炸式增长, 传统的单模态检索已经难以满足人们在多媒体领域下的检索需求, 用户迫切希望能够利用其中一种模态数据作为查询对象来检索语义相似的另一种模态数据内容, 例如用图像检索文本、文本检索图像或视频等, 即跨模态检索。

跨模态检索需要同时处理不同模态的数据, 这些数据在语义上具有一定的相似性, 但是其底层特征是异构的, 难以直接计算它们的相似度, 即存在“语义鸿沟”的问题。基于联合嵌入空间的方法在大多数跨模态检索算法中被使用, 因为它能建立起不同“语义鸿沟”的桥梁^[1]。在这个联

合嵌入空间中, 多模态数据以点的形式表示, 两点的距离反映了对应的原始数据之间的语义相似性。

近年来, 基于联合嵌入空间方法的跨模态视频检索算法受到了人们的广泛关注^[2-5]。现有的检索算法大多与图像-文本检索算法^[6-7]在设计上非常相似, 这些算法一方面通过设计合理的损失函数, 使视频信息和文本信息能够映射到统一共享子空间中, 使不同模态的数据能够在同一度量空间中计算相似度。例如, Faghri 等^[7]设计了一种基于共享子空间学习的跨模态检索算法, 该算法采用了一种精心设计的三元组距离度量损失函

收稿日期: 2020-08-26; 录用日期: 2020-09-04; 网络出版时间: 2020-10-30 10:22

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201030.1006.001.html

基金项目: 国家自然科学基金 (61902093); 广东省自然科学基金 (2020A1515010652)

* 通信作者。E-mail: shuhanqi@cs.hitsz.edu.cn

引用格式: 丁洛, 李逸凡, 于成龙, 等. 基于多语义线索的跨模态视频检索算法[J]. 北京航空航天大学学报, 2021, 47(3): 596-604. DING L, LI Y F, YU C L, et al. Cross-modal video retrieval algorithm based on multi-semantic clues[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 596-604 (in Chinese).

数,改善了视觉-语义特征判别力不足的问题,提升了跨媒体检索的准确率。Torabi 等^[5]提出了一种非对称的相似性度量方法,该方法可以捕获视觉内容相同但描述细节不同的文本之间的相似度。另一方面,这些算法采用了图像-文本检索的方式,通过视频帧的均值池化特性,将视频特征转化为图像特征,然后在现有的图像-文本嵌入方法^[6]基础上进行简单的微调即可扩展为视频文本跨模态检索模型^[2-4]。然而,这些算法存在一定的局限性,一是仅仅考虑视频中的全局信息,而忽略了视频内部丰富的时空动态信息,导致模型无法区分视频帧之间的语义交互关系,二是在对视频帧进行池化过程中,每个视频帧对视频语义的重要程度都视作相同的,使得模型难以捕捉到视频中的重要信息,因此这些算法的检索性能难以令人满意。

为了解决以上问题,本文提出基于多语义线索的跨模态视频检索(Multi-semantic Clue Cross modal Video Retrieval, MCCVR)框架。在该框架中,同时提取了视频中的全局、局部和上下文信息,并对此3种信息进行有效融合,构成了关于视频内容的多语义线索,将其应用于跨模态检索中有效提升了特征的表现力。在提取全局信息的过程中,为了有效提取视频中重要的视频帧,采用一种自下而上的多头自注意力机制捕捉视频模态数据内部的细微交互,有选择性地关注视频内部的重要信息增强模型的表征能力。为了利用视频内容中一定时空范围内的动态信息,本文对视频帧相邻帧之间具有细微差别的局部数据进行了联合编码,从而更好地挖掘出视频和文本数据中的语义信息。不仅如此,本框架还采用了双向GRU以捕捉多模态数据内部上下文之间的交互信息。除此之外,针对传统的三元组距离度量损失函数中存在的正负样本不平衡问题,提出了一种基于相似性排序的困难负样本挖掘方法,有效提升了生成特征的判别力。本文的主要贡献如下:

1) 提出了一种基于多语义线索的跨模态视频检索框架,采用多级耦合的方式综合考察了视觉/文本的全局、局部和上下文信息,从而能够生成具有更强判别力的模态特征。

2) 提出了一种改进的三元组距离度量损失函数,采用了基于相似性排序的困难负样本挖掘方法,改善了正负样本不平衡问题,提升了跨模态特征学习的效果。

3) 在多项测试数据集中与现有的最好方法对比,总召回率取得了一定的提升,其中,在MSVD的

文本检索视频任务上,总召回率提升了11.1%。

1 相关工作

1.1 跨模态检索

跨模态检索是指在不同模态的数据中检索与提供的查询样本语义相关的结果。常见的跨模态检索问题有通过关键词检索图片,或者通过图片检索其相关描述等。跨模态检索问题的主要困难在于如何度量不同模态的数据之间的相似度。较为普遍的一种做法是将不同模态数据的元素映射到一个共享的隐空间(Latent Space)中,在其中将语义相似的元素对齐,使在该隐空间中不同模态数据之间可以进行比较。跨模态检索领域的一种经典方法是由Rasiwasia等^[6]提出的典型关联分析(Canonical Correlation Analysis, CCA),最大限度地提高了图像和文本之间的相关性。基于CCA,很多学者提出了改进的方法^[8-10]。然而,基于CCA的方法是不稳定的,并由于协方差矩阵计算^[11]导致高内存成本。

最近大多数与文本和图像模态相关的工作都是用基于排名的损失函数进行训练的^[11-15]。Frome等^[12]提出了一种将单词和视觉内容投影到联合空间的方法,利用排名损失,当非匹配单词的排名高于匹配单词时,对其进行惩罚。Kiros等^[13]提出了一种跨模态的图像-文本检索方法,利用三重排序损失对图像特征进行投影,并利用基于RNN的句子描述对一个共同的潜在空间进行检索。Wang等^[11]提出了一种多模态注意机制,有选择地关注句子片段和特定的图像区域进行相似性计算。

1.2 视频文本跨模态检索

与视频和文本2种模态相关的计算机视觉的2个主要任务是视频文本检索(Video-Text Retrieval)和视频字幕生成(Video Captioning),本文只关注检索任务。与图像-文本检索算法相似,大多数视频-文本检索算法使用共享子空间的方法。Xu等^[14]通过word2vec模型^[15]对从给定句子中提取的每个主语-动词-宾语三元组进行矢量化,然后将其聚合,使用卷积神经网络将主语、动词、宾语(SVO)向量转换成句子级向量,视频特征向量是通过将每帧特征进行平均池化得到的。然后利用最小二乘损失函数进行训练,将句子表示和视频表示投影到一个联合嵌入空间中。Faghri等^[7]提出了一个寻求联合使用多个句子编码策略的方法,这些策略包括bag-of-words、word2vec和GRU。然而,该工作只是使用视频的平均池化

编码。Dong 等^[2]提出了一种针对视频句子匹配任务的 word2vec 方法,利用均值平方损失将向量化的句子投射到视觉特征空间中。Mithun 等^[3]提出了一种基于图像、文本和声音形态的共享空间,并将其应用于视频文本检索任务中。

这些视频文本检索算法大多与图像文本检索算法在设计上非常相似,未利用多模态数据中丰富的特征进行检索。与现有的工作相比,本文算法能够同时提取出多模态特征中的全局特征、上下文语义特征以及局部特征以实现更有效的检索。

2 模型设计

2.1 模型框架

本文提出的 MCCR 框架如图 1 所示,该框架

采用了一个对称式的网络结构,即视频编码层和文本编码层具有相似的结构,这也意味着本框架在处理视觉信息和文本信息上采用了类似的处理方法。因此,为了方便描述,这里将以视频模态编码为例,详细描述所提出的多语义线索特征,并在此基础上对如何提取文本内容的多语义特征进行阐述。在视频编码模块,基于平均池化和多头自注意力机制,分别挖掘出多模态数据中重复出现和对语义起重要作用的特征,得到全局特征;基于双向 GRU 捕捉到上下文语义特征,挖掘出多模态数据中上下文之间的交互特征;基于一维位神经卷积神经网络(CNN)捕捉局部特征,挖掘出局部视频帧之间的交互特征。通过全局特征、上下文交互特征和局部特征的耦合,可以得到多模态数据中丰富的语义线索。

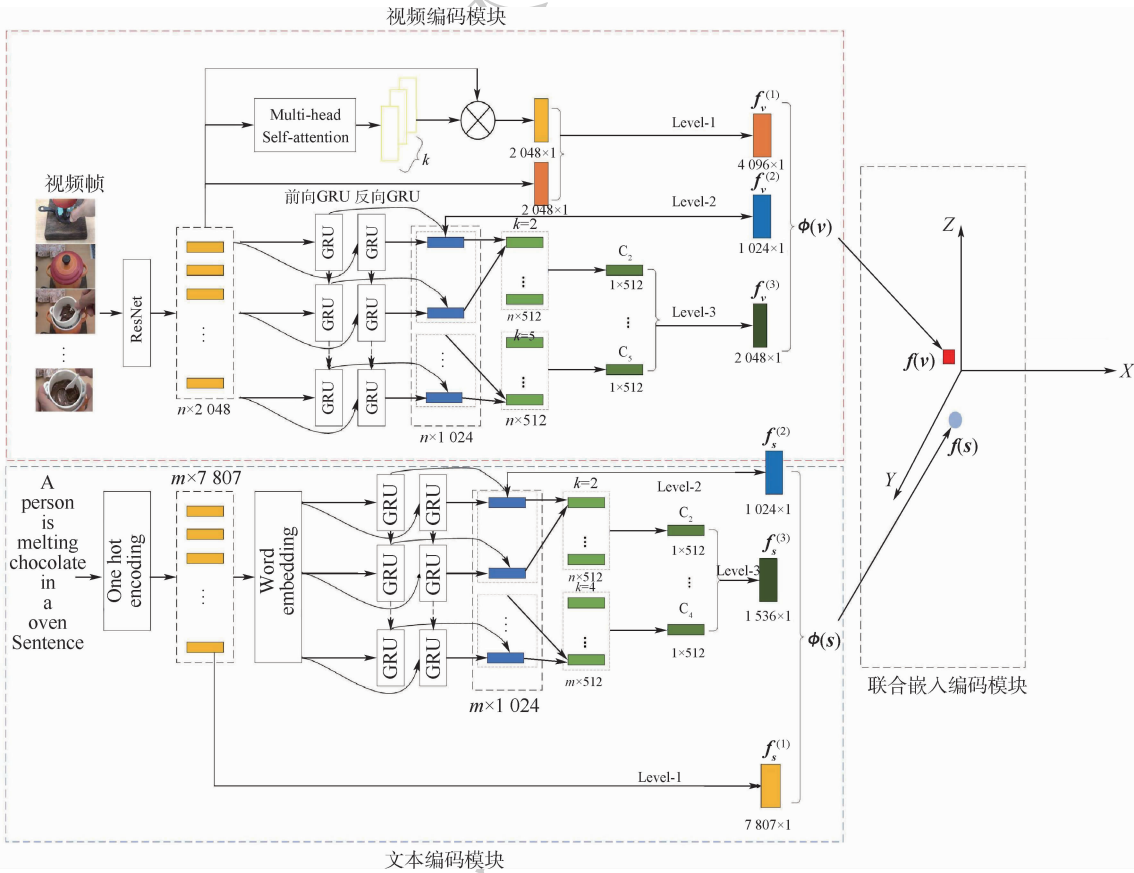


图 1 MCCR 模型示意图

Fig.1 Schematic diagram of proposed MCCR model

2.2 视频编码模块

2.2.1 Level-1:视频全局特征编码

视频编码层采用注意力机制和平均池化相结合的方式计算视频的全局编码。

平均池化的方法在视频跨模态检索中被广泛使用,它能捕获在视频内容中重复出现的视觉模态。这些重复出现的视觉模态往往是全局

性的,对整体的语义有着重要作用。假设每隔 0.5 s 抽取一帧,每帧通过 ImageNet 预训练后的 ResNet 模型^[16]的倒数第二层作为图片特征编码(维度为 2 048 维),可以得到 $\{v_1, v_2, \dots, v_n\}$ 表示视频帧的图像特征, v_t 为视频第 t 帧的特征, n 为视频的帧数。本文用 $\bar{v}_t$ 表示平均池化的编码结果:

$$\bar{\mathbf{v}}_v = \frac{1}{n} \sum_{t=1}^n \mathbf{v}_t \quad (1)$$

在一段视频中,往往会出现其中的某几帧会对描述整段视频语义其关键作用的情况。例如,在一段跨栏田径比赛的视频中,包含跨栏动作的视频往往只占整段视频的一小部分,但是能够将跨栏的视频与跑步的视频区分开来。对于这种情况,平均池化方法往往会忽略关键帧的作用,导致全局特征无法有效表示视频全局信息。鉴于此,提出了一种关键帧加权的池化方法,将不同视频帧根据其重要性赋予不同权重,从而在池化的过程中能够重点关注关键帧的信息。受 Song 和 Soleymani^[17] 的启发,本文采用一种自下而上的多头目自注意力机制捕捉视频模态数据内部的细微交互,有选择性地关注视频内部的重要信息来增强模型的表征能力。它是通过 2 个偏置项为 0 的全连接层和一个 softmax 层实现的:

$$\beta = \text{softmax}(\omega_2 \tanh(\omega_1 \phi(\mathbf{v}))) \quad (2)$$

式中: $\omega_1 \in \mathbf{R}^{A \times D}$ 、 $\omega_2 \in \mathbf{R}^{k \times A}$ 、 $\phi(\mathbf{v}) \in \mathbf{R}^{n \times D}$ 为注意力机制的超参数,用于学习视频特征的权重, D 为每一帧特征的维度,根据经验,设置超参数 $A = D/2$,超参数 $k = 3$, $\mathbf{v}$ 为视频帧组成的矩阵。注意力机制最终得到每一帧的 k 个权重,权重的大小表示不同维度注意力机制对视频每帧特征的选择倾向,可以自动学习视频中的重要部分,过滤掉非重要部分。最终,视频的编码由视频中每帧的加权平均获得:

$$\mathbf{v}_v = \frac{1}{k} \sum_{i=1}^n \sum_{t=1}^k \beta_{t,i} \mathbf{v}_t \quad (3)$$

式中: $\beta_{t,i}$ 为第 i 个注意力机制判定的第 t 帧的权重。

通过结合视频中平均池化特征和加权池化特征,可以得到视频全局特征编码:

$$\mathbf{f}_v^{(1)} = [\bar{\mathbf{v}}_v, \mathbf{v}_v] \quad (4)$$

2.2.2 Level-2: 视频上下文特征编码

Level-1 中只提取了视频中的全局特征,忽略了视频内部丰富的时空动态信息,导致模型无法区分视频帧之间的语义交互关系。本文采用双向 GRU (bi-GRU)^[18] 对视频上下文特征进行编码,其参数比 LSTM 少,因此只需要较少的训练数据。

$$\begin{cases} \vec{\mathbf{h}}_t = \overrightarrow{\text{GRU}}(\mathbf{v}_t, \vec{\mathbf{h}}_{t-1}) \\ \overleftarrow{\mathbf{h}}_t = \overleftarrow{\text{GRU}}(\mathbf{v}_t, \overleftarrow{\mathbf{h}}_{t-1}) \\ \mathbf{h}_t = [\vec{\mathbf{h}}_t, \overleftarrow{\mathbf{h}}_t] \end{cases} \quad (5)$$

一个双向 GRU 由 2 个分离的 GRU 层组成,即前向 GRU 层和一个反向的 GRU。前向 GRU 将

每帧视频的特征按照正常的先后顺序输入,而反向 GRU 按照反向的顺序输入。用 $\vec{\mathbf{h}}_t$ 和 $\overleftarrow{\mathbf{h}}_t$ 表示在相应的时间步 $t = 1, 2, \dots, n$ 的隐含状态,隐含状态可以用式(5)表示。

$\overrightarrow{\text{GRU}}$ 和 $\overleftarrow{\text{GRU}}$ 分别表示前向和后向 GRU,通过 $\vec{\mathbf{h}}_t$ 和 $\overleftarrow{\mathbf{h}}_t$ 传递时间 t 的状态,通过 $\vec{\mathbf{h}}_t$ 和 $\overleftarrow{\mathbf{h}}_t$ 可以得到 GRU 在时刻 t 的输出。根据经验,将前向 GRU 和后向 GRU 中的隐藏向量大小设置为 512。因此, $\mathbf{h}_t$ 的大小为 1 024。将所有的输出结合,可以得到一个特征表示列表 $\mathbf{H} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$,维度为 $1\,024 \times n$ 。通过平均池化,可以得到视频上下文特征编码为

$$\mathbf{f}_v^{(2)} = \frac{1}{n} \sum_{t=1}^n \mathbf{h}_t \quad (6)$$

2.2.3 Level-3: 视频局部特征编码

在考虑其他视频帧和某一帧的交互特征时,Level-2 层将其他视频帧的输出权重都设为相同值,这会导致模型难以捕捉视频中的重要信息。大多数情况下,某一帧只与它相邻几帧有较大的交互作用,与距离比较远的交互作用较小。充分利用相邻帧之间的局部交互信息,将形成能够区分视频细微差别的细粒度视频特征。因此,受文本分类领域^[19] 方法启发,本文在 Level-2 上下文特征的基础上,采用了 1 维卷积对相邻帧进行联合编码,以捕捉视频的局部交互信息。

CNN 的输入是 2.2.2 节双向 GRU 模块生成的特征列表 $\mathbf{H}$ 。用 $\text{Conv1d}_{k,r}$ 表示一维卷积模块,包含 $r = 512$ 个大小为 k 的卷积核。将 $\mathbf{H}$ 通过 $\text{Conv1d}_{k,r}$ 和 padding 操作,可以产生 $n \times k$ 个特征。用 RELU 激活函数进行非线性变换,增强模型的拟合能力。因为每个视频的帧数 n 不一样,可以利用 max pooling 将特征列表 $\mathbf{H}$ 转换为固定长度为 r 的向量 $\mathbf{c}_k$:

$$\mathbf{c}_k = \max \text{pooling}(\text{RELU}(\text{Conv1d}_{k,r}(\mathbf{H}))) \quad (7)$$

一个 $k = 2$ 的卷积核可以用作求 2 个在 $\mathbf{H}$ 中的相邻行之间的交互,同样的, k 表示相邻的 k 个视频帧之间的交互。用卷积核 $k = 2, 3, 4, 5$ 生成多个范围的特征表示。它们的输出被串联起来,形成基于双向 GRU-CNN 的编码:

$$\mathbf{f}_v^{(3)} = [\mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4, \mathbf{c}_5] \quad (8)$$

由于 $\mathbf{f}_v^{(1)}$ 、 $\mathbf{f}_v^{(2)}$ 、 $\mathbf{f}_v^{(3)}$ 是通过特定的编码方法在不同的层次上依次得到的,所以 3 个编码结果存在一定的互补性,一定的冗余是合理的。因此,将 3 层的输出串联起来,得到输入视频的多级编码:

$$\phi(\mathbf{v}) = [\mathbf{f}_v^{(1)}, \mathbf{f}_v^{(2)}, \mathbf{f}_v^{(3)}] \quad (9)$$

2.3 文本编码模块

视频编码模块经过细微修改后,同样适用于文本。给定一个长度为 m 的句子 s ,用独热编码表示它的每个单词。因此,可以生成一个独热编码向量序列 $\{w_1, w_2, \dots, w_m\}$, w_t 表示第 t 个单词的向量。在文字编码模块,仅用所有独热编码向量求平均得到全局编码 $f_s^{(1)}$,相当于经典的词袋法。

对于 Level-2 的编码,首先通过用 3 000 万张 Flickr 图片的英语描述进行预训练^[20] 的 word2vec^[15] 模型,将其转换为一个词向量。将基于双向 GRU 的句子编码的特征表示为 $f_s^{(2)}$ 。类似地,本文将基于一维 CNN 的特征编码记为 $f_s^{(3)}$ 。这里,使用 3 个一维卷积块,卷积核大小 $k=2,3,4$,将文本编码层的所有编码结果串联,得到句子的多级编码:

$$\phi(s) = [f_s^{(1)}, f_s^{(2)}, f_s^{(3)}] \quad (10)$$

$\phi(v)$ 和 $\phi(s)$ 没有直接的关系,不直接进行比较。在视频文本相似度计算中,需要将向量投影到公共空间中,然后给出学习算法。

2.4 联合嵌入编码模块

对于编码后得到的视频特征 $\phi(v)$ 和文本特征 $\phi(s)$,将它们通过仿射转换映射到一个共享空间中,通常的做法是利用全连接层进行映射。经过一些理论研究^[7],本文采用在全连接后加上批量正则层,因为这样效果更好。通过下面的变换,将得到共享空间的视频特征向量 $f(v)$ 和 $f(s)$:

$$\begin{cases} f(v) = \text{BN}(W_v \phi(v) + b_v) \\ f(s) = \text{BN}(W_s \phi(s) + b_s) \end{cases} \quad (11)$$

式中: W_v 和 W_s 分别为视频侧和文本侧的全连接的参数; b_v 和 b_s 分别为 W_v 和 W_s 的参数。

2.5 损失函数

在联合嵌入空间中,视频与文本对之间的相似度能反映出原视频与对应文本之间的语义相似度,以往的许多方法结合双向排序项,最小化三元组损失函数,以学习最大化视频嵌入和相应文本嵌入之间的相似度,并最小化所有其他非匹配项的相似度。优化问题可以写为

$$\min_{\theta} \sum_v \sum_{s^-} [\tau - S(v, s) + S(v, s^-)]_+ + \sum_s \sum_{v^-} [\tau - S(s, v) + S(s, v^-)]_+ \quad (12)$$

式中: $[\cdot]_+ = \max(f, 0)$; s^- 为和视频不匹配的文本; s 为和视频匹配的文本;对于视频 v^- 和 v 的表示方法类似; τ 为距离常数; θ 为神经网络所有参数; $S(v, s)$ 定义为相似度函数,测量视频与文本在共享空间内的相似度。

传统的三元组损失面临正负样本不平衡的问题,即在实际情况当中,负样本的个数往往远大于正样本的个数。然而,数量庞大的负样本中仅有少量样本能够对训练起到促进作用,如果随机选取负样本容易造成模型训练不稳定的情况。为了解决这个问题, Faghri 等^[7] 提出了硬负样本挖掘的方法,从负样本集合中挑选较有价值的困难负样本参与训练。然而硬负样本挖掘容易受到数据中的噪声影响,且忽略了负样本集合中其他负样本的作用。在此基础上,本文通过加权排序的思想,对损失函数进行优化,根据匹配的视频文本对的相对排序对损失函数进行加权,提出基于距离度量的三元组损失函数:

$$\min_{\theta} \sum_v L(r_v) [\alpha - S(v, s) + S(v, \bar{s})]_+ + \sum_s L(r_s) [\alpha - S(s, v) + S(s, \bar{v})]_+ \quad (13)$$

式中: $L(\cdot)$ 为不同秩的加权函数;对于视频嵌入 v , r_v 为所有跟 v 比较的句子中的相似度排名。同样,对于一个文本嵌入 s , r_s 为所有跟 s 比较的视频中的相似度排名。将加权函数定义为 $L(r) = 1 + 1/(N - r + 1)$, N 为 batch 的大小,在本文中为 128。

从式(13)中可以明显看出,损失函数是基于检索结果中语义相似样本对相似度排名权重的函数。如果语义相似的样本对的相似度在所有样本对的相似度中排名靠前,那么 $L(\cdot)$ 将把一小部分重量分配给损失,而不会使损失函数过大。然而,如果语义相似的样本对的相似度排在靠后的位置, $L(\cdot)$ 将分配一个大得多的权重给损失,这最终会将语义相似的样本对的相似度的排名靠前。

3 实验

3.1 数据集

本文在 2 个基准数据集上进行了实验: Microsoft Research Video to Text (MSR-VTT)^[21] 数据集和 Microsoft Video Description Dataset (MSVD)^[22] 数据集,以评估所提框架性能。

MSR-VTT 是一个大型的视频描述数据集,包含 10 000 个视频剪辑。数据集分为 6 513 个训练视频、2 990 个测试视频和 497 个验证集视频,每个视频有 20 句话的描述。就句子的数量和词汇量而言,这是最大的视频字幕数据集之一。

MSVD 数据集包含 1 970 个 Youtube 剪辑,每个视频都有大约 40 个句子的注释。本文只使用英语描述,利用 1 200 个训练视频,100 个验证视

频,670 个测试视频。

3.2 实验设置

3.2.1 实验细节

本文代码使用 Pytorch 框架实现,模型使用 Adam 优化器进行训练,学习率为 10^{-4} , α 为 0.3,批尺寸为 128。以上超参数均在验证集上进行选择,本文实验均是多次实验求均值后的结果。

在计算损失函数时,将一个批(batch)中所有的样本均视为锚样本,对每个锚样本,在该 batch 中寻找所有符合条件的正负例构建三元组,并对每个批中所有视频和文本之间的相似度进行排序,以计算检索损失和语义损失。

3.2.2 评价指标

本文实验选取了使用视频检索文本(Video-to-Text Retrieval)及使用文本检索视频(Text-to-Video Retrieval)2 个任务来衡量模型的准确率。实验采用中位序数(Median Rank, MedR)以及 top K 召回率(Recall rate at top K, R@K)作为准确率评价指标。中位序数指所有检索样本对应的被检索样本在检索结果中的序数的中位数,其值越低,代表模型的准确率越高;top K 召回率,例如视频检索文本任务中的 top 5 召回率,指使用视频检索文本时,视频对应的文本出现在检索结果前 5 位的比率,其值越高,代表模型准确率越高。

3.3 结果比较

本文选择了近年来提出的 3 种跨模态视频检索领域模型作为基准算法,与本文提出的算法进行准确率比较。另外,为了证明本文提出的损失函数的有效性,加入本文(VSE++)。

1) VSE++^[7]:在跨模态检索中,提出硬负样本挖掘的方法。其中,用 ResNet-152 提取视频特征,用平均池化的方法表示视频侧总特征。

2) W2VV^[4]:将不同模态数据都映射到视觉语义空间的方法。

3) Fusion^[3]:利用视频模态特征与语音特征融合的表示视频侧特征,因为在数据集 MSVD 中,视频是消音的,所以仅提取视频中的视觉特征,而在 MSR-VTT 数据集中该方法加入了音频特征。

4) 本文(VSE++):在本文模型的基础上,用式(12)的损失函数进行训练。

3.3.1 跨模态视频检索

由表 1 和表 2 可知,本文模型的准确率在不同的任务、不同的数据集下均显著地优于 3 种基准算法。其中,在 MSVD 数据集上,Recall Sum 相比目前最优的算法,提高了 3.7%,视频检索文本任务中召回率相比最优的算法提高了 5.0%,在 MSR-VTT 数据集上,Recall Sum 相比于目前最优的算法,提高了 4.0%,视频检索文本任务中召回率相比最优的算法提高了 11.1%,注意到 Fusion 在视频检索文本任务上表现较本文提出的算法效果好,可能是因为在 MSR-VTT 数据集上,存在音频数据,给该算法提供了额外的语音特征。本文提出的损失函数相较于硬负样本挖掘的方法在 MSVD 数据集上总召回率提升了 2.0%,在 MSR-VTT 提升了 2.4%。3 种基准算法在图像和文本编码模块上均未能充分挖掘出数据特征,而本文模型通过多语义线索和改进的三元组距离度量损失函数,取得了更高的检索准确率。

表 1 在 MSR-VTT 数据集上的结果

Table 1 Results on MSR-VTT dataset

算法	文本检索视频				视频检索文件				Recall Sum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
VSE++	5.0	16.4	24.6	47	7.7	20.3	31.2	28	105.2
W2VV	5.5	17.6	25.9	51	9.1	24.6	36.0	23	118.7
Fusion	7.0	20.9	29.7	38	12.5	31.3	42.4	14	143.8
本文(VSE++)	7.6	21.7	31.2	31	12.2	29.4	42.4	18	144.5
本文	7.8	23.0	33.1	29	13.1	30.7	43.1	15	150.8

表 2 在 MSVD 数据集上的结果

Table 2 Results on MSVD dataset

算法	文本检索视频				视频检索文本				Recall Sum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
VSE++	15.4	39.6	53.3	9	21.2	43.4	52.2	9	225.1
W2VV	15.4	39.2	51.4	10	16.3	33.4	44.8	14	200.5
Fusion	18.9	46.1	60.9	6	30.6	49.1	61.5	6	267.1
本文(VSE++)	19.7	48.2	61.0	6	31.7	50.7	61.8	6	272.5
本文	20.9	49.0	62.6	5	32.2	51.1	62.2	5	278.0

3.3.2 消融分析

表3展示MCCR的各个编码部分(Level-1、Level-2、Level-3)及其组合在MSR-VTT数据集上对检索结果的影响,该结果可以反映模型中各个部分的重要程度。图2为MSR-VTT数据集的6个视频检索文本的测试样例,直观地展示了各编码部分及其组合的检索效果,其中GT表示随机抽取的正样本,各算法后的数字表示该算法检索时正例的最佳排名。

可以看到,单独使用一个编码部分进行训练时,在3个部分中,使用Level-3进行训练得到的准确率最高,这反映了局部特征编码的重要性。

表3 在MSR-VTT数据集上的消融分析结果
Table 3 Ablation analysis results on MSR-VTT dataset

方法	文本检索视频				视频检索文本				Recall Sum
	R@1	R@5	R@10	MedR	R@1	R@5	R@10	MedR	
Level-1	6.4	18.9	27.1	46	11.9	28.3	39.2	22	131.8
Level-2	6.3	19.7	28.8	38	10.0	26.2	38.3	20	128.8
Level-3	7.3	21.5	31.2	32	10.6	27.3	38.5	20	136.4
Level-(1+2)	7.2	21.3	29.6	37	12.1	30.5	40.9	17	141.6
Level-(1+3)	7.4	21.2	32.3	30	12.4	29.9	42.5	16	147.1
Level-(2+3)	7.6	22.4	32.2	31	11.9	30.6	42.4	16	147.2
Level-(1+2+3)	7.8	23.0	33.1	29	13.1	30.7	43.1	15	150.8



图2 MSR-VTT数据集的6个视频检索文本的测试样例,以及MCCR的各个编码部分(Level-1、Level-2、Level-3)及其组合的检索结果

Fig.2 Test samples of 6 video retrieval words in MSR-VTT dataset, as well as the retrieval results of each coding part (Level-1, Level-2, level-3) of MCCR and its combination.

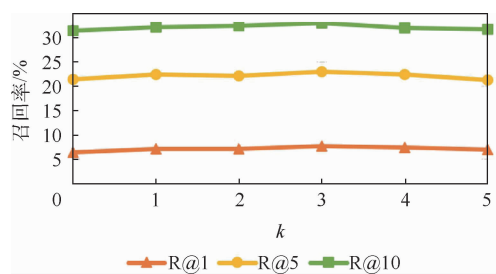


图3 文本检索视频
Fig.3 Text to video retrieval

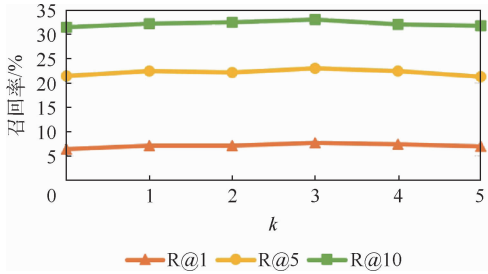


图4 视频检索文本
Fig.4 Video to text retrieval

由图3和图4可知,通过适当地增加注意力模型的个数,可以提升模型的检索准确率。当 k 小于3时,随着注意力模型个数的增加,检索准确度存在一定程度的上升,当 k 大于3时,检索准确度下降。原因可能是在一定范围内,随着注意力模型个数的增加,提取出重要视频帧的信息也越来越多,因此可以提升检索效果。然而注意力机制越多,需要训练的参数越多,模型欠拟合,检索效果下降。

4 结 论

针对当前人们对不同模态数据检索的需求,本文提出了基于多语义线索的跨模态视频检索算法,可以根据给定的句子检索出语义相似的视频,同时也可以根据给定的视频检索出语义相似的句子,进而达到跨模态检索的目的。本文所做贡献如下:

1) 将每种模态数据都通过了数层的编码,将各层的编码结果耦合连接得到多种语义特征。在视频编码模块,基于平均池化和多头自注意力机制,分别挖掘出多模态数据中重复出现和对语义起重要作用的特征,得到全局特征;基于双向GRU捕捉到上下文语义特征,挖掘出多模态数据局部特征,挖掘出局部视频帧之间的交互特征。通过全局特征、上下文交互特征和局部特征的耦合,可以得到丰富的语义线索描述这2种模态数据。

2) 在模型的基础上提出了基于加权排序的三元组损失函数,在一定程度上解决了负样本对比正样本对相似度高的问题,进而提升了模型的检索效果。

尽管本文模型相对现有的算法已经有了明显进步,但还没有达到尽善尽美的地步。当前算法在视频模态中未利用语音模态数据,对视频的语义也起到了重要的作用。将来会在算法中考虑语音模态数据,进而提高检索效果。

参考文献 (References)

[1] 张鸿,吴飞,庄越挺. 跨媒体相关性推理与检索研究[J]. 计算机研究与发展,2008,45(5):869.
ZHANG H, WU F, ZHUANG Y T. Research on cross-media correlation inference and retrieval[J]. Computer Research and Development,2008,45(5):869(in Chinese).
[2] DONG J, LI X, SNOEK C G M. Predicting visual features from text for image and video caption retrieval[J]. IEEE Transactions on Multimedia,2018,20(12):3377-3388.
[3] MITHUN N C, LI J, METZE F, et al. Learning joint embedding

with multimodal cues for cross-modal video-text retrieval[C]// Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. New York: ACM Press,2018:19-27.
[4] DONG J, LI X, SNOEK C G M. Word2VisualVec: Image and video to sentence matching by visual feature prediction [EB/OL]. (2016-04-23) [2020-08-01]. https://arxiv.org/abs/1604.06838.
[5] TORABI A, TANDON N, SIGAL L. Learning language-visual embedding for movie understanding with natural-language[EB/OL]. (2016-08-26) [2020-08-01]. https://arxiv.org/abs/1609.08124.
[6] RASIWASIA N, COSTA P J, COVIELLO E, et al. A new approach to cross-modal multimedia retrieval[C]// Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press,2010:251-260.
[7] FAGHRI F, FLEET D J, KIROS J R, et al. VSE++: Improving visual-semantic embeddings with hard negatives [EB/OL]. (2017-07-18) [2020-08-01]. https://arxiv.org/abs/1707.05612.
[8] GONG Y, KE Q, ISARD M, et al. A multi-view embedding space for modeling internet images, tags, and their semantics [J]. International Journal of Computer Vision,2014,106(2):210-233.
[9] HODOSH M, YOUNG P, HOCKENMAIER J. Framing image description as a ranking task: Data, models and evaluation metrics[J]. Journal of Artificial Intelligence Research,2013,47(24):853-899.
[10] 李志欣,施智平,陈宏朝,等. 基于语义学习的图像多模态检索[J]. 计算机工程,2013,39(3):258-263.
LI Z X, SHI Z P, CHEN H C, et al. Multi-modal image retrieval based on semantic learning[J]. Computer Engineering,2013,39(3):258-263(in Chinese).
[11] WANG L, LI Y, HUANG J, et al. Learning two-branch neural networks for image-text matching tasks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2018,41(2):394-407.
[12] FROME A, CORRADO G S, SHLENS J, et al. Devise: A deep visual-semantic embedding model[C]// Advances in Neural Information Processing Systems. New York: ACM Press,2013:2121-2129.
[13] KIROS R, SALAKHUTDINOV R, ZEMEL R S. Unifying visual-semantic embeddings with multimodal neural language models [EB/OL]. (2014-11-10) [2020-08-01]. https://arxiv.org/abs/1411.2539.
[14] XU R, XIONG C, CHEN W, et al. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework[C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence,2015:6.
[15] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-01-16) [2020-08-03]. https://arxiv.org/abs/1301.3781.
[16] 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.

- [17] SONG Y, SOLEYMANI M. Polysemous visual-semantic embedding for cross-modal retrieval [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 1979-1988.
- [18] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [EB/OL]. (2014-01-03) [2020-08-01]. <https://arxiv.org/abs/1406.1078>.
- [19] KIM Y. Convolutional neural networks for sentence classification [EB/OL]. (2014-08-25) [2020-08-01]. <https://arxiv.org/abs/1408.5882>.
- [20] YOUNG P, LAI A, HODOSH M, et al. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions [J]. Transactions of the Association for Computational Linguistics, 2014, 2(1): 67-78.
- [21] XU J, MEI T, YAO T, et al. MSR-VTT: A large video description dataset for bridging video and language [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 5288-5296.
- [22] CHEN D, DOLAN W B. Collecting highly parallel data for paraphrase evaluation [C] // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 2011: 190-200.

作者简介:

丁洛 男, 硕士研究生。主要研究方向: 多模态检索、目标检测。

李逸凡 男, 博士研究生。主要研究方向: 视觉问答、目标识别技术。

漆舒汉 男, 博士, 教授, 硕士生导师。主要研究方向: 计算机视觉、多媒体信息检索和机器博弈。

Cross-modal video retrieval algorithm based on multi-semantic clues

DING Luo¹, LI Yifan¹, YU Chenglong², LIU Yang¹, WANG Xuan^{1,3}, QI Shuhan^{1,3,*}

(1. School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China;

2. School of Digital Media, Shenzhen Institute of Information Technology, Shenzhen 518172, China;

3. Peng Cheng Laboratory, Shenzhen 518055, China)

Abstract: Most of the existing cross-modal video retrieval algorithms map heterogeneous data to a space, so that semantically similar data are close to each other and semantically dissimilar data are far from each other, that is, the global similarity relationship of different modal data is established. However, these methods ignore the rich semantic clues in the data, which makes the performance of feature generation poor. To solve this problem, we propose a cross-modal retrieval model based on multi-semantic clues. This model captures the data frames that play an important role in semantics within video model through multi-head self-attention mechanism, and pays attention to the important information of video data to obtain the global characteristics of the data. Bidirectional Gate Recurrent Unit (GRU) is used to capture the interaction characteristics between contexts within multimodal data. Our method can also mine the local information in video and text data through the joint coding of the slight differences between the local data. Through the global features, context interaction features and local features of the data, the multi-semantic clues of the multi-modal data are formed to better mine the semantic information in the data and improve the retrieval effect. Besides this, an improved triplet distance measurement loss function is proposed, which adopts the difficult negative sample mining method based on similarity sorting and improves the learning effect of cross-modal characteristics. Experiments on MSR-VTT dataset show that the proposed method improves the text retrieval video task by 11.1% compared with the state-of-the-art methods. Experiments on MSVD dataset show that the proposed method improves the text retrieval video task by 5.0% compared with the state-of-the-art methods.

Keywords: cross-modal video retrieval; multi-semantic clues; multi-leader attention mechanism; distance measurement loss function; multi-modal

Received: 2020-08-26; Accepted: 2020-09-04; Published online: 2020-10-30 10:22

URL: kns.cnki.net/kcms/detail/11.2625.V.20201030.1006.001.html

Foundation items: National Natural Science Foundation of China (61902093); Natural Science Foundation of Guangdong Province (2020A1515010652)

* Corresponding author. E-mail: shuhanqi@cs.hitsz.edu.cn

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

DOI: 10.13700/j.bh.1001-5965.2020.0471

基于注意力机制的跨分辨率行人重识别

廖华年¹, 徐新^{1, 2, 3, *}

(1. 武汉科技大学 计算机科学与技术学院, 武汉 430065;

2. 武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室, 武汉 430065;

3. 武汉大学 深圳研究院, 深圳 518000)

摘 要: 行人图像分辨率的变化对现有的行人重识别方法带来了很大的挑战。针对这一问题,提出了一种新的跨分辨率行人重识别方法。该方法从两方面解决分辨率变化带来的识别困难:一方面通过通道注意力机制和空间注意力机制捕捉人物特征获取局部区域;另一方面通过核动态上采样模块恢复任意分辨率图像的局部区域信息。为了验证所提方法的有效性,在 Market1501、CUHK03 和 CAVIAR 三个公开数据集上开展了对比实验,实验结果表明:所提方法取得了最佳性能。

关 键 词: 行人重识别; 通道注意力机制; 空间注意力机制; 图像超分辨率; 上采样
中图分类号: TP391

文献标志码: A **文章编号:** 1001-5965(2021)03-0605-08

随着平安城市、雪亮工程、天网工程的推进,视频监控系统得到了发展和普及的同时,视频侦查技术也进入了广泛运用阶段,利用视频智能分析技术,自动地从海量监控数据中对特定的行人目标进行检索、分析、比对的方式正在逐渐取代人工判别。行人重识别^[1](person Re-identification, Re-ID)作为智能视频侦查的关键技术之一,在实际应用方面对于预防犯罪、嫌犯追踪和治安管理都具有积极作用。

行人重识别旨在匹配不同监控摄像头视图下相同身份类信息图像,它不仅需要解决同一摄像头下的行人遮挡、姿势、角度、光线等问题,还需注意到不同摄像头之间存在的摄像头规格不同。无约束的成像条件给行人重识别带来了一定的挑战,大多数现有的行人重识别方法都假设查询图像和图库图像具有相似且足够高的分辨率。然而,由于摄像头和行人之间的距离不受约束且不

同摄像头之间的参数可能不同,导致行人图像往往具有不同的分辨率,这种分辨率不匹配问题给行人重识别带来了困难。与高分辨率(High Resolution, HR)图像相比,低分辨率(Low Resolution, LR)图像包含的身份细节要少得多,直接跨分辨率匹配图像将导致性能显著下降。

为了解决分辨率不匹配问题,许多研究者开展了低分辨率行人重识别的研究^[2-3],但性能提升不大。随着深度学习的不断发展,引入深度学习概念后的跨分辨率行人重识别任务^[4-9]取得了较大的进步。早期的工作主要是通过建立高分辨率图像与低分辨率图像特征之间的映射关系来解决跨分辨率的匹配问题,Jing等^[2]设计了一种半耦合低秩字典学习方法构建高低分辨率之间的联系;Li等^[3]则是先假设同一行人的不同图像在某个特征空间具有相似的结构,从而得到一个跨分辨率图像对齐网络,将高、低分辨率关系引入到距

收稿日期: 2020-08-28; 录用日期: 2020-09-18; 网络出版时间: 2020-11-06 11:06

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201105.1153.003.html

基金项目: 国家自然科学基金(U1803262, 61602349, 61440016); 深圳市科技计划基础研究项目(JCYJ20170818143246278)

* 通信作者。E-mail: xuxin0336@163.com

引用格式: 廖华年, 徐新. 基于注意力机制的跨分辨率行人重识别[J]. 北京航空航天大学学报, 2021, 47(3): 605-612.

LIAO H N, XU X. Cross-resolution person re-identification based on attention mechanism[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 605-612 (in Chinese).

离度量方法中。但是在无约束的成像条件下,行人图像分辨率是不同且多样的,无法一一对应,查询图像与检索图像分辨率比例并不是一个固定的值,在实际场景中,多个行人图像对之间的分辨率比例多样,早期工作中提出的方法并不适用。

受上述工作的启发,Jiao 等^[4]提出的图像超分辨率(Super-Resolution, SR)和行人身份识别联合学习方法,通过 2、3、4 三个尺度的超分辨率网络,恢复对应尺度的低分辨率行人图像中对身份识别有效的高频外表信息,解决低分辨率行人重识别问题。Wang 等^[5]提出将 2、4、8 三个尺度的超分辨率网络级联起来,通过逐步恢复低分辨率图像细节信息,最后与图库图像进行距离度量。尽管这些方法在跨分辨率行人重识别上取得了一定的效果,但这些方法都需要预先定义图像对之间的尺度比例,再通过与之适配的放大因子模块进行图像超分辨率重建工作。

与上述工作不同,受人类视觉注意力机制(Attention mechanism, Attention)^[13]的影响,本文提出了基于注意力机制的局部超分辨率联合身份学习网络解决上述问题。具体来说,Attention 的目的是辅助网络找到更利于识别的局部,但是即使是同一行人的不同分辨率图像的显著区域也会有一定的差异,因此提出了一个基于注意力机制的跨分辨率行人重识别方法,首先查询图像输入到由编码解码网络组成的 Attention 网络,目的是得到唯一的且利于识别任务的注意力图;然后通过核动态上采样的方法^[14],任意尺度的重建低分辨率图像;最后经过行人重识别网络得到分类结果。本文主要贡献如下:

1) 提出了基于通道和空间注意力机制的跨分辨率行人重识别方法。该方法主要意图在通过关注和比较不同分辨率行人图像对相同位置的显著区域,然后利用自编码网络的学习,得到任意分辨率行人图像的利于识别的局部区域。

2) 使用任意上采样因子的跨分辨率重识别方法,使得网络能够处理任意低分辨率的查询图像的重建。经过注意力机制得到的局部区域能够被重建到与图库图像同一分辨率。

1 相关工作

由于在真实场景下,背景^[15]、姿态、照明^[16]、视角、相机^[17]等条件变化很大,在无约束的成像条件下,人的图像分辨率的变化可能是最常见的,匹配不同分辨率的行人样本需要行人重识别算法关注不同的视觉特征。例如,图 1 展示了某商场不同摄像头下的行人图像,在高分辨率的图像样本中,可以通过发型或者衣服标志来区分不同行人,如 C 分别与 A、D 进行匹配,可以得出 A 和 C 是同一行人,D 为不同行人。但在低分辨率图像样本中人眼都无法观察到这些细节,比如 B 难以判定和 A、C 是否为同一行人,对于机器而言这些用于区分行人的标识在低分辨率图像中是不可用的,ReID 方法需要借助于剪影或全局纹理来进行可靠的匹配。而且,同一个人高分辨率和低分辨率样本的差异甚至可能比不同人在相似分辨率下的样本差异更大。因此,需要对 ReID 方法进行专门的处理,以处理人物图像的跨分辨率变化。

解决行人重识别方向中的分辨率问题,最简单的方法是使用更大的数据集覆盖尽可能多的分辨率比例,从而构建高低分辨率之间的关系,但是这需要大量的数据和标注,且难以罗列出所有的分辨率比例场景。

1.1 学习低、高分辨率图像的关系

早期的工作主要是学习低、高分辨率图像之间的映射关系。2015 年,Jing 等^[2]使用字典学习的方法学习低、高分辨率图像之间的映射函数,通过得到的映射函数实现 1/8 尺度(低分辨率、高分辨率图像分辨率比为 1:8)的 LR 图像与 HR 图像之间的转换;Li 等^[3]通过寻找给定的 1/4 尺度

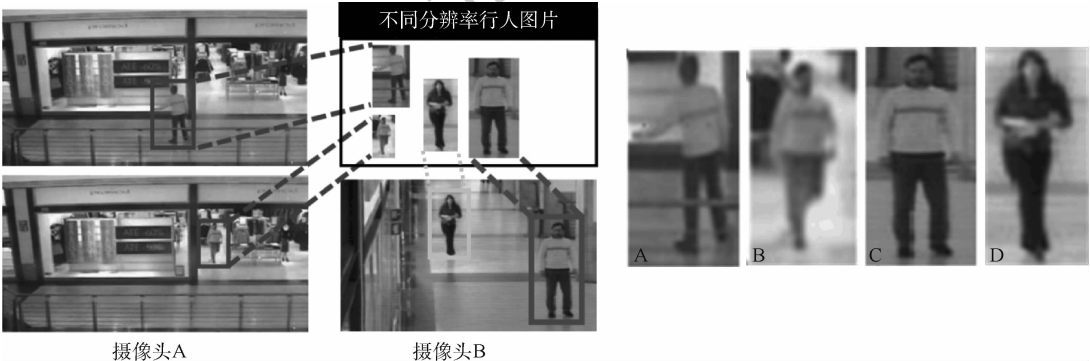


图 1 跨分辨率行人图像
Fig. 1 Cross-resolution pedestrian image

的低、高分辨率图像在某特征空间中的对齐关系计算图像对之间的距离,从而判断是否为同一行人,这种做法必须满足同一行人的不同图像应该在某个特征空间具有相似的结构这一假设;Wang 等^[6]提出了将尺度渐变曲线投影到特征空间分类,解决多低分辨率行人重识别问题;Chen 等^[7]通过对低、高分辨率图像的特征分布进行对齐,解决跨分辨率行人重识别问题。但这几种方法只涉及到粗糙的外观信息和身份信息,丰富的高分辨率图像的细粒度细节在学习过程中被丢弃了。

1.2 级联图像超分辨率

利用图像超分辨率可以恢复图像信息,解决细粒度区别信息丢失的问题。Jiao 等^[4]提出了超分辨率和行人身份识别联合学习方法,能够通过增强低分辨率行人图像中对身份识别有用处的高频外表信息解决低分辨率行人重识别问题中的由于分辨率不同带来的信息量差异的问题。Wang 等^[5]提出了级联超分辨网络通过逐步恢复低分辨率图像细节再与图像检索库中的高分辨率图像匹配。Li 等^[8]在低分辨率图像和高分辨率图像中提取的特征表示上添加了一个对抗损失用来学习分辨率不变的表示,同时通过端到端的方式恢复低分辨率输入图像中缺失的细节。尽管这些方法带来了一些性能提升,但它们需要对预先定义的超分辨率模型进行培训,然而实际问题中查询和图库图像之间的分辨率差异通常是未知的,即无法预先定义图像对之间分辨率差异的倍数,并且梯度在这样一个级联的重模型^[18]中反向传播的难度要大得多,因此这类方法存在模型训练效果不佳的问题,直接使用超分辨率模型不太适合 ReID 任务。

1.3 多任务学习

为了解决超分辨率任务和 ReID 任务之间模型效果训练不佳的问题,受到低分辨率人脸识别工作^[9-10]的启发,Cheng 等^[11]提出了一个正则化方法将超分辨率、ReID 任务联系起来,让超分辨率作为 ReID 任务的辅助任务,通过这种有效的结

合训练方式解决训练不佳的问题。然而这种方法也是无差别的恢复低分辨率图像中缺失的信息,不仅使得超分辨率任务的计算量庞大,而且低分辨率图像中的部分不利于识别的信息被恢复也给行人匹配引入了干扰。Mao 等^[12]通过区分行人图像的前景和背景信息解决了此问题,但这种简单的划分行人为前景和其他物体为背景可能会丢失背景中的有用信息,比如路标、特色建筑、随身物品等。

针对以上问题,本文提出了基于注意力机制的跨分辨率行人重识别方法,利用空间、通道双重注意力机制的特性,得到查询图像中利于身份识别的局部区域,采用动态地预测上采样滤波器权重的方法解决任意放大因子的图像重建任务,精准获取行人图像中缺失的身份识别信息。

2 主要方法

许多前期工作在行人重识别网络前加入了图像超分辨率网络,以恢复低分辨率图像中的信息。这种方法可以通过对低分辨率图像的重建恢复高频细节,但是也放大了其他信息的干扰。受文献^[19]的启发,本文提出基于注意力机制的局部跨分辨率联合身份学习网络,其网络架构主要由 2 部分组成:注意力模块和任意尺度超分辨率模块。通过一个自编码器,跳跃连接训练 ID 损失和注意力损失,逐步学习分辨率不变的特征,得到注意力模块融合得到的前景等显著区域信息,再通过任意尺度超分辨率模块对该部分进行重建,输出特征以此来计算交叉熵损失。

与文献^[20]不同的是,本文首先利用编码解码网络学习分辨率不变特征,不同分辨率图像经过编码网络学习得到中间特征 f_c ,经过解码网络得到特征 F ;然后将特征 F 输入到通道注意力模块中经过一系列操作得到权重系数和特征 F' ,特征 F' 输入到空间注意力模块中得到相应的权重系数和融合特征,即一个包含身份区分信息的局部特征图。网络结构模型如图 2 所示,本文将从 2 方面进行具体的阐述。

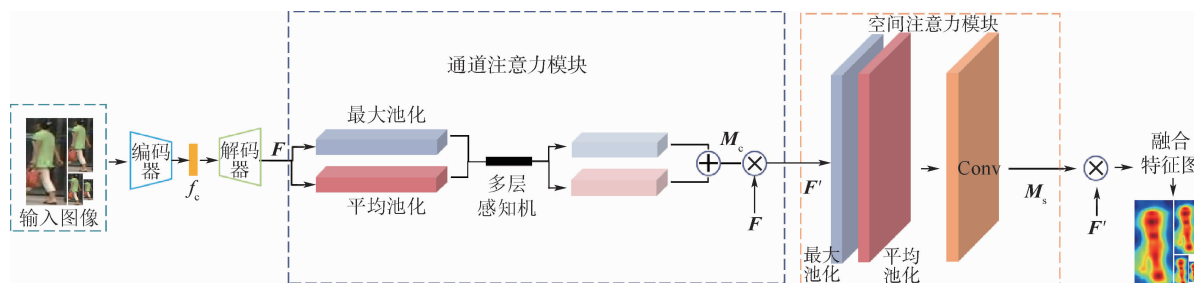


图 2 注意力网络框架

Fig. 2 Framework of attention network

2.1 通道注意力机制

通道注意力的作用是得到有利于识别的特征信息,利用特征的通道间关系生成通道注意图。由于特征图的每个通道都被认为是特征检测器,通道的注意力都集中在对输入图像有意义的地方。为了有效地计算通道注意力,压缩了输入特征图的空间维数,对于空间信息的聚合,一般采用平均池化。具体的流程如图2所示,本文查询图像 x 首先会经过自编码器得到特征 F ,分别经过全局最大池化和全局平均池化,再经过多层感知机输出特征,进行向量相加操作,生成一个通道注意力图 M_c ,与做乘法操作的结果输入到空间注意力模块。 M_c 为

$$M_c(F) = \sigma(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))) = \sigma(W_1(W_0(F_{\text{avg}}^c)) + W_1(W_0(F_{\text{max}}^c)))$$

$$W_0 \in \mathbf{R}^{C/r \times C}, W_1 \in \mathbf{R}^{C \times C/r} \quad (1)$$

式中:AvgPool为全局平均池化;MaxPool为全局最大池化; W 为对应的权重系数;MLP为多层感知机; r 为减少率; σ 为sigmoid函数; C 为通道数; W_1 和 W_0 分别表示2个权重系数; F_{avg}^c 为特征 F 经过全局池化后的特征向量,上标 c 表示通道注意力模块,与下文上标为 s 的空间注意力模块做区分; F_{max}^c 为特征 F 最大池化操作后的向量,2个特征相加后经过一个sigmoid激活生成最终的 M_c 。 W_0 输出后再接ReLU激活,将与本模块输入的特征做乘法操作,生成空间注意力模块需要的输入特征。

2.2 空间注意力机制

将通道注意力模块的输出特征图作为本模块的输入特征图。空间注意是通道注意的补充,如图2右侧所示,首先做一个基于通道的最大池化和平均池化,然后将结果基于通道做连接操作,通过一个卷积降维成单通道,最后激活生成空间特征 M_s ,并与本模块的输入特征做乘法,得到最终的生成特征。通过对通道注意特征的应用池化可以有效地突出显示信息区域^[21]。 M_s 的计算方式为

$$M_s(F) = \sigma(f^{7 \times 7}([\text{AvgPool}(F); \text{MaxPool}(F)])) = \sigma(f^{7 \times 7}([F_{\text{avg}}^s; F_{\text{max}}^s])) \quad (2)$$

式中: f 表示在注意力模块的最后经过 7×7 的卷积操作。空间注意力机制的输出需要与整个模块输入的特征 F 做一个乘法操作,得到最后的融合特征图。

2.3 图像超分辨率

由于复杂的实际场景和无约束的成像条件,

获取到的行人图像并不一定是相近的分辨率,往往获得的分辨率跨度比较大,因此无法预先定义一个尺度因子解决所有场景的图像重建问题。对于跨分辨率行人重识别而言,如何将任意分辨率的查询图像转换至与图库图像为同一分辨率是关键。受文献[22-24]启发,采用一个动态的上采样模块代替传统的放大模块^[14],动态的预测上采样滤波器的权重,然后用这些权重生成高分辨率图像,即能够以任意的上采样因子放大任意的查询图像。

首先通过特征学习^[25]模块提取到特征,对于超分辨率图像中的每一个像素,都是由查询图像在像素上的特征和对应的滤波器权重决定的。通过 F_{LR} 和 I_{SR} 之间的映射函数—上采样模块,得到最终的超分辨率图像。

上采样模块需要一个特定的卷积核或滤波器映射 (i_1, j_1) 和 (i, j) 的值,映射函数如下:

$$I_{\text{SR}}(i, j) = f(F_{\text{LR}}(i_1, j_1), w(i, j)) \quad (3)$$

式中: $I_{\text{SR}}(i, j)$ 为超分辨率图像在 (i, j) 的像素值; $f(\cdot)$ 表示计算像素值的特征映射函数; $w(i, j)$ 为像素点 (i, j) 的权重预测模块(与式(5)相对应); $F_{\text{LR}}(i_1, j_1)$ 表示在低分辨率图像中像素点 (i_1, j_1) 的特征向量。

对于超分辨率图像中的每个像素 (i, j) ,可以通过一个投影转换函数 T 得到:

$$(i_1, j_1) = T(i, j) = (\lfloor i/s \rfloor, \lfloor j/s \rfloor) \quad (4)$$

具体的可以看作一种可变步长机制,比如说当尺度因子 s 为2时,一个 (i_1, j_1) 像素决定超分辨率图像上的2个点。若尺度因子为非整数的1.5,则一些像素决定2个像素,一些像素决定一个像素。无论如何,每一个超分辨率图像上的像素都能找到一个 (i_1, j_1) 。

确定查询图像和超分辨率图像之间的位置关系后还需要得到两者之间特定的权重以及偏移量,可以通过如下公式得到:

$$W(i, j) = \varphi(v_{ij}; \theta) \quad (5)$$

$$v_{ij} = (i/s - \lfloor i/s \rfloor, j/s - \lfloor j/s \rfloor) \quad (6)$$

式中: $W(i, j)$ 为超分辨率图像上像素 (i, j) 对应的卷积核权重; v_{ij} 为和 (i, j) 关联的向量; φ 为权重预测网络; θ 为权重预测网络的权重。

最后需要获取 (i_1, j_1) 像素点的像素值。其特征映射表述为

$$\Phi(F_{\text{LR}}(i_1, j_1), W(i, j)) = F_{\text{LR}}(i_1, j_1) W(i, j) \quad (7)$$

将输出的超分辨率图像与图库图像输入到基线网络^[26]中得到最后的匹配结果。

3 实 验

3.1 实验环境和数据集

实验是在 2 块 TITAN Xp GPU 上进行的。该网络基于 Pytorch 框架,网络基本结构为 Res-Net^[26],基线网络参考了文献[12],并使用 Adam 优化器优化参数并将原始学习速率设置为 10^{-3} 。通过 3 种主流的数据集对本文方法进行评价:Market1501^[27]、CUHK03^[28-29] 和 CAVIAR^[30]。首先对这 3 个数据集以及相应评价标准进行说明。

MLRMarket1501: 数据集包括了来自 6 个不同摄像机拍摄的 1 501 个行人。使用了 DPM 方法,将视频中的行人裁剪出来。数据集划分为训练集和测试集,其中训练集有 751 人,共 12 936 张图片;测试集有 750 人,查询图像 3 368 张,图库图像 19 732 张。然而该数据集所有图像分辨率被处理至统一的大小 128×64 。因此通过下采样的方法将数据集中的图像处理为原尺度的 1、1/2、1/4、1/8、1/16 五种尺度。

MLRCUHK03: 在实验中采用了新的数据集协议,新协议将 CUHK03 数据集分为类似于 Market1501 的划分方法,将来自 10 个摄像头的 1 467 个行,划分为由 767 个身份和 700 个身份组成的训练集和测试集。数据集提供 2 种标注:第 1 种是人类手工标注行人框,第 2 种是通过 DPM 方法检测得到行人框。虽然数据集中图像分辨率是多样的,但分辨率跨度不大,且其中尺度较低的图像相较于低分辨率数据集分辨率偏高,因此通过下采样的方法将数据集中的图像处理至原尺度的 1、1/2、1/4、1/8、1/16 五种尺度。

CAVIAR: 数据集包含由 2 台摄像机捕获的 72 个身份的 1 220 张图像。丢弃了 22 个只出现在相机中的人,将剩下的人分成 2 部分,这 2 部分的身份标签没有重叠。

3.2 结果分析

在本文实验中,为了评价 Re-ID 的方法,计算出所有候选数据集的累积匹配曲线^[27] (Cumulative Matching Characteristics, CMC) 的 Rank1 和 Rank5。表 1 和表 2 展示了本文方法与主流的跨分辨率行人重识别方法在 3 个数据集上的实验结果对比。表中最佳的 2 个结果分别用加粗和下划线形式突出显示。

在 Market1501 数据集上,本文方法和其他主流方法的定量结果对比如表 1 所示。对比近年来效果较好的行人重识别方法^[31],以及大部分跨分辨率行人重识别方法,其中 INTACT 为近年来处

表 1 现有方法在 Market1501 和 CUHK03 数据集上的定量结果对比

Table 1 Quantitative result comparison of existing methods on Market1501 和 CUHK03 datasets %

方法	Market1501		CUHK03	
	Rank1	Rank5	Rank1	Rank5
FD-GAN ^[31]	79.6	91.6	73.4	93.8
JUDEA ^[3]			26.2	58.0
SDF ^[6]			22.2	48.0
SING ^[4]	74.4	87.8	67.7	90.7
CSR-GAN ^[5]	76.4	88.5	71.3	92.1
RAIN ^[7]			78.9	97.3
CAD ^[8]	83.7	92.7	82.1	97.4
INTACT ^[11]	<u>88.1</u>	95.0	<u>86.4</u>	<u>97.4</u>
RIPR ^[12]	66.9	84.7	73.3	92.6
本文	90.2	<u>94.3</u>	89.2	97.5

表 2 现有方法在 CAVIAR 数据集上的定量结果对比

Table 2 Quantitative result comparison of existing methods on CAVIAR dataset %

方法	Rank1	Rank5	Rank10
FD-GAN ^[31]	32.3	72.3	85.9
SLD ² L ^[2]	18.4	44.8	61.2
JUDEA ^[3]	22.0	60.1	80.8
SDF ^[6]	14.3	37.5	62.5
SING ^[4]	33.5	72.7	89.0
CSR-GAN ^[5]	34.7	72.5	87.4
RAIN ^[7]	42.0	77.3	89.6
CAD ^[8]	42.8	76.2	91.5
INTACT ^[11]	<u>44.0</u>	<u>81.8</u>	93.9
RIPR ^[12]	36.4	72.0	90.0
本文	49.3	83.7	<u>91.2</u>

理跨分辨率行人重识别问题的性能最优的方法,该方法为基于多任务学习的方法,通过正则化改进方法,使得模型训练更简单,其 Rank1 和 Rank5 分别为 88.1% 和 95.0%,而本文方法获得了 90.2% 和 94.3%,Rank1 准确率提高了 2.1%,优于前面对比的方法。除了客观数据的对比,还在 Market1501 数据集上分别进行了实验,由图 3 可以看出 RIPR^[12] 中几个不匹配样本的例子,使用本文方法得到了解决。图 3 中三角标“△”是指不匹

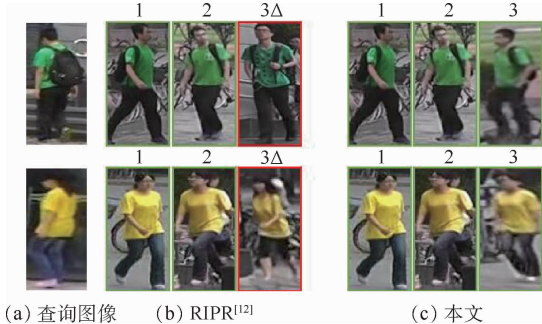


图 3 各模型主观性能对比

Fig. 3 Subjective performance comparison of various models

配的样本,“1”、“2”和“3”对应于前 3 个检索到的图库样本。

在 CUHK03 数据集上,本文方法使用了新的 CUHK03 协议^[29]进行训练,由于 CUHK03 数据集更接近真实情况,图像是来自于几个月来录制的一系列视频,无约束的成像条件也导致了图像分辨率的多样性。因此本文方法在此数据集上取得了比 MLRMarket1501 数据集更大的进步。同基于局部超分辨率 RIPR^[12]方法对比,本文方法 Rank1 提高了 15.9%,进一步说明了只恢复前景区域会丢失背景中有利于识别的信息。相较于性能最优的 INTACT 方法 Rank1 和 Rank5 分别提高了 2.8% 和 0.1%。这些数据都证明了本文方法对在跨分辨率场景的有效性。

CAVIAR 数据集是早期行人重识别主流数据集,由于早期的摄像头性能较差,且摄像头分布并不密集,采集数据也比较困难,数据集内图像数量较少,且质量较差,一般的行人重识别方法在这一真实的数据集上性能很差。因此本文方法在此数据集上取得了比上述数据集更大的进步:①同基于单一低分辨率的方法 SLD²L、JUDEA 对比,本文方法 Rank1 提高了 30.9% 和 27.3%,体现了任意上采样因子的超分辨率重建在跨分辨率行人重识别方法中的优越效果;②与基于级联超分辨率的方法 CSR-GAN、SING 相比,本文方法 Rank1 提升了 14.6% 和 15.8%,进一步说明了任意上采样因子的超分辨率重建和级联超分辨率重建相比的有效性;③CAD 是基于分辨率不变表示的方法,可以处理未训练过的尺度,本文方法在 Rank1 上也提高了 6.5%,这有力地证明了基于注意力机制的方法和任意尺度因子的超分辨率重建相结合应用于行人重识别方法上对于跨分辨率问题的有效性。

3.3 消融实验

为了验证本文方法的注意力模块和图像超分辨率模块在跨分辨率行人重识别问题上的有效性,采用不同的模型参数进行训练,使用单一查询模式在 CUHK03 数据集上进行实验。如表 3 所示,ResNet50 为基线模型,CAM 为通道注意力模块,SAM 为空间注意力模块,Non-Local^[32]、SENet^[20]是常用于图像分类领域的 2 个易于集成的注意力模块,对于图像分类有一定的作用,MASR 表示任意放大因子的图像超分辨模块。所有的实验均只改变了一个设置,其余设置与本文方法的设置相同。

首先,通过表 3 中不同注意力模块的性能对比验证通道注意力模块和空间注意力模块联合作

表 3 各模块消融实验结果

Table 3 Ablation experimental results of each module

模块	MLRCUHK03	
	Rank1	Rank5
ResNet50	58.1	79.3
ResNet50 + CAM	63.2	85.0
ResNet50 + SAM	65.0	87.1
ResNet50 + NonLocal	70.7	87.7
ResNet50 + SENet	71.3	89.1
ResNet50 + CAM + SAM	78.9	90.3
ResNet50 + MASR + ID	83.3	93.1
本文	89.2	97.5

用的有效性,单独使用通道注意力模块或者空间注意力模块,都显著地提高了跨分辨率行人重识别的性能,当它们联合使用时 Rank1 比主流的注意力机制 Non-Local、SENet 分别提升了 8.2%、7.6%。同时,由于空间注意力机制将图像中的空间域信息做对应的空间变换并保留关键信息,进一步挖掘了图像中显著且关键的区域,相比于通道注意力机制更有助于模型鲁棒性的提高,对于模型性能有更加显著的提升。

4 结束语

本文提出了一个用于跨分辨率行人重识别的方法。使用基于通道注意力机制和空间注意力机制相融合的方法来获取更利于识别的特征和区域,同时解决了不同分辨率查询图像关注区域不同的问题。网络中任意上采样模块在超分辨率重建上的应用极其有力地解决了图像信息恢复过程中由于分辨率多样性导致的网络级联训练困难、计算量大、模型复杂等问题,从而使本文方法适用于更广泛的场景。这 2 个互补的模块被联合训练来优化行人重识别方法,在 3 个公开的数据集上与大量先进方法对比取得了最优或次优的效果,充分证明了本文引入的模块的有效性。

参考文献 (References)

[1] ZAJDEL W,ZIVKOVIC Z,KROSE B J A. Keeping track of humans: Have I seen this person before[C]//Proceedings of the 2005 IEEE International Conference on Robotics and Automation. Piscataway:IEEE Press,2005:2081-2086.

[2] JING X Y,ZHU X K,FEI W,et al. Super-resolution person re-identification with semi-coupled low-rank discriminant dictionary learning[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2015:695-704.

[3] LI X,ZHENG W S,WANG X J,et al. Multi-scale learning for

- low-resolution person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 3765-3773.
- [4] JIAO J N, ZHENG W S, WU A C, et al. Deep low-resolution person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 6967-6974.
- [5] WANG Z, YE M, YANG F, et al. Cascaded SR-GAN for scale-adaptive low resolution person re-identification[C]//Proceedings of the International Joint Conference on Artificial Intelligence, 2018: 3891-3897.
- [6] WANG Z, HU R M, YU Y, et al. Scale-adaptive low-resolution person re-identification via learning a discriminating surface [C]//Proceedings of the International Joint Conference on Artificial Intelligence, 2016: 2669-2675.
- [7] CHEN Y C, LI Y J, DU X F, et al. Learning resolution-invariant deep representations for person re-identification[C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 8215-8222.
- [8] LI Y J, CHEN Y C, LIN Y Y, et al. Recover and identify: A generative dual model for cross-resolution person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 8090-8099.
- [9] CHENG Z Y, ZHU X T, GONG S G. Low-resolution face recognition[C]//Proceedings of the Asian Conference on Computer Vision, 2018: 605-621.
- [10] LU Z, JIANG X D, ALEX K. Deep coupled resnet for low-resolution face recognition[J]. IEEE Signal Processing Letters, 2018, 25(4): 526-530.
- [11] CHENG Z Y, DONG Q, GONG S G, et al. Inter-task association critic for cross-resolution person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2020: 2605-2615.
- [12] MAO S N, ZHANG S L, YANG M. Resolution-invariant person re-identification[C]//Proceedings of the International Joint Conference on Artificial Intelligence, 2019: 883-889.
- [13] ZHAO R, OUYANG W L, WANG X G. Person re-identification by salience matching[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2014: 2528-2535.
- [14] HU X K, MU H Y, ZHANG X Y, et al. Meta-SR: A magnification-arbitrary network for super-resolution[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019: 1575-1584.
- [15] WANG Z, JIANG J, WU Y, et al. Learning sparse and identity-preserved hidden attributes for person re-identification[J]. IEEE Transactions on Image Processing, 2019, 29(1): 2013-2025.
- [16] ZENG Z, WANG Z, WANG Z, et al. Illumination-adaptive person re-identification[J]. IEEE Transactions on Multimedia, 2020, 22(12): 3064-3074.
- [17] WANG Z, WANG Z, ZHENG Y, et al. Beyond intra-modality: A survey of heterogeneous person re-identification[EB/OL]. (2020-04-21)[2020-07-14]. <https://arxiv.org/abs/1905.10048>.
- [18] SARATH C, CHINNADHURAI S, EUGENE V, et al. Towards non-saturating recurrent units for modelling long-term dependencies[EB/OL]. (2020-07-14)[2020-08-01]. <https://arxiv.org/abs/1902.06704>.
- [19] SANGHYUN W, JONGCHAN P, JOONYOUNG L, et al. CBAM: Convolutional block attention module[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2018: 3-19.
- [20] HU J, SHEN L, SUN G. Squeeze and excitation networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2018: 7132-7141.
- [21] KOMODAKIS N, ZAGORUYKO S. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[EB/OL]. (2011-02-12)[2020-07-12]. <https://arxiv.org/abs/1612.03928>.
- [22] IAN G, JEAN P A, MEHDI M, et al. Generative adversarial nets[C]//Proceedings of the Advances in Neural Information Processing Systems, 2014: 2672-2680.
- [23] CHRISTIAN L, LUCAS T, FERENC H, et al. Photo-realistic single image super-resolution using a generative adversarial network[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 4681-4690.
- [24] ZHANG Y L, LI K P, LI K, et al. Image super-resolution using very deep residual channel attention networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2018: 286-301.
- [25] ZHANG Y, TIAN Y, KONG Y, et al. Residual dense network for image super-resolution[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2018: 2472-2481.
- [26] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2016: 770-778.
- [27] ZHENG L, SHEN L Y, TIAN L, et al. Scalable person re-identification: A benchmark[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1116-1124.
- [28] LI W, ZHAO R, XIAO T, et al. DeepReID: Deep filter pairing neural network for person re-identification[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2014: 152-159.
- [29] ZHONG Z, ZHENG L, CAO D, et al. Re-ranking person re-identification with k -reciprocal encoding[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017: 3652-3661.
- [30] CHENG D S, CRISTANI M, STOPPA M, et al. Custom pictorial structures for re-identification[C]//British Machine Vision Conference, 2011: 6.
- [31] GE Y X, LI Z W, ZHAO H Y, et al. FD-GAN: Pose-guided feature distilling gan for robust person re-identification[C]//Proceedings of the Advances in Neural Information Processing Systems, 2018: 1222-1233.
- [32] WANG X S, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2018: 7794-

7803.

识别。

作者简介:

廖华年 女,硕士研究生。主要研究方向:计算机视觉、行人重

徐新 男,博士,教授,博士生导师。主要研究方向:计算机视觉、机器学习、行人重识别。

Cross-resolution person re-identification based on attention mechanism

LIAO Huanian¹, XU Xin^{1,2,3,*}

(1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;

2. Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China;

3. Shenzhen Institute, Wuhan University, Shenzhen 518000, China)

Abstract: The resolution variation of person images poses great challenges to current person re-identification methods. To address this problem, this paper presents a cross-resolution person re-identification method. This method solves the resolution variation from two aspects. On the one hand, the spatial and channel attention mechanisms are utilized to capture person features and obtain local region; On the other hand, local information of any resolution image is recovered by the nuclear dynamic upsampling module. Comparative experiments have been conducted to verify the effectiveness of the proposed method against state-of-the-art methods on Market1501, CUHK03, and CAVIAR person re-identification datasets. The experimental results show that the proposed method has the best performance.

Keywords: person re-identification; channel attention mechanism; spatial attention mechanism; image super resolution; up-sampling

Received: 2020-08-28; Accepted: 2020-09-18; Published online: 2020-11-06 11:06

URL: kns.cnki.net/kcms/detail/11.2625.V.20201105.1153.003.html

Foundation items: National Natural Science Foundation of China (U1803262,61602349,61440016); Basic Research Project of Science and Technology Plan of Shenzhen (JCYJ20170818143246278)

* Corresponding author. E-mail: xuxin0336@163.com

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

DOI: 10.13700/j.bh.1001-5965.2020.0445

基于多尺度联合学习的行人重识别



谢彭宇¹, 徐新^{1,2,3,*}

(1. 武汉科技大学 计算机科学与技术学院, 武汉 430065;

2. 武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室, 武汉 430065;

3. 上海交通大学 电子信息与电气工程学院, 上海 200240)

摘 要: 现有的行人重识别方法主要关注于学习行人的局部特征来实现跨摄像机条件下的行人辨识。然而在人体部件存在运动或遮挡、背景干扰等行人数据非完备条件下,会导致行人局部辨识信息丢失概率的增加。针对这个问题,提出了一种多尺度联合学习方法对行人辨识特征进行精细化表达。该方法包含3个分支网络,分别提取行人的粗粒度全局特征、细粒度全局特征和细粒度局部特征。其中粗粒度全局分支通过融合不同层次的语义信息来增强全局特征的丰富性;细粒度全局分支通过联合全部局部特征,在对全局特征进行细粒度描述的同时学习行人局部部件间的相关性;细粒度局部分支则通过遍历局部特征来挖掘行人非显著性的信息以增强局部特征的鲁棒性。为了验证所提方法的有效性,在Market1501、DukeMTMC-ReID和CUHK03三个公开数据集上开展了对比实验,实验结果表明:所提方法取得了最佳性能。

关 键 词: 行人重识别; 多尺度; 联合学习; 多分支网络; 深度学习

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2021)03-0613-10

近年来中国安防形势日趋严峻,各类大案需要对嫌疑目标进行跨摄像机的检索、分析、比对。行人重识别作为视频侦查的关键技术以处理跨摄像机的行人外观差异为基础,旨在从多摄像机条件下检索特定身份的行人。因此,探索具备跨摄像机行人检索能力的行人重识别方法具有重要意义。

随着深度学习^[1]的迅速发展,行人重识别方法对行人的表征能力得到了极大提升。然而实际监控场景中的行人姿态、遮挡、背景等方面存在着极大差异,给现有行人重识别方法带来了极大挑战。因此,现有行人重识别方法主要关注于提取行人的局部部件特征,进而对行人进行精细化表示。这些方法主要分为两类:第一类方法利用人

体姿态估计模型,分别对各个行人部件进行精细化表达。基于此,Zhao等^[2]设计了一个树形融合网络,通过选取各个部件最显著的特征进行融合,以增强部件级特征的鉴别性。然而这类方法首先需要图像进行预处理,增加了方法的复杂性。并且这类方法高度依赖于姿态估计模型的鲁棒性,错误的估计会大大影响识别的结果。第二类方法则假定行人图像已对齐,进而将特征图裁剪为条状或者分块。Sun等^[3]通过把网络输出的特征图进行水平划分,从而学习不同条状区域的局部特征。此类方法通过提取行人的局部部件特征实现了精细化的行人特征表示。以上方法忽略了部件之间的联系,缺乏行人结构性信息。在摄像

收稿日期: 2020-08-24; 录用日期: 2020-09-19; 网络出版时间: 2020-10-16 16:43

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201016.1507.002.html

基金项目: 国家自然科学基金(U1803262,61602349,61440016)

* 通信作者。E-mail: xuxin0336@163.com

引用格式: 谢彭宇, 徐新. 基于多尺度联合学习的行人重识别[J]. 北京航空航天大学学报, 2021, 47(3): 613-622.

XIE P Y, XU X. Multi-scale joint learning for person re-identification[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 613-622 (in Chinese).

机拍摄角度受限、人体部件存在运动或遮挡等行人数据非完备条件下,会导致行人局部辨识特征信息丢失概率的增加。

针对这个问题,近年来研究者开始尝试结合行人的局部和全局特征。例如,Zheng 等^[4]和 Wei 等^[5]在使用姿态估计模型提取局部特征的基础上,结合全局特征来避免模型误检对识别精度的影响。Zheng 等^[6]和 Fu 等^[7]通过金字塔模型对特征图进行不同尺度的划分,从而获得多尺度的部件信息。Wang 等^[8]则通过一个全局分支和 2 个局部分支,对行人进行多粒度描述。这些方法通过结合离散的局部特征与全局特征,能提高行人辨识的性能。然而,实际监控场景中的遮挡、背景等因素会对行人信息带来干扰,进而降低行人的全局特征和局部部件特征的辨识性能。

本文的贡献主要包括 3 个方面:

- 1) 提出了一种多尺度联合学习方法,该方法结合了行人的粗粒度全局特征、细粒度全局特征和细粒度局部特征。
- 2) 结合细粒度全局特征和细粒度局部特征能够有效学习行人局部部件间的相关性并挖掘非显著信息,从而提高遮挡、背景差异等条件下的行人辨识性能。
- 3) 通过在 3 个行人重识别数据集上的实验测试,综合比较了本文方法与 12 种目前的主流模型的性能。

1 相关工作

由于在真实场景下背景^[9]、姿态、照明^[10]、视角、相机^[11]等条件变化很大,行人重识别在计算

机视觉中是一项十分具有挑战性的任务。图 1 展示了在真实场景下受遮挡的相似行人图像。以前,人们通过手工特征,如颜色、HOG^[12]等,通过 XQDA^[13]或 KISSME^[14]来学习最佳的相似度量。然而,传统的手工特征描述能力有限。近年来随着深度学习的兴起,深度网络的特征学习已成为行人重识别的常见做法。Zheng 等^[15]提出的 IDE (ID-discriminative Embedding) 把 ResNet-50^[16]做为骨干网络,将行人重识别视为分类问题,通过 ID-Loss 训练。在此之后,一系列基于深度学习的行人重识别方法被提出。最近一些研究工作意识到行人局部特征有助于行人精细化表示。

1.1 局部特征

通过总结,基于局部特征主要分为以下 2 种方式:①通过预定义区域对行人进行划分,基于人体姿态估计的方法获得身体部件。Zhao 等^[2]根据姿态估计模型定位的部件区域,对图像进行部件分割,以提取部件级特征。Su 等^[17]通过学习各个部件级特征的权重,以增强局部特征的鲁棒性。不同于上述方法直接对行人图像进行分割。Suh 等^[18]和 Xu 等^[19]利用双流网络实现行人部件匹配,其中上流网络用来获取不同行人部件的特征区域,并引导下流网络来增强指定区域的关注度。Sarfraz 等^[20]引入了行人的姿态信息进行训练,以获得有鉴别性的特征。然而这种方法由于姿态估计和行人重识别数据集之间的间隙,其精度很大程度上依赖估计器的性能。②假设行人已经对齐,根据人体结构信息进行划分。一些方法^[21]通过把网络输出的特征图进行水平划分获得局部特征,Sun 等^[3]提出了 PCB (Part-based Convolutional Baseline) 以获得精细化描述。为了避

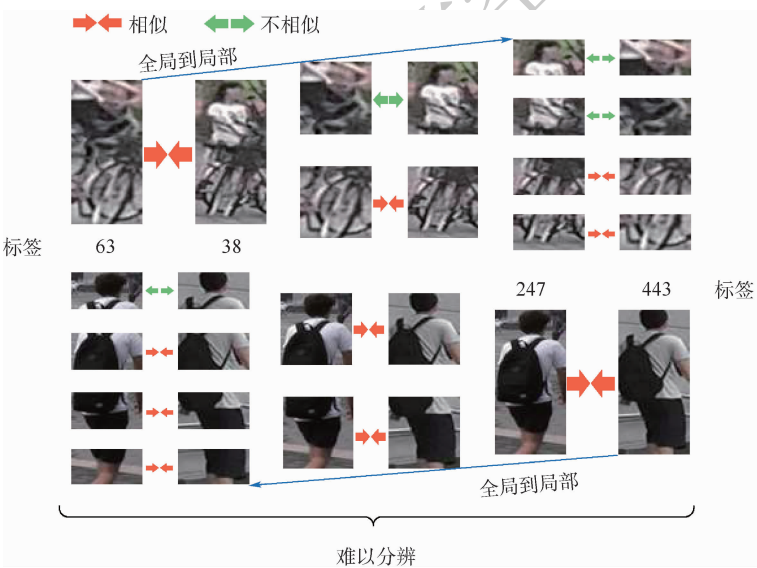


图 1 真实场景下受遮挡的相似行人图像

Fig. 1 Obscured images of similar pedestrians in real scenes

免行人不对齐,导致各个部件之间产生误匹配。Luo 等^[22]基于动态规划的思想,通过计算各局部特征之间的最短路径,对行人部件进行对齐。Sun 等^[23]提出了部件感知模型,避免由于部件被遮挡产生的噪声。本文方法属于此类,与其他方法不同,考虑了身体部位之间的相关性,而不是直接的使用局部特征。

1.2 联合特征

虽然局部特征能对行人进行精细化表示,但无法对行人显著性特征进行描述。还有一些方法^[24-25]联合行人的局部和全局特征,增强行人特征的完整性。Zheng 等^[4]和 Wei 等^[5]通过提取行人较粗部件特征并联合全局特征提高模型的鲁棒性。Li 等^[24]设计了一个注意力模块以增强联合特征的鉴别性。基于 PCB, Wang 等^[8]提出了 MGN(Multiple Granularity Network)通过多分支网络,将全局特征和局部特征相结合,对行人进行多粒度描述。Zheng 等^[6]和 Fu 等^[7]则基于金字塔模型,通过不同尺度的水平划分来学习多尺度的局部和全局特征。

2 多尺度联合学习方法

多尺度联合学习网络由粗粒度全局分支、细粒度全局分支和细粒度局部分支构成。粗粒度全局分支用于增强全局特征的丰富性。而细粒度全局分支在对全局特征精细化描述的同时又学习了相邻部件之间的相关性。细粒度局部分支则通过遍历组合局部特征,加强学习局部特征

之间的非显著性信息。图 2 展示了多尺度联合学习网络的框架结构,对其框架和各个模块的详细说明如下。

2.1 粗粒度全局分支

给定一组图像 $X = \{a_1, a_2, \dots, a_k\}$ 为监控系统中摄像头所捕获的人, k 为图像的数量。使用 ResNet-50^[16] 作为多尺度联合学习的骨干网络,不同于其他方法只提取网络的高级特征对行人进行描述。考虑到网络不同阶段的输出所带来的语义差异,通过融合不同层次的语义,增强行人全局特征的丰富性。笔者做法如下:对于粗粒度全局分支,分别将 res_conv4x 和 res_conv5x 层的特征图进行全局平均池化(GAP)和全局最大池化(GMP)操作,分别得到全局特征 $d_{\max, \text{avg}}$ 和 $f_{\max, \text{avg}}$, 之后将其数值相加得到特征向量 D 和 F :

$$D = d_{\max} + d_{\text{avg}} \tag{1}$$

$$F = f_{\max} + f_{\text{avg}} \tag{2}$$

对于 $d_{\max, \text{avg}}$ 和 $f_{\max, \text{avg}}$, 将这 4 个特征向量分别通过三元组损失进行训练。特征向量 D 和 F 分别通过 1×1 的卷积减少至 512 维,并通过 softmax 函数进行训练。

2.2 细粒度全局分支

通过对特征图进行水平划分,以获得行人的精细化描述是一种常见的方式。在现有的做法中,只对每个局部特征单独的进行学习,并将其连接起来,以产生行人的描述。虽然这种方式可以使行人获得更精细化的表示,但由于行人图像中存在不对齐的现象,容易使各个部件之间产生误

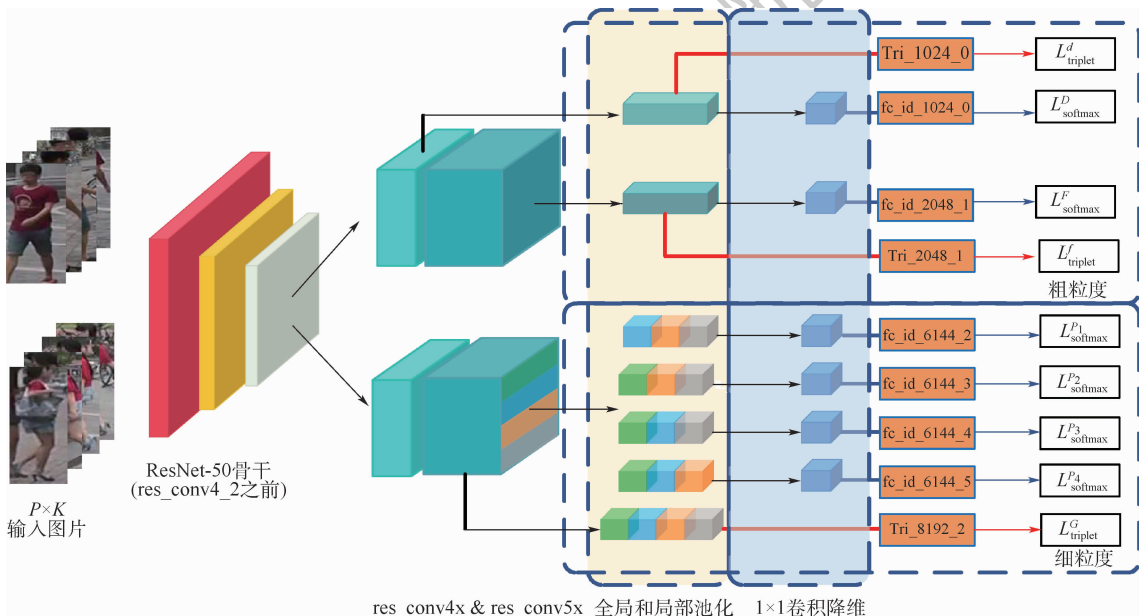


图 2 多尺度联合学习网络框架

Fig. 2 Multi-scale joint learning network framework

匹配。更重要的是,由于各个局部特征是独立存在的,虽然获得了精细化的行人描述,但是缺少了行人特征的完整性。这会导致在相同部件具有相似属性的不同行人难以进行区分,造成误判。为此,考虑各个部件之间的相关性,通过对全局特征进行细粒度描述来解决此问题。

具体的,行人图像通过骨干网络可以得到一个大小为 $C \times H \times W$ 的三维特征张量, C 为特征通道数, H 和 W 分别为特征空间的高度和宽度。根据特征图的 H 轴将特征图划分为 n 个部分,每个部分的大小为 $C \times (H/n) \times W$,分别对每部分进行平均和最大池化操作,可以得到长度为 C 的特征向量 $\mathbf{g}_{\max i}$ ($i = 1, 2, \dots, n$) 和 $\mathbf{g}_{\text{avg} i}$ ($i = 1, 2, \dots, n$)。将 $\mathbf{g}_{\max i}$ 和 $\mathbf{g}_{\text{avg} i}$ 分别连接起来,得到了长度为 $n \times C$ 的向量 $\mathbf{G}_{\max}$ 和 $\mathbf{G}_{\text{avg}}$ 。通过将局部特征互相关联,使其在既保证全局特征完整性的同时,又对行人特征进行细粒度描述,对于 $\mathbf{G}_{\max}$ 和 $\mathbf{G}_{\text{avg}}$ 使用三元组损失^[26]进行训练。通过考虑各部件之间的相关性,以缓解部件之间的误匹配,并增强相似部件之间的判别。

2.3 细粒度局部分支

经过划分的局部特征,通过身体各个部件之间相互联系,可以增强局部特征的鉴别力。虽然细粒度全局分支将2个相邻部件之间互相关联,以学习到相邻部件之间的相关性。但是,对于不相邻的2个部件,是否也存在着潜在的相关性。此外,由于部件相互之间间隔的尺度不同,就会形成不同尺度的局部特征。那么对于局部特征在什么尺度下,能够 and 全局特征进行有效联合,仍需要进一步研究。

笔者的做法如下:对于长度为 C 的特征向量 $\mathbf{g}_{\max i}$ ($i = 1, 2, \dots, n$) 和 $\mathbf{g}_{\text{avg} i}$ ($i = 1, 2, \dots, n$),将 n 设置为4,并分别将每个 $\mathbf{g}_{\max i}$ 和 $\mathbf{g}_{\text{avg} i}$ 数值相加得到局部特征向量 $\mathbf{g}_i$ ($i = 1, 2, \dots, n$)。

$$\mathbf{g}_i = \mathbf{g}_{\max i} + \mathbf{g}_{\text{avg} i} \quad (3)$$

为了挖掘不相邻的局部特征之间相关性,分别对 $\mathbf{g}_i$ 进行丢弃,根据丢弃的 $\mathbf{g}_i$ 的不同,可以得到多组包含不同 $\mathbf{g}_i$ 的局部特征。假设每次丢弃

1个 $\mathbf{g}_i$,并对 $\mathbf{g}_i$ 进行遍历的丢弃,那么就会获得4组由不同的 $\mathbf{g}_i$ 组成的局部特征向量 $\mathbf{P}_n$ ($n = 1, 2, 3, 4$)。

$$\mathbf{P}_n = \{\mathbf{g}_i | i = 1, 2, 3, 4 \text{ and } i \neq n\} \quad (4)$$

对于每组 $\mathbf{P}_n$ 将其分别通过多部件相关性进行训练。图3为细粒度局部分支丢弃1个 $\mathbf{g}_i$ 时的示意图。

对于每组 $\mathbf{P}_n$,由于都缺少了行人的某个关键部件,当对 $\mathbf{P}_n$ 挖掘相关性时,就会使原本不相邻的2个局部特征之间相互连接。通过利用不相邻部件之间的相关性,从而可以挖掘更多关键但非显著的潜在信息。基于此,进一步探索了局部尺度特征和全局尺度特征联合判别的有效性。具体的,对于每组 $\mathbf{P}_n$ 分别通过改变丢弃的尺度进行训练。通过实验可以发现,不论丢弃尺度为多少细粒度局部特征都有助于提高精确度。但由于不同数据集行人图像的边界框(Bounding Box)质量不同,不同尺度的特征嵌入的有效性也不相同,具体实验将在3.4节进行详细阐述。

2.4 损失函数

为了提高该网络学习行人特征表示的辨别能力,采用了基于 softmax 的 ID-Loss 和最大三元组损失作为训练阶段的损失函数,这2种函数被广泛的应用于各种行人重识别的方法。

首先将行人重识别当作一个分类任务。因此,笔者的目标是预测每个人的标签,通过优化分类模型学习行人的具体特征表示。对于第 β 个学习到的特征 $\mathbf{f}_\beta$, softmax loss 阐述如下:

$$L_{\text{softmax}} = - \sum_{\beta=1}^N \ln \frac{\exp(W_y^T \mathbf{f}_\beta)}{\sum_{k=1}^L \exp(W_k^T \mathbf{f}_\beta)} \quad (5)$$

式中: W_k 表示对应于第 k 类的权重;在训练中一个训练批次数量为 N ; W_y^T 表示对应于第 y (ground truth) 类的权重;训练数据集的分类数为 L 。不同于传统的 softmax,抛弃了线性多分类中的 bias^[27],这有助于获得更好的分类性能。在训练的过程中将该损失用于全局特征 $\mathbf{D}$ 、 $\mathbf{F}$,以及局部特征 $\mathbf{P}_n$ ($n = 1, 2, 3, 4$)。

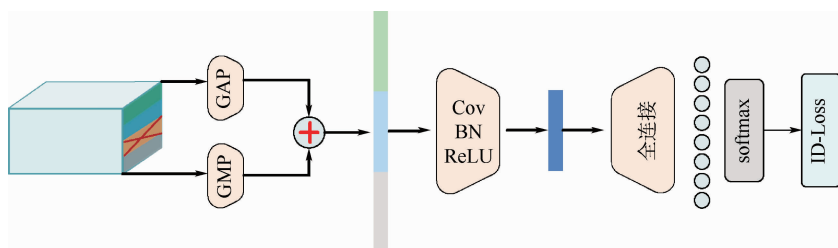


图3 细粒度局部分支

Fig.3 Fine-grained local branch

对于所有的全局特征以及不进行丢弃的局部相关性特征 $\{d_{\max}, d_{\text{avg}}, f_{\max}, f_{\text{avg}}, G_{\max}, G_{\text{avg}}\}$ 通过三元组损失训练来提高性能。使用 hard triplet-loss^[28] 阐述如下:

$$L_{\text{triplet}} = - \sum_{b=1}^P \sum_{a=1}^K [\alpha + \max_{p=1,2,\dots,K} \|f_a^b - f_p^b\|_2 - \min_{\substack{q=1,2,\dots,K \\ c=1,2,\dots,P \\ c \neq b}} \|f_a^b - f_n^c\|_2] + \quad (6)$$

f_a^b , f_p^b 和 f_n^c 分别为通过锚 (Anchor)、积极 (Positive) 和消极 (Negative) 进行采样得到的特征,积极和消极是代表和锚有相同标签和不同标签的行人; α 为控制类之间的距离。候选的三元组由距离最远的有相同标签的对和最近的有不同标签的对构成。最难的相同标签样本和不同标签样本分别在一个批次中,这个批次有 P 个标签的行人,每个标签有 K 张图片。损失函数鼓励最远的相同标签样本的距离小于最近的不同标签样本的距离。

3 实验

3.1 数据集

Market1501^[29] 是在清华大学校园中采集的数据集,图像来自于6个不同的摄像头,其中有一个摄像头为低分辨率。同时该数据集提供训练集和测试集。训练集包含12 936张行人图像,测试集包含19 732张行人图像。图像由检测器自动检测并切割,所以存在一些检测误差。训练数据中一共有751人,测试集中有750人。所以在训练集中,平均每类(每个人)有17.2张训练数据。

CUHK03^[30] 是在香港中文大学采集,行人图像来自2个不同的摄像头。该数据集提供机器自动检测和手动检测2个数据集。其中检测数据集包含一些检测误差,更接近实际情况。数据集总共包括1 467个行人的14 097张图片,平均每个人有9.6张训练数据。最初整个数据集被划分为20个随机分组用于交叉验证,但这是针对手工制作的方法而设计的。因此,笔者采用提出的新培训/测试协议^[31]。

DukeMTMC-ReID^[32] 是在杜克大学内采集,图像来自8个不同摄像头,行人图像的边框由人工标注完成。训练集包含16 522张图像,测试集包含17 661张图像。训练数据中一共有702人,平均每个人有23.5张训练数据。

以上3个数据集,采用目前行人重识别方法普遍使用的首中准确率 (Rank-1) 和平均匹配度

(mAP) 2种评价指标评估方法的性能。所有实验都会使用单一查询方式。此外,为简单起见不使用显著改善 mAP 的重新排序算法。

3.2 实验设置

为了从行人图像中获得全局和局部足够的信息,将所有图像的大小调整为 288×144 , 并使用通过 ImageNet^[33] 分类的 ResNet-50 作为骨干网络。与原始版本不同将 res_conv4_1 块之后的后续部分化为2个独立的分支,并与原始的 ResNet-50 共享相似的体系结构。将最后一个卷积层的 stride 从2设置为1,并通过水平移动,标准化和随机擦除来增强训练图像。批量大小设置为32, $P=8, K=4, \alpha=0.3, n=4$ 。笔者的模型训练220个 epoch。基础学习率设定为0.03,并在60个时期后衰减至0.003,130个时期后衰减至0.0003直至训练结束。在每个批量中使用具有0.9动量的随机梯度下降 (SGD) 方法以更新参数。本文方法在 pytorch 框架上实现,并使用单个 GTX1060 GPU 进行训练,所有数据集共享与上述相同的实验设置。

3.3 现有方法的对比实验

本节中将所提方法与现有的先进的方法进行比较,以表示对于其他方法的优势,这些方法大多都是最近发布的方法,具体情况如表1所示,针对每个数据集详细说明如下。

3.3.1 Market1501 数据集

本文方法在该数据集上实现了95.9% Rank-1 和89.1% mAP。对比仅仅使用了单一局部特征的 PCB^[3], 本文方法分别在 Rank-1 和 mAP 提高了3.6% 和11.7%。MGN 虽然考虑了多个分支结构,但是却忽略了局部信息之间的联系,作为该数据集上最好的方法分别提高了0.2% Rank-1, 2.2% mAP。

图4展示了查询前10名的结果。分别选择了行人被遮挡、背景复杂、图片模糊等复杂条件下情况。第1、2排行人的查询图像在被小包遮挡和背景杂乱的情况下,网络仍然可以健壮的表示其判别信息。第3行查询图像是在低分辨率下进行捕获的,丢失了大量精细的信息,但本文网络却可以通过挖掘局部特征之间的潜在联系,找出正确的行人图像。最后一位行人,背景十分杂乱,身体大部分被自行车遮挡且照片也十分的模糊。但本文方法仍然有较好的效果。可以看到,在 Rank-9 中即使行人出现了较大的不对齐现象,仍然可以将其正确的找出,这得益于细粒度局部分支对网络的影响。

表 1 多尺度联合学习方法和其他方法性能对比

Table 1		Performance comparison of multi-scale joint learning method and other methods								%
方法		CUHK03				Market1501		DukeMTMC-ReID		
		Labeled		Detected						
		Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	
基于部件	IDE ^[15]	22.0	21.0	21.3	19.7	72.5	46.0	67.7	47.1	
	MGN ^[8]	68.0	67.4	66.8	66.0	95.7	86.9	88.7	78.4	
	PCB ^[3]	61.9	56.8	60.6	54.4	92.3	77.4	81.7	66.1	
	Pyramid ^[6]	78.9	76.9	78.9	74.8	95.7	88.2	89.0	79.0	
	GFLF-S ^[34]	76.6	73.5	74.4	69.6	94.8	88.0	89.3	77.1	
基于注意力 机制	CASN ^[35]	73.7	68.0	71.5	64.4	94.4	82.8	87.7	73.7	
	MitB ^[36]	70.1	66.5	66.6	64.2	94.7	84.5	85.8	72.9	
	Manes ^[37]	69.0	63.9	65.5	60.5	93.1	82.3	84.9	71.8	
	HACNN ^[24]	44.4	41.0	41.7	38.6	91.2	75.7	80.5	63.9	
其他	DPFL ^[38]	43.0	40.5	40.7	37.0	88.9	73.1	79.2	60.0	
	BDB ^[39]	73.6	71.7	72.8	69.3	94.2	84.3	86.8	72.1	
	SVDNet ^[40]	40.9	37.8	41.5	37.3	82.3	62.1	76.7	56.8	
本文	多尺度联合	80.7	77.0	78.0	73.4	95.9	89.1	90.0	80.4	



图 4 Market1501 数据集部分图像查询结果
Fig. 4 Partial image query results on Market1501 dataset

3.3.3.2 DukeMTMC-ReID 数据集

可以看到本文方法仍然在 Rank-1 和 mAP 达到了最好的效果分别为:90.0%和 80.4%。采用金字塔模型的 Pyramid 最接近本文方法,但仍然低于 1.0% Rank-1 和 1.4% mAP。和只考虑局部特征的 PCB 相比,本文方法分别超过 PCB 的 8.3% Rank-1 和 14.3% mAP。

3.3.3.3 CUHK03 数据集

在新协议下的该数据集是现在最具有挑战的

数据集之一,该数据集的边界框用 2 种不同的方法进行标注分别为 CUHK03 Labeled 和 CUHK03 Detected。该方法在 Labeled 上达到了 80.7% Rank-1 和 77.0% mAP,在 Detected 上达到了78.0% Rank-1 和 73.4% mAP。本文方法在 Labeled 上相比于其他方法,提高了 1.8% 的 Rank-1。对比同样采用多分支结构 MGN 将其原有性能提高了大约 20%,这得益于本文多尺度联合学习方法,增强了不同粒度联合判别的有效性。

3.4 消融实验

为了验证本文方法每个组件的有效性以及探索细粒度局部分支丢弃尺度问题。使用单一查询模式在 Market1501、DukeMTMC-ReID 和 CUHK03 这 3 个数据集上设计了几个不同设置的消融实验。各个指标的结果 Rank-1、mAP 显示在表 2 中。CG 表示粗粒度全局分支,FG 表示细粒度全局分支,FP 表示细粒度局部分支,1、2 分别表示丢弃尺度为 1、2。所有结果只更改一个设置,其余设置均与默认设置相同。

首先,在表 2 中显示了不同组件对本文模型的消融实验。在表格的前三行可以看到,为了验证单个分支的有效性,单独使用粗粒度全局分支或细粒度全局分支,都显著提高了行人重识别的性能。特别的在 CUHK03 数据集上,使用细粒度全局分支分别使 mAP 和 Rank-1 的精确度提高了 12.9% 和 11.8%。第四行将它们联合使用,进一步提高了精确度。为了验证不同粗细粒度分支的组合对模型性能的影响,从第五行开始增加了细粒度局部分支,分别对不同分支相互组合进行实验。可以看出不论以哪一种方式组合相比于不使用该分支,都有助于模型性能的提升。由于细粒度局部分支利用了不相邻部件间的相关性,进一步挖掘了各局部关键但不显著的信息。相比于不同粒度的全局分支,细粒度局部分支更有助于模型鲁棒性的提高,对模型的性能有更加显著的提升。

其次,分析了细粒度局部分支丢弃不同尺度的部件情况。可以发现,不论以何种尺度丢弃部件来挖掘更多关键但非显著的潜在信息。相比于

只使用细粒度全局分支,都有助于增强和全局尺度特征联合判别的有效性。特别的,相比于不使用细粒度局部特征,在 DukeMTMC-ReID 数据集上进一步提高了 1.3% Rank-1 和 1.3% mAP。根据实验可以发现当丢弃尺度为 2 时,在 Market1501、DukeMTMC-ReID 和 CUHK03 Labeled 数据集上更有助于增强局部尺度特征和全局尺度特征联合判别的有效性。对于 CUHK03 Detected 数据集,丢弃尺度为 1 更有效。在行人图像对齐较好的 Market1501、DukeMTMC-ReID 和 CUHK03 Labeled 数据集,相比于丢弃尺度为 1,丢弃尺度为 2 时由于相关联的行人部件更少,更有助于模型挖掘更多关键但非显著的信息,使得行人特征更有鲁棒性。而 CUHK03 Detected 数据集由于行人图像存在大量不对齐的现象,减小丢弃尺度有助于避免具有相似部件的不同行人容易混淆的问题。

接着,笔者基于热力图可视化了一些图像在不同分支下模型关注的区域情况。如图 5 所示,第 1 列为输入到模型的原始图像,第 2 至 5 列以基线为标准并依次递增不同的分支,展示了不同情况下的热力图。可知,当添加了粗粒度全局分支后,由于融合了不同层次的语义信息,增强了模型关注区域的丰富性。当进一步增加了细粒度全局分支后,模型将相邻部件之间相互关联,在保证区域丰富性的同时增强了相邻部件之间的关注度以增强相似部件间的辨别。最后当同时利用 3 个分支时,通过利用不相邻部件之间的相关性,挖掘了更多关键但非显著的信息,增强了关注区域的鲁棒性。

表 2 多尺度联合学习方法消融实验
Table 2 Ablation experiment of multi-scale joint learning method

方法	CUHK03				Market1501		DukeMTMC-ReID	
	Labeled		Detected		Rank-1	mAP	Rank-1	mAP
	Rank-1	mAP	Rank-1	mAP				
基线	59.1	54.2	55.1	50.2	93.5	82.4	85.3	72.0
基线 + CG	69.8	66.1	66.9	62.6	94.8	86.9	87.9	76.7
基线 + FG	70.9	67.1	68.2	63.3	95.1	87.3	88.2	77.9
基线 + CG + FG	76.4	72.1	73.0	68.4	95.3	88.7	88.7	79.1
基线 + CG + FP1	78.4	75.1	76.0	72.2	95.6	88.7	89.1	78.5
基线 + CG + FP2	78.7	75.2	75.5	71.5	95.6	89.0	89.2	79.6
基线 + FG + FP1	77.3	73.1	76.4	71.8	95.6	88.5	89.5	78.9
基线 + FG + FP2	77.6	74.2	75.0	71.4	95.7	88.8	89.5	79.8
基线 + CG + FG + FP1	80.7	77.0	78.0	73.4	95.9	88.8	89.6	79.2
基线 + CG + FG + FP2	80.8	76.7	76.0	71.8	95.9	89.1	90.0	80.4

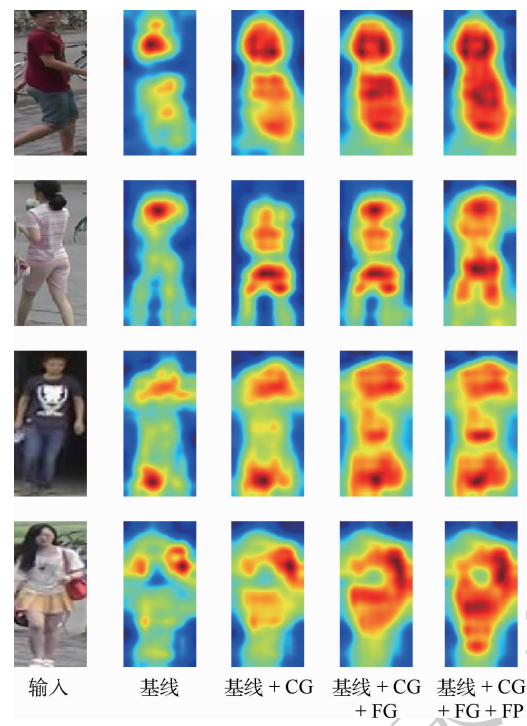


图5 Market1501 数据集部分图像热力图
Fig.5 Partial image heatmap on Market1501 dataset

4 结束语

本文提出了多尺度联合学习方法,通过 3 个分支网络,分别提取行人的粗粒度全局特征、细粒度全局特征和细粒度局部特征,对行人不同粒度下的信息联合学习,使其特征更具有区分性。此外通过挖掘各个部件之间不同尺度下的潜在关系,联合全局特征形成了更有鉴别性的行人特征。大量实验证明,本文方法不仅可以在 3 个主流的行人重识别数据集上实现最好的结果,而且和现有方法相比可以将性能大幅度提高。

参考文献 (References)

[1] LECUN Y,BENGIO Y,HINTON G. Deep learning[J]. Nature, 2015,521(7553):436-444.

[2] ZHAO H,TIAN M,SUN S,et al. Spindle Net: Person re-identification with human body region guided feature decomposition and fusion[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017:907-915.

[3] SUN Y,ZHENG L,YANG Y,et al. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline) [C] // Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018:480-496.

[4] 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.

[5] WEI L,ZHANG S,YAO H,et al. GLAD: Global-local-align-

ment descriptor for pedestrian retrieval[C] // Proceedings of the 25th ACM International Conference on Multimedia. New York: ACM Press, 2017:420-428.

[6] ZHENG F,DENG C,SUN X,et al. Pyramidal person re-identification via multi-loss dynamic training[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019:8514-8522.

[7] FU Y,WEI Y,ZHOU Y,et al. Horizontal pyramid matching for person re-identification[C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2019:8295-8302.

[8] WANG G,YUAN Y,CHEN X,et al. Learning discriminative features with multiple granularities for person re-identification [C] // Proceedings of the 26th ACM International Conference on Multimedia. New York: ACM Press, 2018:274-282.

[9] WANG Z,JIANG J,WU Y,et al. Learning sparse and identity-preserved hidden attributes for person re-identification [J]. IEEE Transactions on Image Processing, 2019,29(1):2013-2025.

[10] ZENG Z,WANG Z,WANG Z,et al. Illumination-adaptive person re-identification [J]. IEEE Transactions on Multimedia, 2020,22(12):3064-3074.

[11] WANG Z,WANG Z,ZHENG Y,et al. Beyond intra-modality: A survey of heterogeneous person re-identification [EB/OL]. (2020-04-27) [2020-07-23]. <https://arxiv.org/abs/1905.10048v4>.

[12] DALAL N,TRIGGS B. Histograms of oriented gradients for human detection[C] // 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2005:886-893.

[13] LIAO S,HU Y,ZHU X,et al. Person re-identification by local maximal occurrence representation and metric learning[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015:2197-2206.

[14] KOESTINGER M,HIRZER M,WOHLHART P,et al. Large scale metric learning from equivalence constraints[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012:2288-2295.

[15] ZHENG L,YANG Y,HAUPTMANN A G. Person re-identification: Past, present and future[EB/OL]. [2020-07-23]. <https://arxiv.org/abs/1610.02984>.

[16] 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.

[17] SU C,LI J,ZHANG S,et al. Pose-driven deep convolutional model for person re-identification [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2017:3980-3989.

[18] SUH Y,WANG J,TANG S,et al. Part-aligned bilinear representations for person re-identification[C] // Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2018:418-437.

[19] XU J,ZHAO R,ZHU F,et al. Attention-aware compositional network for person re-identification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

- Piscataway:IEEE Press,2018:2119-2128.
- [20] SARFRAZ M S, SCHUMANN A, EBERLE A, et al. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2018:420-429.
- [21] ZHENG W S, LI X, XIANG T, et al. Partial person re-identification [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2015:4678-4686.
- [22] LUO H, JIANG W, ZHANG X, et al. AlignedReID++: Dynamically matching local information for person re-identification [J]. Pattern Recognition,2019,94:53-61.
- [23] SUN Y, XU Q, LI Y, et al. Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:393-402.
- [24] LI W, ZHU X, GONG S. Harmonious attention network for person re-identification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2018:2285-2294.
- [25] LIU X, ZHAO H, TIAN M, et al. HydraPlus-Net: Attentive deep features for pedestrian analysis [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2017:350-359.
- [26] 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.
- [27] WANG F, XIANG X, CHENG J, et al. Normface: L2 hypersphere embedding for face verification [C] // Proceedings of the 25th ACM International Conference on Multimedia. New York:ACM Press,2017:1041-1049.
- [28] HERMANS A, BEYER L, LEIBE B. In defense of the triplet loss for person re-identification [EB/OL]. (2017-11-17) [2020-07-23]. <https://arxiv.org/abs/1703.07737>.
- [29] ZHENG L, SHEN L, TIAN L, et al. Scalable person re-identification: A benchmark [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2015:1116-1124.
- [30] LI W, ZHAO R, XIAO T, et al. DeepReID: Deep filter pairing neural network for person re-identification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2014:152-159.
- [31] ZHONG Z, ZHENG L, CAO D, et al. Re-ranking person re-identification with k -reciprocal encoding [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2017:1318-1327.
- [32] ZHENG Z, ZHENG L, YANG Y. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2017:3754-3762.
- [33] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2009:248-255.
- [34] PARK H, HAM B. Relation network for person re-identification [EB/OL]. (2017-08-22) [2020-07-23]. <https://arxiv.org/abs/1701.07717>.
- [35] ZHENG M, KARANAM S, WU Z, et al. Re-identification with consistent attentive siamese networks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:5735-5744.
- [36] YANG W, HUANG H, ZHANG Z, et al. Towards rich feature discovery with class activation maps augmentation for person re-identification [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE Press,2019:1389-1398.
- [37] WANG C, ZHANG Q, HUANG C, et al. Manacs: A multi-task attentional network with curriculum sampling for person re-identification [C] // Proceedings of the European Conference on Computer Vision. Berlin:Springer,2018:365-381.
- [38] CHEN Y, ZHU X, GONG S. Person re-identification by deep learning multi-scale representations [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2017:2590-2600.
- [39] DAI Z, CHEN M, GU X, et al. Batch DropBlock network for person re-identification and beyond [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2019:3690-3700.
- [40] SUN Y, ZHENG L, DENG W, et al. SVDNet for pedestrian retrieval [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway:IEEE Press,2017:3820-3828.

作者简介:

谢彭宇 男,硕士研究生。主要研究方向:计算机视觉、行人重识别。

徐新 男,博士,教授,博士生导师。主要研究方向:计算机视觉、机器学习、行人重识别。

Multi-scale joint learning for person re-identification

XIE Pengyu¹, XU Xin^{1,2,3,*}

- (1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;
- 2. Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China;
- 3. School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: The existing person re-identification approaches mainly focus on learning person’s local features to match a specific pedestrian across different cameras. However, in the presence of incomplete conditions of pedestrian data such as motion or occlusion of human body parts, background interference, etc., it leads to an increase in the probability of partial loss of pedestrian recognition information. This paper presents a multi-scale joint learning method to extract the fine-grained person feature. This method consists of three subnets, i. e. coarse-grained global feature extraction subnet, fine-grained global feature extraction subnet, and fine-grained local feature extraction subnet. The coarse-grained global feature extraction subnet enhances the diversity of the global feature by fusing semantic information at different levels. The fine-grained global branching unites all local features to learn the correlation among local components of a pedestrian while describing the global features at a fine-grained level. The fine-grained local feature extraction subnet enhances robustness by traversing local features and finding out pedestrian non-significant information. Comparative experiments have been conducted to evaluate the performance of the proposed method against state-of-the-art methods on Market1501, DukeMTMC-ReID, and CUHK03 person re-identification datasets. The experimental results show that the proposed method has the best performance.

Keywords: person re-identification; multi-scale; joint learning; multi-branch network; deep learning

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

DOI: 10.13700/j.bh.1001-5965.2020.0460

基于块对角投影表示的人脸识别

刘保龙^{1,2,3}, 王勇^{3,4}, 李丹萍⁵, 王磊^{2,3,*}

(1. 西安电子科技大学青岛计算技术研究院, 青岛 266000;

2. 上海交通大学 海洋智能装备与系统教育部重点实验室, 上海 200240;

3. 西安电子科技大学 电子工程学院, 西安 710071; 4. 中国电子科技集团公司第27研究所, 郑州 450047;

5. 西安电子科技大学 通信工程学院, 西安 710071)

摘

要: 针对大多数特征表示算法在挖掘高维数据内在结构时容易受到噪声的影响,

以及特征学习与分类器设计割裂导致分类性能降低的问题,提出了一种新的基于特征表示的人脸识别方法,称为块对角投影表示(BDPR)学习。首先,利用样本信息对每类样本的编码系数施加一个加权矩阵,通过局部约束来加强表示系数之间的相似性,从而降低噪声对系数学习的影响,使所提方法能够更好地保持数据的局部结构。其次,为了实现数据与编码系数相关联,降低表示系数的学习难度,构造了块对角化判别约束项来学习一个判别投影,通过投影从低维数据中提取样本表示系数,使系数包含更多的样本间全局结构信息且具有更低的计算复杂度。最后,将系数学习和分类器学习整合到同一框架下,同时增大不同类别样本间的“标签距离”,采用迭代求解的方式交替更新判别投影和分类器,最终得到最适合当前表示特征的分

关 键 词: 图像分类; 特征表示; 局部约束; 判别投影; 块对角化结构

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2021)03-0623-09

近年来,随着计算机技术的迅速发展,人脸识别技术得到广泛研究与应用,人脸识别成为近30年里模式识别和图像处理中最热门的研究主题之一。人脸识别的目的是从人脸图像中抽取人的个性化特征,并以此来识别人的身份。

随着人们获取数据能力提高的同时,数据维数的增加引起了“维数灾难”问题,限制了人脸识别的准确度和识别效率,因此对数据进行降维预处理变得尤为必要。其中,主成分分析(PCA)^[1]和线性判别分析(LDA)^[2]是最早的降维方法,有效缓解了维数灾难问题,提高了实验的精确度,在

目标识别、图像分割^[3]和信号处理^[4]等领域有着广泛的应用。为了增强投影的判别力,He等^[5]考虑样本的局部信息,提出局部保持投影(LPP),进一步,根据重构误差的最小,其提出了近邻保持嵌入(NPE)^[6],但是该算法并没有考虑样本的全局结构;Sugiyama^[7]将LPP算法与LDA算法结合,提出了局部费舍尔判别分析(LFDA),通过最大化类间可分性和保持类内局部结构来适当地降低多聚类结构数据的维数;2007年,Yan等^[8]提出了图嵌入框架,定义了本征图和惩罚图来描述数据的特征,并提出了边界费舍尔分析(MFA)。上述算法

收稿日期: 2020-08-25; 录用日期: 2020-09-04; 网络出版时间: 2020-10-22 15:15

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201022.1412.001.html

基金项目: 国家重点研发计划(2016YFE0207000); 国家自然科学基金(61203137,61401328); 陕西省自然科学基金基础研究计划(2014JQ8306,2015JM6279)

* 通信作者. E-mail: leiwang@mail.xidian.edu.cn

引用格式: 刘保龙, 王勇, 李丹萍, 等. 基于块对角投影表示的人脸识别[J]. 北京航空航天大学学报, 2021, 47(3): 623-631.

LIU B L, WANG Y, LI D P, et al. Block-diagonal projective representation for face recognition[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 623-631 (in Chinese).

被广泛应用于维数约简,具有良好的降维效果。

随着进一步的研究,基于表示的分类在机器学习和模式识别等领域备受关注^[9],常见的有稀疏表示、协作表示、低秩表示等。稀疏表示技术起源于压缩感知理论,最初应用于信号处理领域,对信号起压缩作用^[10]。2009年,Wright等^[11]首次将稀疏表示应用于图像分类中,提出基于稀疏表示的分类(Sparse Representation based Classification, SRC),其主要思想是:在 l_1 范数的约束下,将测试样本表示为训练样本的一个线性组合,通过重建误差最小对测试样本进行分类,强调样本表示系数的稀疏性。鉴于SRC的成功,许多基于稀疏表示的降维方法相继出现。Qiao等^[12]提出了一种基于SRC的稀疏保持投影(SPP)来保持数据的稀疏构造关系。Ly等^[13]提出了基于图的稀疏判别分析(SGDA)算法,并应用于遥感图像的特征降维,该算法与SPP思想相近,不同的是SGDA是有监督的算法,利用了数据的判别信息。但是解决 l_1 范数最小化问题往往需要较大的计算量,这使得基于稀疏表示的方法(无论是原始的SRC还是降维的SPP等)均具有很高的计算复杂度。

虽然SRC强调了表示系数稀疏性的重要,但是Zhang等^[14]证明了真正提高人脸识别精度的机制可能来自协作表示,而非 l_1 范数稀疏性约束,进而提出了基于协作表示的分类(Collaborative Representation based Classification, CRC),其采用 l_2 范数约束,不仅提高了运算效率,而且通过理论证明了CRC同样具有稀疏分类的效果。但是严格来说,与SRC一样,CRC并不算有监督分类方法并且作用于数据的原始空间。为了充分利用样本的标签信息,将CRC的思想与有监督投影算法相结合产生了一系列工作。2014年, Ly等^[15]提出了基于图的协作判别分析(CGDA)方法,其通过同类样本之间的协作表示来构造类内权值矩阵,使样本映射到低维空间后能够保持样本之间的局部信息。2016年, Yang等^[16]提出了基于判别投影的正则化最小回归(RLSDP),首先利用所有训练样本用 l_2 范数正则化最小二乘法重构每个训练样本,通过最大化LDA的类间散射和最小化同一类的重构残差,寻找一个判别子空间。除上述算法外,许多学者在协作表示基础上,不断地推出了新的方法和应用场景^[17-18]。

尽管上述算法在图像分类上取得了不错的效果,但是它们主要从最小化重构误差角度出发,无法保证采用组系数约束得到相同类别样本的编码系数间是否相似。另一方面,上述算法均为两阶

段模型,实际上分类器的更新反馈对于算法的优化过程具有引导作用,将两者整合到同一框架下,通过联合优化,可获得更优的编码系数、判别投影矩阵以及更适合当前任务的分类器,从而有助于模型识别性能的提升。

基于此,本文提出了一种基于块对角投影表示(BDPR)学习方法,其主要优势如下:

1) 充分利用了同类样本之间的相似性,通过同类样本间的距离构造系数约束项,增强相同类别样本表示系数的相似性,同时使个别受噪声污染严重的样本在数据重构中的贡献最小化,以加强干净特征在联合表示中的作用。

2) 构造编码系数块对角化判别逼近项来学习判别投影,实现数据与编码系数相关联,使得判别投影具有提取数据相关表示的能力,同时使近似编码系数与标签信息相关联,确保近似编码系数尽可能具有块对角化结构。

3) 通过提取的具有块对角化结构的近似编码系数训练线性分类器,利用分类器的迭代反馈更新算法的优化过程以提升模型的最终分类性能。

在多个公开人脸数据集上测试的结果表明,本文所提BDPR方法的分类性能优于目前基于协作表示的分类方法和多个主流的子空间学习方法。

值得一提的是,上述子空间投影方法可以归纳入一种特殊的度量学习框架^[19],随着深度学习的方兴未艾,以深度度量学习为代表的系列方法在图像分类中取得了瞩目的效果^[20],构建了一系列度量函数,驱动各类深度模型在人脸识别^[21]、车辆识别^[22]、故障诊断^[23]等领域展现出卓越的性能。

1 基于块对角化结构的投影表示

1.1 方法描述

1) 基于编码系数块对角化结构的重构误差项

本文BDPR方法的首要目标是单独地学习每一类人脸图像数据的编码系数,以信息损失最少为代价重构原始图像数据,同时获得判别投影矩阵 $\mathbf{P}$ 用于从每个人的图像数据中提取具有判别性的编码系数,从而达到区分人脸图像的作用。因此,重构误差项可设计为

$$\min \sum_{i=1}^c \left\{ \|\mathbf{X}_i - \mathbf{X}_i \mathbf{Z}_i\|_F^2 + \lambda \|\mathbf{Z}_i\|_F^2 + \alpha \|\mathbf{R}_i \odot \mathbf{Z}_i\|_F^2 + \beta \|\mathbf{P} \mathbf{X}_i - \mathbf{Q}_i \mathbf{Z}_i\|_F^2 \right\}$$

$$\text{s. t. } \mathbf{Z}_i \geq \mathbf{0}, \text{diag}(\mathbf{Z}_i) = \mathbf{0}, \mathbf{Z}_i \mathbf{1} = \mathbf{1} \quad (1)$$

式中: $\mathbf{X}_i \in \mathbf{R}^{d \times N_i}$ 为第 i 类的训练样本, d 为样本维度, N_i 为第 i 类样本的个数, $i = 1, 2, \dots, c$, c 为整

个训练集 $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_c] \in \mathbf{R}^{d \times N}$ 的样本类别数, N 为样本总数; $\mathbf{Z}_i \in \mathbf{R}^{N_i \times N_i}$ 和 $\mathbf{R}_i \in \mathbf{R}^{N_i \times N_i}$ 分别为第 i 类样本的编码系数和样本之间的欧氏距离; $\mathbf{Q} = [\mathbf{Q}_1, \mathbf{Q}_2, \dots, \mathbf{Q}_c] \in \mathbf{R}^{N \times N}$ 为判别转换矩阵; α, β 和 λ 为正则化参数; $\mathbf{1} \in \mathbf{R}^{N_i \times 1}$ 为所有元素全为 1 的列向量。在 BDPR 方法中通过 $\mathbf{Q}_i$ 构建第 i 类理想化的样本编码系数,即每个样本应该仅由相同类样本表示。假设 $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4, \mathbf{x}_5, \mathbf{x}_6]$, $\mathbf{x}_1, \mathbf{x}_2$ 属于第 1 类, $\mathbf{x}_3, \mathbf{x}_4$ 属于第 2 类, $\mathbf{x}_5, \mathbf{x}_6$ 属于第 3 类,那么 $\mathbf{Q}$ 的构建方式为

$$\mathbf{Q} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad (2)$$

$\underbrace{\quad}_{\mathbf{Q}_1} \quad \underbrace{\quad}_{\mathbf{Q}_2} \quad \underbrace{\quad}_{\mathbf{Q}_3}$

式(1)中第 1 项 $\|\mathbf{X}_i - \mathbf{X}_i \mathbf{Z}_i\|_F^2$ 为样本重构误差项,借鉴自编码器的学习原理,使得每个样本仅由相同类别样本重构;第 2 项 $\|\mathbf{Z}_i\|_F^2$ 则是为了防止方法过拟合;第 3 项 $\|\mathbf{R}_i \odot \mathbf{Z}_i\|_F^2$ 为编码系数局部约束项,目的在于增强相同类别样本表示系数间的相似性,同时最小化噪声对编码系数影响;第 4 项 $\|\mathbf{P}\mathbf{X}_i - \mathbf{Q}_i \mathbf{Z}_i\|_F^2$ 为编码系数近似误差项,利用投影矩阵 $\mathbf{P}$ 将数据特征与编码系数进行关联,使近似编码系数矩阵 $\mathbf{P}\mathbf{X}$ 近似于块对角化系数矩阵 $\mathbf{Z}$,从而使投影矩阵 $\mathbf{P}$ 具有提取数据稀疏表示的能力。

为避免编码系数为负,以及样本被自己表示的极端情况,进一步约束 $\mathbf{Z}_i$ 非负,并强制其对角元素为 0。同时为了避免平凡解,即一些样本在线性表示中不被选择,约束 $\mathbf{Z}_i$ 每一行的和为 1。

2) 基于编码系数块对角化结构的分类误差项

为了训练更适合于有监督人脸识别任务的分类器,在式(1)的基础上,进一步引入人脸图像数据的标签信息进行分类器学习,在最小化样本重构误差的同时将分类器的分类误差作为反馈指导编码系数的优化,提升编码系数的表示能力。假设引入每个样本 $\mathbf{x}$ 对应的标签向量为 $\mathbf{y} = [0, \dots, 1, \dots, 0]^T \in \mathbf{R}^{c \times 1}$, 1 所在的位置对应 $\mathbf{x}$ 的标签。假设可将近似编码系数 $\mathbf{P}\mathbf{X}$ 投影到标签矩阵上来学习一个线性分类器,那么分类误差项可写为

$$\min \tau \left(\|\mathbf{W}\mathbf{P}\mathbf{X} - \mathbf{Y} - \mathbf{B} \odot \mathbf{H}\|_F^2 + \|\mathbf{W}\|_F^2 \right) \quad (3)$$

式中: $\mathbf{W} \in \mathbf{R}^{c \times N}$ 为线性分类器; $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N] \in \mathbf{R}^{c \times N}$ 为标签矩阵; τ 为正则化参数。需要注意,从线性回归的角度分析,以 0-1 标签矩阵 $\mathbf{Y}$ 为回归目标会使得任意一对不同类别编码系数的回归目标之间

的距离为 $\sqrt{2}$,实际上不同类别编码系数的回归目标之间的可区分性应当是越大越好。基于此,在式(3)中利用 ε -dragging 技术构建回归误差项,使学习到的分类器具有更强的判别性。“ $\odot$ ”表示矩阵的哈达马积(Hadamard Product, HP); $\mathbf{H} \in \mathbf{R}^{c \times N}$ 中的元素即为每个样本针对不同类别的 ε 项; $\mathbf{B} \in \mathbf{R}^{c \times N}$ 为一个常数矩阵, $\mathbf{B}$ 的第 i 行第 j 列的元素为

$$\mathbf{B}_{ij} = \begin{cases} +1 & i = l_j \\ -1 & \text{其他} \end{cases} \quad (4)$$

其中: l_j 表示 $\mathbf{x}_j$ 所属的类别。例如对于上述 $\mathbf{x}_1 \sim \mathbf{x}_6$ 构成的数据,其 $\mathbf{B}$ 矩阵构造为

$$\mathbf{B} = \begin{bmatrix} +1 & +1 & -1 & -1 & -1 & -1 \\ -1 & -1 & +1 & +1 & -1 & -1 \\ -1 & -1 & -1 & -1 & +1 & +1 \end{bmatrix} \quad (5)$$

通过结合式(1)和式(3),可以得到所提 BD-PR 方法的最终目标函数为

$$\begin{aligned} \min \quad & \sum_{i=1}^c \left\{ \|\mathbf{X}_i - \mathbf{X}_i \mathbf{Z}_i\|_F^2 + \lambda \|\mathbf{Z}_i\|_F^2 + \right. \\ & \alpha \|\mathbf{R}_i \odot \mathbf{Z}_i\|_F^2 + \beta \|\mathbf{P}\mathbf{X}_i - \mathbf{Q}_i \mathbf{Z}_i\|_F^2 \left. \right\} + \\ & \tau \left(\|\mathbf{W}\mathbf{P}\mathbf{X} - \mathbf{Y} - \mathbf{B} \odot \mathbf{H}\|_F^2 + \|\mathbf{W}\|_F^2 \right) \\ \text{s. t.} \quad & \mathbf{Z}_i \geq \mathbf{0}, \text{diag}(\mathbf{Z}_i) = \mathbf{0}, \mathbf{Z}_i \mathbf{1} = \mathbf{1} \end{aligned} \quad (6)$$

为了验证 BDPR 提取的近似编码系数 $\mathbf{P}\mathbf{X}$ 为块对角化结构,在 AR 人脸数据集上进行测试。采用 CRC 作为对比方法,实验结果如图 1 所示。图 1(a)为 CRC 所提取编码系数,图 1(b)为 BDPR 所提取近似编码系数 $\mathbf{P}\mathbf{X}$ 。可以看出, CRC 方法提取的编码系数块对角结构不明显,且主要分量集中于对角线元素,即样本重构过程中存在自我表示问题; BDPR 所提取的编码系数则具有较为明显的块对角化结构,表现出更明确的样本关系信息。

1.2 优化过程

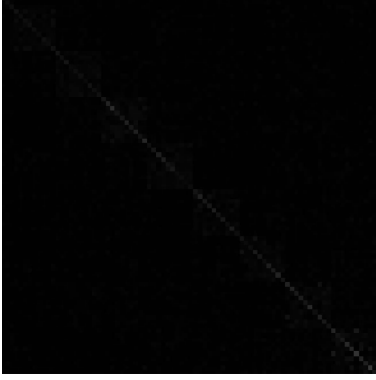
首先初始化 $\mathbf{P}$ 和 $\mathbf{W}$ 为随机矩阵。为了便于求解,引入一个辅助变量 $\mathbf{A}$,式(6)可转化为如下的等价问题:

$$\begin{aligned} \min \quad & \sum_{i=1}^c \left\{ \|\mathbf{X}_i - \mathbf{X}_i \mathbf{Z}_i\|_F^2 + \lambda \|\mathbf{Z}_i\|_F^2 + \alpha \|\mathbf{R}_i \odot \mathbf{A}_i\|_F^2 + \right. \\ & \beta \|\mathbf{P}\mathbf{X}_i - \mathbf{Q}_i \mathbf{Z}_i\|_F^2 \left. \right\} + \\ & \tau \left(\|\mathbf{W}\mathbf{P}\mathbf{X} - \mathbf{Y} - \mathbf{B} \odot \mathbf{H}\|_F^2 + \|\mathbf{W}\|_F^2 \right) \\ \text{s. t.} \quad & \mathbf{Z}_i \geq \mathbf{0}, \text{diag}(\mathbf{Z}_i) = \mathbf{0}, \mathbf{Z}_i \mathbf{1} = \mathbf{1}, \mathbf{Z}_i = \mathbf{A}_i \end{aligned} \quad (7)$$

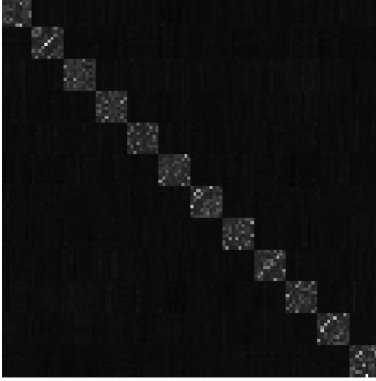
式中: $\mathbf{A}$ 的结构为

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{A}_c \end{pmatrix} \quad (8)$$

可以构建如下拉格朗日函数:



(a) CRC所提取编码系数



(b) BDPR所提取编码系数

图1 CRC与BDPR在AR数据集上的样本编码系数对比

Fig. 1 Comparison of sample coding coefficients learned by CRC and BDPR on AR dataset

$$L = \sum_{i=1}^c \left\{ \left\| X_i - X_i Z_i \right\|_F^2 + \lambda \left\| Z_i \right\|_F^2 + \alpha \left\| R_i \odot A_i \right\|_F^2 + \beta \left\| P X_i - Q_i Z_i \right\|_F^2 + \frac{\mu}{2} \left\| Z_i - A_i + \frac{C_i}{\mu} \right\|_F^2 \right\} + \tau \left(\left\| W P X - Y - B \odot H \right\|_F^2 + \left\| W \right\|_F^2 \right)$$

$$\text{s. t. } Z_i \geq 0, \text{diag}(Z_i) = 0, Z_i \mathbf{1} = \mathbf{1} \quad (9)$$

式中: μ 为惩罚参数; C_i 为拉格朗日乘子。对于式(9),可以采用迭代求解的方式进行。

步骤1 固定 Z 、 P 、 W ,更新 A ,对应子问题为

$$\min_A \sum_{i=1}^c \left\{ \alpha \left\| R_i \odot A_i \right\|_F^2 + \frac{\mu}{2} \left\| Z_i - A_i + \frac{C_i}{\mu} \right\|_F^2 \right\} \quad (10)$$

对于第 i 类样本:

$$f(A_i) = \alpha \left\| R_i \odot A_i \right\|_F^2 + \frac{\mu}{2} \left\| Z_i - A_i + \frac{C_i}{\mu} \right\|_F^2 \quad (11)$$

定义 $G_i = Z_i + \frac{C_i}{\mu}$,式(11)可以等价于

$$\min_{A_i} \alpha \left\| R_i \odot A_i \right\|_F^2 + \frac{\mu}{2} \left\| G_i - A_i \right\|_F^2 \Leftrightarrow$$

$$\min_{A_i} \sum_{k=1}^{N_i} \sum_{j=1}^{N_i} \left(r_{kj}^2 a_{kj}^2 + \frac{\mu}{2\alpha} (a_{kj} - g_{kj})^2 \right) \Leftrightarrow$$

$$\sum_{k=1}^{N_i} \sum_{j=1}^{N_i} \min_{a_{kj}} \left(a_{kj} - \frac{\frac{\mu}{\alpha} g_{kj}}{2r_{kj}^2 + \frac{\mu}{\alpha}} \right)^2 \quad (12)$$

式中: a_{kj} 、 g_{kj} 、 r_{kj} 分别为 A_i 、 G_i 、 R_i 的第 k 行第 j 列元素。对于式(12),可以得到最优解 A_i 的每个元素为

$$a_{kj} = \frac{\mu}{2\alpha r_{kj}^2 + \mu} g_{kj} \quad (13)$$

结合式(8)和式(13),则可求解出变量 A 。

步骤2 固定 A 、 P 、 W ,更新 Z ,其中 Z 为 $[Z_1, Z_2, \dots, Z_c]$ 构成的块对角化矩阵。求解 Z_i ($i=1, 2, \dots, c$)对应子问题为

$$\min_{Z_i} \sum_{i=1}^c \left\{ \left\| X_i - X_i Z_i \right\|_F^2 + \lambda \left\| Z_i \right\|_F^2 + \beta \left\| P X_i - Q_i Z_i \right\|_F^2 + \frac{\mu}{2} \left\| Z_i - A_i + \frac{C_i}{\mu} \right\|_F^2 \right\}$$

$$\text{s. t. } Z_i \geq 0, \text{diag}(Z_i) = 0, Z_i \mathbf{1} = \mathbf{1} \quad (14)$$

为了减小计算复杂度和提高计算效率,参考Xu等^[24]的方法,首先通过式(15)计算式(14)的近似闭式解 $\hat{Z}_i$:

$$\hat{Z}_i = \min_{Z_i} \sum_{i=1}^c \left\{ \left\| X_i - X_i Z_i \right\|_F^2 + \lambda \left\| Z_i \right\|_F^2 + \beta \left\| P X_i - Q_i Z_i \right\|_F^2 + \frac{\mu}{2} \left\| Z_i - A_i + \frac{C_i}{\mu} \right\|_F^2 \right\} \quad (15)$$

式中: $\hat{Z}$ 的结构与 A 一致,均为块对角化结构。对于第 i 类样本

$$f(Z_i) = \left\| X_i - X_i Z_i \right\|_F^2 + \lambda \left\| Z_i \right\|_F^2 + \beta \left\| P X_i - Q_i Z_i \right\|_F^2 + \frac{\mu}{2} \left\| Z_i - A_i + \frac{C_i}{\mu} \right\|_F^2 \quad (16)$$

对式(16)求 Z_i 的导数,并令 $\partial f(Z_i) / \partial Z_i = 0$ 得

$$\hat{Z}_i = \left(X_i^T X_i + \beta Q_i^T Q_i + \left(\lambda + \frac{\mu}{2} \right) I \right)^{-1} \cdot \left(X_i^T X_i + \beta Q_i^T P X_i + \frac{\mu}{2} \left(A_i - \frac{C_i}{\mu} \right) \right) \quad (17)$$

式中: I 为单位矩阵。

参考Xu等^[24]的方法,可以根据式(18)求解目标解 Z_i :

$$\min_{Z_i \geq 0, \text{diag}(Z_i) = 0, Z_i \mathbf{1} = \mathbf{1}} \left\| Z_i - \hat{Z}_i \right\|_F^2 \quad (18)$$

对于式(18),可以得到 Z_i 的每一行为

$$z_k = \max(\zeta_k \bar{\mathbf{1}}^T + \bar{z}_k, \mathbf{0}) \quad (19)$$

式中: $\bar{z}_k = [\bar{z}_{k1}, \dots, \bar{z}_{kn}, \dots, \bar{z}_{kn}] \in \mathbf{R}^{1 \times n}$ 为 $\hat{Z}_i$ 的第 k 行且 $\bar{z}_{kk}$ 被设置为0; $\bar{\mathbf{1}}$ 为第 k 个元素为0,其余全为1的列向量; ζ_k 为拉格朗日乘子,其可根据式(20)计算:

$$\zeta_k = \frac{1 + \bar{z}_k \mathbf{1}}{n - 1} \quad (20)$$

Z 的形式与 A 一致,结合式(8)和式(19),可以得到编码系数 Z 。

步骤 3 固定 A, Z, W , 更新 P , 对应子问题为

$$\min_P \sum_{i=1}^c \{ \beta \|PX_i - Q_i Z_i\|_F^2 \} + \tau \|WPX - Y - B \odot H\|_F^2 \quad (21)$$

假设 $X = [X_1, X_2, \dots, X_c]$, $M = [Q_1 Z_1, Q_2 Z_2, \dots, Q_c Z_c]$, 则式(21)等价于

$$\min_P \beta \|P[X_1, X_2, \dots, X_c] - [Q_1 Z_1, Q_2 Z_2, \dots, Q_c Z_c]\|_F^2 + \tau \|WPX - Y - B \odot H\|_F^2 = \beta \|PX - M\|_F^2 + \tau \|WPX - Y - B \odot H\|_F^2 \quad (22)$$

对式(22)求 P 的导数, 并令其为 0 可得

$$P = (\beta I + \tau W^T W)^{-1} (\beta M X^T + \tau W^T (Y + B \odot H) X^T) (X X^T)^{-1} \quad (23)$$

步骤 4 固定 A, Z, P , 更新 W , 对应子问题为

$$\min_W \|WPX - Y - B \odot H\|_F^2 + \|W\|_F^2 \quad (24)$$

令式(24)导数为 0, 可以得到其最优解 W 为

$$W = ((Y + B \odot H) X^T P^T) (P X X^T P^T + I)^{-1} \quad (25)$$

步骤 5 固定 A, Z, P, W , 更新 H , 对应子问题为

$$\min_H \|WPX - Y - B \odot H\|_F^2 \quad (26)$$

令 $S = WPX - Y$, 式(26)可被表示为

$$\min_H \|S - B \odot H\|_F^2 \quad (27)$$

式(27)可以等效地解耦为 $c \times N$ 个子问题, 对于矩阵 H 的第 i 行第 j 列个元素, 有

$$\min_{H_{i,j}} (S_{i,j} - B_{i,j} H_{i,j})^2 \quad \text{s. t.} \quad H_{i,j} \geq 0 \quad (28)$$

式中: $S_{i,j}$ 和 $B_{i,j}$ 分别为 S 和 B 的第 i 行第 j 列个元素。容易得到最优解 $H_{i,j} = \max(B_{i,j} S_{i,j}, 0)$, 即

$$H = \max(B \odot S, 0) \quad (29)$$

1.3 测试过程

给定测试样本 x_{test} , 首先利用最终学习到的投影矩阵计算 x_{test} 的近似编码系数 Px_{test} , 然后使用学习到的线性分类器 W 得到 x_{test} 的软标签向量 f_{test} , 其具体形式为

$$f_{\text{test}} = WPx_{\text{test}} \in \mathbf{R}^{c \times 1} \quad (30)$$

最后找到 f_{test} 最大元素所处的索引值, 并将其作为 x_{test} 的预测类别标签。

将 BDPR 的方流程总结如下。

算法 1 BDPR 方法。

输入: 训练集 X , 距离度量矩阵 R , 类标签矩阵 Y , 调节参数 $\lambda, \alpha, \beta, \tau$ 。

输出: 编码系数 Z , 投影矩阵 P , 线性分类器 W 。

初始化: $C_i = 0 (i = 1, 2, \dots, c)$, $\mu = 0.2$, $\mu_{\max} = 10^8$, $\rho = 1.1$, $Z = I$ 。

训练:

1. 按照式(13)更新 A 。
2. 按照式(19)更新 Z 。

3. 按照式(23)更新 P 。

4. 按照式(25)更新 W 。

5. 按照式(29)更新 H 。

6. 按照下式更新 C_i 和 μ :

$$C_i^{t+1} = C_i^t + \mu(Z_i - A_i)$$

$$\mu^{t+1} = \min(\mu_{\max}, \rho \mu^t)$$

7. 检查式(6)函数值是否满足收敛条件, 若满足或者方法达到最大迭代次数, 迭代停止, 否则跳到步骤 1 继续执行。

测试:

1. 根据式(30)计算测试样例的软标签向量 f_{test} 。
2. 将 f_{test} 中最大元素的位置索引作为测试样例的预测类别标签。

2 实验结果与分析

本节对所提 BDPR 方法在 3 个不同的人脸数据集 (BANCA、AR、YaleB) 进行实验, BANCA 为小样本数据集, AR 和 YaleB 为中等样本数据集, 具体数据集信息如表 1 所示。对比方法为标准的协作表示分类 CRC^[14] 和经典的子空间投影算法, 包括 PCA^[1]、LPP^[5]、MFA^[8]、CGDA^[15]、RLSDP^[16]、RLSL^[25] 等。在预处理阶段, 统一先采用 PCA 方法保留样本 99% 的能量对样本进行初步降维。在训练阶段, 本文在 $\{10^{-4}, 10^{-3}, \dots, 10^4\}$ 区间内采用网格搜索的方式寻找各方法最优参数组合。所有实验均在软件环境为 MATLAB R2019a、硬件环境为 8.00 GB RAM、2.40 GHz CPU、64 位 Windows 操作系统的计算机上进行。

表 1 实验采用的数据集信息

Table 1 Information of dataset used in experiment

数据集	类别	样本数	数据维度
BANCA	52	520	2 576
AR	50	1 300	2 200
YaleB	38	2 414	1 024

2.1 实验结果对比

本节将详细介绍实验中所采用的人脸数据集 (BANCA、AR、YaleB) 及各方法的实验结果。

BANCA 数据集^[26]: 包含 208 人共 2 496 幅图像, 这些图像包含不同图像质量、不同时间段、不同光照等变化条件。实验中选取了 52 个人, 每人包含 10 幅图像, 共 520 幅图像组成数据集。将原始数据图像裁剪为 56×46 。实验中每类选取 4、5、6 幅图像为训练样本, 其余样本构建测试集。

AR 数据集: 包含 4 000 多种人脸图像, 分别对应 126 人的面部, 包含不同的面部表情、照明条件和遮挡的情况。实验中, 选取 50 个人, 每人包

含 26 幅图像,共 1 300 幅图像组成数据集。将原始数据图像裁剪为 55×40 。实验中,每类分别选取 5、10、15 幅图像作为训练样本,其余样本构建测试集。

YaleB 数据集^[27]:包含 38 个人共 2 414 幅图像、9 种姿态和 64 种光照的变化。每幅图像剪裁

为 32×32 。实验中,每类随机选取 20、30、40 幅图像作为训练集,其余样本构建测试集。

实验中,对数据进行 10 次随机划分,表 2 为 3 个数据集上各方法在最优参数下的平均正确率,加粗数据表示同一实验条件下的最高分类正确率。

表 2 各方法在不同数据集的最优分类正确率

方法	BANCA			AR			YaleB		
	4 train	5 train	6 train	5 train	10 train	15 train	20 train	30 train	40 train
CRC	46.03	45.92	47.74	68.79	81.85	87.58	92.67	88.69	96.02
PCA	48.40	52.42	56.63	33.73	48.09	57.60	61.90	68.31	72.19
LPP	41.12	38.78	42.84	38.00	48.44	61.20	83.97	88.69	90.83
MFA	73.01	77.62	80.67	86.49	94.43	97.24	93.97	96.04	97.13
CGDA	67.30	73.73	79.47	90.69	96.66	98.69	93.86	95.00	98.02
RLSDP	68.97	76.38	82.12	88.96	96.69	98.69	91.90	95.49	96.86
RLSL	72.18	77.50	82.13	84.97	95.63	97.87	93.47	96.85	98.15
BDPR	75.80	79.88	84.38	93.01	97.38	98.60	94.33	97.37	98.48

从实验结果可以得出以下结论:
1) PCA 和 LPP 算法没有利用样本的标签信息,属于无监督算法,因此其分类性能明显低于有监督算法。

2) MFA、CGDA 和 RLSDP 虽然都是有监督算法,但仅考虑了样本的类内和类间特性,而 BDPR 不仅利用标签信息构建了样本间的协作表示,而且考虑了相同类样本编码系数的相似性构造了编码系数局部约束项,并且该项可降低野值点样本对其他同类样本的重构贡献,因而编码系数具有更鲁棒的数据关系表征能力。

3) 上述对比方法的分类器是不参与训练的, BDPR 则将系数学习和分类器学习整合到一个联合框架下,利用分类过程的反馈信息引导算法优化,学习到的分类器更适合当前样本特点且可直接用来分类,无需另外选择分类算法,避免了两阶段方法带来的系统误差。

2.2 方法收敛性

为了考察方法的收敛性,分别在上述 3 个数据集上记录了方法在整个训练过程中目标函数值的变化。图 2 给出 BDPR 在各数据集上(前述实验各最高训练数目下)方法的收敛性能曲线。可知 BDPR 在各数据集上均具有良好的收敛性能,并且从第 3 次迭代开始,收敛曲线均趋于平缓,说明 BDPR 具有较快的收敛速度。

2.3 参数敏感性

为了衡量不同的参数(λ 、 α 、 β 和 τ)对 BDPR 分类性能的影响,考察其在不同的参数组合情况下、10 次随机实验的平均分类正确率。选取 AR 人脸数据库,每类随机选取 10 个样例作为训练样

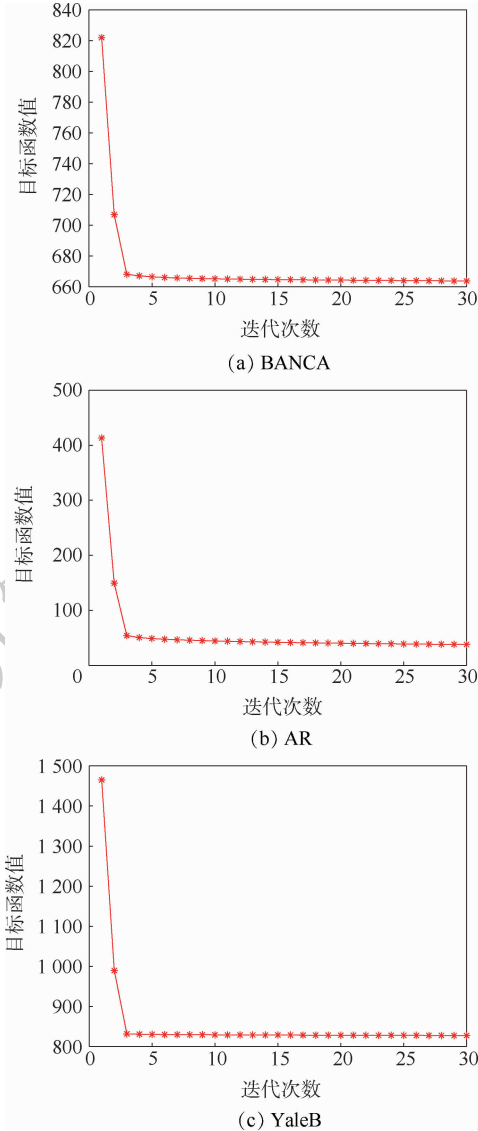


图 2 不同数据集上 BDPR 的函数收敛曲线
Fig.2 Function convergence curves of BDPR method on different datasets

本,其余设置和 2.1 节一致,参数选取范围为 $\{10^{-3}, 10^{-2}, \dots, 10^3\}$ 。鉴于 BDPR 中包含 4 个重要参数,因此采取固定 2 个参数调整另外 2 个参

数的组合遍历测试。图 3 为 BDPR 分类正确率随参数变化的三维柱状图。可知,BDPR 对参数的选择并不敏感。

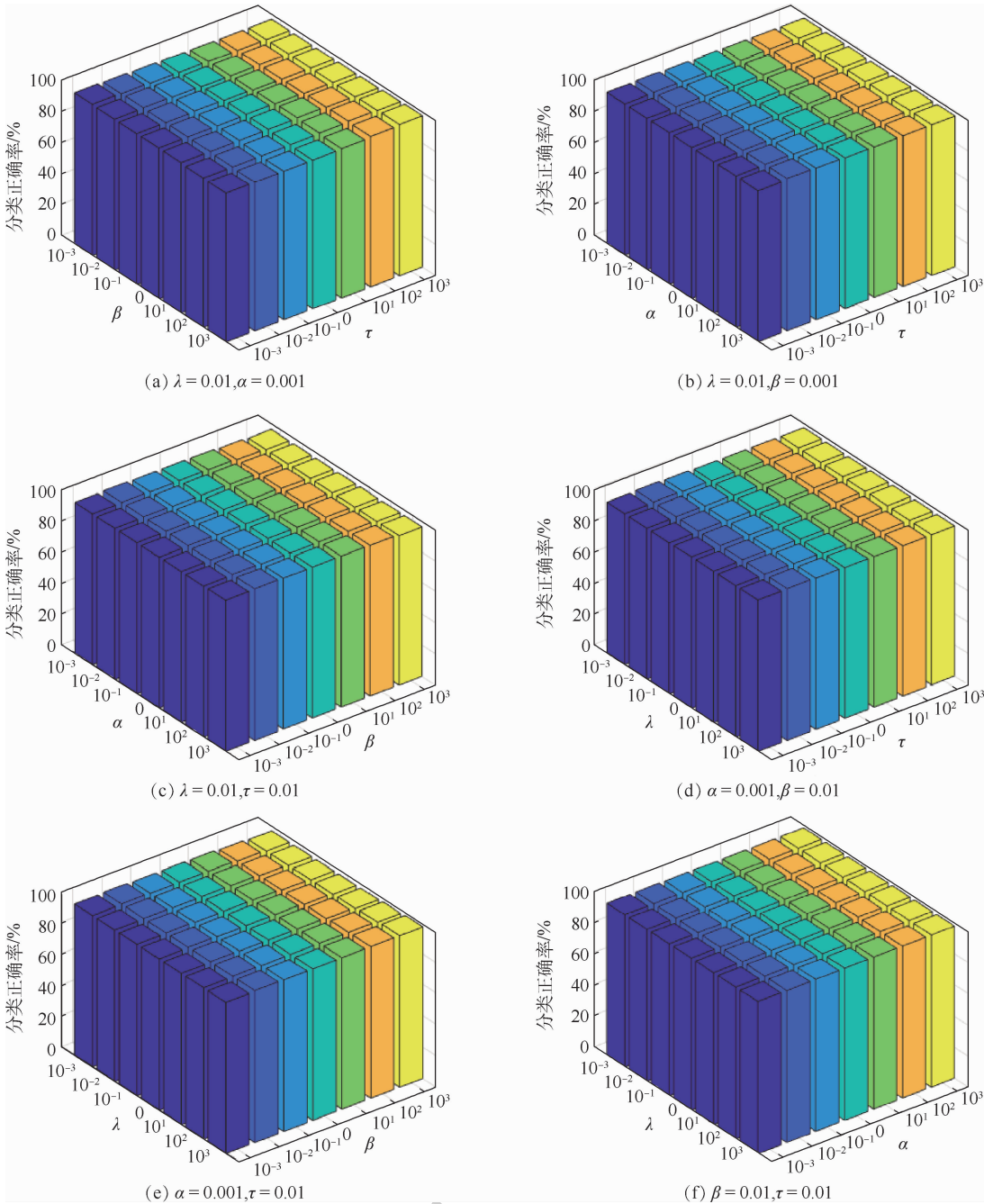


图 3 AR 数据集上 BDPR 分类正确率随参数变化的三维柱状图

Fig. 3 Three-dimensional histogram of classification accuracy of BDPR method changing with different parameters on AR dataset

3 结 论

本文提出了一种基于块对角结构的投影表示(BDPR)学习方法,通过与经典的基于协作表示分类和子空间学习方法相比,可以得到以下结论:

1) BDPR 在传统协作表示的基础上,构造了

编码系数约束项,在增强了编码系数表征数据关系能力的同时,使得噪声和异常值对数据重构的影响最小。

2) 本文方法通过构造编码系数判别逼近项来学习一个判别投影,利用该投影将数据与编码系数相关联,实现了子空间里的最优判别表示。实验结果表明,相比于经典的 CRC 算法,BDPR

所提取的编码系数具有更加明显的块对角化结构,有助于为后续的分类决策过程提供更加丰富的判别信息。

3) BDPR 构建了投影表示学习和分类器学习联合优化的算法框架,学习更适合当前分类任务的线性分类器,避免了传统的两阶段模型在分类阶段的分类器抉择困难和在现实应用中其级联性能可能会下降的问题。

在多个人脸数据集上的测试结果表明,相比于现有的协作表示分类和子空间学习方法,本文方法具有更优的识别效果。

未来工作中,本文方法可以与深度网络相结合,对深度模型输出的高维特征向量进行子空间学习,不仅可以有效降低深度特征的维度,去除冗余信息,而且可以寻找更具判别力的投影表示向量,提升最终分类性能,这也是笔者未来的研究重点。

参考文献 (References)

[1] TURK M, PENTLAND A. Eigenfaces for recognition[J]. Journal of Cognitive Neuroscience, 1991, 3(1) : 71-86.

[2] FISHER R A. The use of multiple measurements in taxonomic problems[J]. Annals of Human Genetics, 1936, 7(2) : 179-188.

[3] 蒋晨琛, 霍宏涛, 冯琦. 一种基于 PCA 的面向对象多尺度分割优化算法[J]. 北京航空航天大学学报, 2020, 46(6) : 1192-1203.

JIANG C C, HUO H T, FENG Q. An object-oriented multi-scale segmentation optimization algorithm based on PCA[J]. Journal of Beijing University of Aeronautics and Astronautics, 2020, 46(6) : 1192-1203 (in Chinese).

[4] 李可, 刘伟, 杜少毅, 等. 基于 PCA 和 WPSVM 的航天器电特性识别方法[J]. 北京航空航天大学学报, 2015, 41(7) : 1177-1182.

LI K, LIU Y, DU S Y, et al. Spacecraft electrical characteristics identification method based on PCA feature extraction and WPSVM[J]. Journal of Beijing University of Aeronautics and Astronautics, 2015, 41(7) : 1177-1182 (in Chinese).

[5] HE X F, YAN S C, HU Y X, et al. Face recognition using Laplacianfaces[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(3) : 328-340.

[6] HE X F, CAI D, YAN S C, et al. Neighborhood preserving embedding[C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2005 : 1208-1213.

[7] SUGIYAMA M. Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis[J]. Journal of Machine Learning Research, 2007, 8(1) : 1027-1061.

[8] YAN S C, XU D, ZHANG B Y, et al. Graph embedding and extensions: A general framework for dimensionality reduction[J]. IEEE Transactions on Pattern Analysis and Machine Intelli-

gence, 2007, 29(1) : 40-51.

[9] BENGIO Y, COURVILLE A, VINCENT P, et al. Representation learning: A review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8) : 1798-1828.

[10] HUANG K, AVIYENTE S. Sparse representation for signal classification[C] // Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2007 : 609-616.

[11] WRIGHT J, YANG A Y, GANESH A, et al. Robust face recognition via sparse representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2) : 210-227.

[12] QIAO L S, CHEN S C, TAN X Y. Sparsity preserving projections with applications to face recognition[J]. Pattern Recognition, 2010, 43(1) : 331-341.

[13] LY N H, DU Q, FOWLER J E. Sparse graph-based discriminant analysis for hyperspectral imagery[J]. IEEE Transactions on Geoscience and Remote Sensing, 2014, 52(7) : 3872-3884.

[14] ZHANG L, YANG M, FENG X C, et al. Sparse representation or collaborative representation: Which helps face recognition? [C] // Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2011 : 471-478.

[15] LY N H, DU Q, FOWLER J E. Collaborative graph-based discriminant analysis for hyperspectral imagery[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2014, 7(6) : 2688-2696.

[16] YANG W K, SUN C Y, ZHENG W M. A regularized least square based discriminative projections for feature extraction[J]. Neurocomputing, 2016, 175 : 198-205.

[17] ZHENG C Y, WANG N N. Collaborative representation with k-nearest classes for classification[J]. Pattern Recognition Letters, 2019, 117 : 30-36.

[18] LI W, DU Q. Collaborative representation for hyperspectral anomaly detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 53(3) : 1463-1474.

[19] KAYA M, BILGE H S. Deep metric learning: A survey[J]. Symmetry, 2019, 11(9) : 1066.

[20] ZHONG G Q, WANG L N, LING X, et al. An overview on data representation learning: From traditional feature learning to recent deep learning[J]. The Journal of Finance and Data Science, 2016, 2(4) : 265-278.

[21] ZHOU X Z, JIN K, XU M, et al. Learning deep compact similarity metric for kinship verification from face images[J]. Information Fusion, 2019, 48 : 84-94.

[22] CHU R H, SUN Y F, LI Y D, 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.

[23] 车畅畅, 王华伟, 倪晓梅, 等. 基于深度学习的航空发动机故障融合诊断[J]. 北京航空航天大学学报, 2018, 44(3) : 621-628.

CHE C C, WANG H W, NI X M, et al. Fault fusion diagnosis of aero-engine based on deep learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2018, 44(3) : 621-628 (in Chinese).

[24] WEN J, ZHANG B, XU Y, et al. Adaptive weighted nonnegative

- low-rank representation [J]. Pattern Recognition, 2018, 81: 326-340.
- [25] FANG X Z, TENG S H, LAI Z H, et al. Robust latent subspace learning for image classification [J]. IEEE Transactions on Neural Networks, 2018, 29(6): 2502-2515.
- [26] BAILLY-BAILLIÈRE E, BENGIO S, BIMBOT F, et al. The BANCA database and evaluation protocol [C] // International Conference on Audio- and Video-based Biometric Person Authentication. Berlin: Springer, 2003: 625-638.
- [27] GEORGHIADES A S, BELHUMEUR P N, KRIEGMAN D J.

From few to many: Illumination cone models for face recognition under variable lighting and pose [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 643-660.

作者简介:

刘保龙 男, 硕士研究生。主要研究方向: 机器学习、模式识别。

王磊 男, 博士, 副教授, 硕士生导师。主要研究方向: 模式识别与分类、机器学习、计算机视觉等。

Block-diagonal projective representation for face recognition

LIU Baolong^{1,2,3}, WANG Yong^{3,4}, LI Danping⁵, WANG Lei^{2,3,*}

(1. Xidian University Qingdao Institute of Computing Technology, Qingdao 266000, China;

2. Key Laboratory of Marine Intelligent Equipment and System of Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China;

3. School of Electronic Engineering, Xidian University, Xi'an 710071, China;

4. The 27th Research Institute of China Electronics Technology Group Corporation, Zhengzhou 450047, China;

5. School of Telecommunications Engineering, Xidian University, Xi'an 710071, China)

Abstract: Most feature representation algorithms are susceptible to noise when mining the internal structure of the high-dimensional data. Meanwhile, their feature learning and classifier design are separated, resulting in the limited classification performance in practice. Aimed at this issue, a new feature representation method, Block-Diagonal Projective Representation (BDPR), is proposed in this paper. First, a weighted matrix is imposed on the coding coefficients of samples over each class. By using such local constraints to enhance the similarity between the coefficients and reduce the impact of noise on coefficient learning, the proposed BDPR can well maintain the internal data structure. Second, to closely correlate the data with their coding coefficients and reduce the difficulty of learning the representation coefficients, we construct a block-diagonal constraint to learn a discriminative projection. In this way, the sample representation coefficients can be obtained in the low-dimensional projected subspace, which contains more global structure information between samples and enjoys lower computational complexity. Finally, the representation learning and classifier learning are integrated into the same framework. By increasing the "label distance" between samples of different classes, BDPR updates the discriminative projection and classifier in an iterative manner. In this way, the most suitable classifier can be found for the current optimal feature representation, making the proposed algorithm automatically realize the classification task. The results of experiments on multiple benchmark face datasets show that BDPR has achieved better recognition performance, compared to traditional collaborative representation based classification and several mainstream subspace learning algorithms.

Keywords: image classification; feature representation; locality constraint; discriminative projection; block-diagonal structure

Received: 2020-08-25; Accepted: 2020-09-04; Published online: 2020-10-22 15:15

URL: kns.cnki.net/kcms/detail/11.2625.V.20201022.1412.001.html

Foundation items: National Key R & D Program of China (2016YFE0207000); National Natural Science Foundation of China (61203137, 61401328); Natural Science Basic Research Program of Shaanxi (2014JQ8306, 2015JM6279)

* **Corresponding author.** E-mail: leiwang@mail.xidian.edu.cn

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0454

基于噪声柯西分布的社交图像标签优化与标注



练连荣, 项欣光*

(南京理工大学 计算机科学与工程学院, 南京 210094)

摘 要: 随着社交网络的快速发展, 带有用户提供标签的社交网络图像呈现爆炸式增长。但是用户提供的标签是不准确的, 存在很多不相关以及错误的标签。这势必会增加相关多媒体任务的困难。针对标签噪声无序性以及常用的高斯分布对标签噪声中大噪声过于敏感的问题, 但是高斯分布对大噪声比较敏感。鉴于此, 采用对各种噪声都具有鲁棒性的柯西分布拟合噪声, 提出了一个基于噪声柯西分布的弱监督非负低秩深度学习 (CDNL) 模型, 通过柯西分布建模标签噪声来获得理想标签, 并利用深度神经网络模块学习视觉特征和理想标签之间的内在联系, 来得到图像对应的正确标签, 从而大幅提高社交网络图像的标签准确率。所提模型不仅可以修正错误标签、补充缺失标签, 也可以对新图像进行标注。在 2 个公开的社交网络图像数据集上进行了验证, 并且与一些最新的相关工作进行了对比, 证实了所提模型的有效性。

关 键 词: 社交标签; 柯西分布; 深度神经网络; 图像标注; 矩阵分解

中图分类号: TB553

文献标志码: A **文章编号:** 1001-5965 (2021) 03-0632-09

近年来随着社交网络的蓬勃发展, 海量的社交网络图像在被网络用户共享和浏览。海量的图像数据使得精确检索变得困难, 故现在迫切需要有效的图像检索技术。基于标签的图像检索是通过建立图像和标签之间的语义关系来进行的。图像标注对图像检索十分重要。目前, 用户提供的标签虽然可以在一定程度上描述视觉内容信息, 但是这些标签是不准确的。如文献[1]中所述, 用户提供的标签只有一半可以描述图像的视觉内容。这是因为现实中社交网络标签通常是不完整、不准确的, 甚至有非常大比例的图像是没有标签的 (文献[2]中 MIRFlickr 数据集中有超过 50% 的图片没有标签)。也就是说, 用户提供的标签是弱监督的。而这会增加相关多媒体任务的困难, 所以通过学习图像视觉信息和标签语义之间的内在联系来提高社交图像的标签质量是非常必要的。

对于社交网络图像的重标注问题, 已提出了

多种解决方法。如文献[3]采用了矩阵分解来最小化噪声, 从而学习图像-标签的内在关系。文献[4]验证了重标注标签矩阵的低秩性, 并且考虑了图像视觉特征的一致性和标签之间的相关性。基于文献[4]中的非负低秩模型, 文献[5-7]通过引入 2 个潜在因子矩阵来分离优化函数, 更好地减小重标注标签和观测标签之间的差别, 并得到更理想的图像-标签关系模型。上述工作均验证了非负低秩模型在社交图像重标注任务中的有效性。

文献[3-5]中都采用了矩阵分解的方法, 并且基于模型泛化能力的考虑, 通常采用平方损失函数作为目标函数。然而这其中隐含的基本假设是社交标签中的噪声是服从高斯分布的^[8]。由于中心极限定理^[9]的特点, 高斯概率密度函数被广泛应用于信号处理和图像分析等领域中。高斯分布可以很好地拟合最常出现的白噪声。但是在现实中, 数据噪声的内在概率分布是未知的, 可能存

收稿日期: 2020-08-24; 录用日期: 2020-08-28; 网络出版时间: 2020-10-16 16:00
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201016.1445.001.html
* 通信作者. E-mail: xgxian@njust.edu.cn

引用格式: 练连荣, 项欣光. 基于噪声柯西分布的社交图像标签优化与标注[J]. 北京航空航天大学学报, 2021, 47(3): 632-640.
LIAN L R, XIANG X G. Social image tag refinement and annotation based on noise Cauchy distribution[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 632-640 (in Chinese).

在各种类型的噪声。然而高斯分布能够很好地拟合小噪声的分布情况,而对大噪声比较敏感^[10]。鉴于上述情况,本文选用对大噪声更加鲁棒的柯西分布建模标签噪声,并由于噪声的稀疏性,采用 l_1 范数来优化噪声矩阵。

另外在图像标注任务中,图像特征对模型训练具有显著意义,具有更深语义的图像特征能显著提升训练效果。在文献[3,5,11]中都采用了传统的特征提取方法(如 Gist、SIFT、HOG),这会导致深层语义的丢失。鉴于此,本文根据文献[12]提出利用预训练好的 Resnet50 网络来提取视觉特征。基于提出的视觉特征和观测标签,本文采用三层卷积神经网络来训练图像-标签模型。此外,为了更快地训练模型,会利用特征矩阵和观测标签矩阵对模型进行预训练。

鉴于此,本文提出了一种基于噪声柯西分布的弱监督非负低秩深度学习(CDNL)模型,来同时解决错误标签重新标注及新图像标注问题。非负低秩模型主要用于对理想标签的优化和对噪声的抑制,通过引入2个潜在因子分离优化目标。柯西分布能够更好地拟合标签噪声,通过优化损失函数,减少理想标签矩阵和观测标签矩阵的差异。另外,为了更好地学习图像和理想标签的内在映射关系,本文选用 Resnet50 网络来提取深层图像特征,以及 CNN 网络来学习特征矩阵和标签矩阵的内在关系。本文实验基于2个广泛使用的社交图像数据集,并且和一些最新的先进工作进行对比,来证实本文方法的有效性。

本文主要贡献如下:

1) 提出了一种基于噪声柯西分布的弱监督非负低秩深度学习模型,创新性地利用柯西分布模型建模标签噪声,学习得到视觉特征和标签语义之间更深层次的内在关系。

2) 通过建立不同尺度的柯西分布模型,模拟标签噪声,选择出最拟合标签噪声的柯西模型,从而更好地减少理想标签矩阵和观测标签矩阵之间的差异。

3) 采用预训练的 Resnet50 模型来提取特征,并利用 CNN 网络来学习特征与标签之间的映射关系。

4) 本文模型不仅能够细化、补全、重标注社交图像,而且可以标注新图像。

1 相关工作

在图像标注领域,许多先前的工作致力于通过学习图像-标签关系模型来提高图像标签质量。

其中文献[13-15]是基于监督模型上的,但是随着用户提供的社交图像呈现爆炸式增加,弱监督图像标注与优化成为图像标注领域的难点,大量工作利用传统方法(如文献[6,11,16-18])来解决这个问题。而随着深度学习的发展,文献[19]率先将深度学习方法引入图像标注;文献[20]将深度学习和 CCA 方法结合;文献[3,5,21-23]将深度学习与矩阵分解相结合;文献[24-26]则在卷积神经网络的基础上提出了新的网络结构。传统方法通过学习标签之间的语义相关性来探索用户提供的标签语义信息。比如在文献[6]中,通过邻居投票策略来研究图像标签之间的相关性。文献[11]通过删除不准确的标签和增加相关标签,来探究视觉相似性和语义相关性之间的一致性。文献[17]通过低阶矩阵分解来解决标签优化问题。文献[18]通过结合低秩矩阵复原和最大似然估计,来对缺失的标签进行补充,对噪声标签进行细化。然而上述工作在图像视觉特征和学习到的低秩标签空间的联系上缺乏有效性,使得准确性受到限制。

而深度学习方法的出现在一定程度上解决了这一问题。文献[19]证明了深度学习方法在图像标注领域是十分有效的,近几年深度学习方法也广泛应用于图像标注任务,如在文献[20]中认为修正、补全、重标注社会标签的过程,其本质是广义协同过滤的过程。广义协同方法即:相似的图像应该具有相似的标签,具有相同标签的图像也应该具有相似的内容。因此应用协同过滤的方法来处理社会标签是可行的,并提出了 CCA 模型来对标签进行标注。文献[3,5]则为了优化非负低秩模型,通过引入潜在因子来分离优化函数,达到分别优化理想标签和噪声标签的目的,并且利用了深度学习模型来训练图像-标签的内在关系。其中文献[5]将基于社会标签的图像标注问题表述为:通过获得低秩的重标注矩阵和稀疏的误差矩阵,来合成观测标签矩阵的问题(由于图像和标签之间的语义相关性,重标注矩阵应当为低秩矩阵,而标注误差不会常出现,所以误差矩阵应该是稀疏的),并且对该问题进行了建模优化。文献[23]通过假设噪声为高斯分布,并结合图像相似性、标签一致性以及矩阵分解模型,来优化基于损失平方和的目标函数。文献[24]利用图卷积神经网络在部分标签上进行学习,并允许不需要附加标签的组级联接,借此来学习图像-标签关系模型。在文献[25]中,提出利用语义图嵌入的跨模态注意力机制来学习语义标签嵌入,并显式地利

用了标签之间的关系,在学习标签嵌入的指导下,生成新颖的跨模态注意力图。文献[26]则在非负低秩矩阵分解的基础上,引入深度学习学习方法学习标签语义图。上述工作均验证了深度学习能够有效地建立起图像特征和标签语义之间的联系,并且文献[3,5,23,26]验证了矩阵分解方法优化社交标签的有效性。然而上述工作默认噪声符合高斯分布,而实际数据的内在概率分布是未知的,不一定符合高斯分布,所以需要噪声建模。

在真实噪声环境中,噪声源往往是多样的。假设噪声是由多个不同概率分布的随机变量加合而成,并且每一个随机变量都是独立的,那么根据中心极限定理^[9],噪声分布随着噪声源数量的上

升趋近于高斯分布^[10]。虽然高斯分布能有效解决小噪声但是对于大噪声过于敏感,而社交图像(见文献[2])属于大噪声。柯西分布的重尾特性可以有效地建模大噪声,并且柯西分布在峰值处平滑,使得其对密集噪声也有很好的效果^[27]。所以本文选用柯西分布来拟合标签噪声。

基于上述工作,本文提出在非负低秩矩阵模型的基础上,基于噪声稀疏性原则,利用柯西分布来对噪声建模(见图1)。选择 Rsenet50 网络提取更深层次的视觉特征,为了更好地联系视觉特征和标签空间,还利用 CNN 网络来训练图像和理想标签矩阵模型,并在训练模型前,对 CNN 网络预训练来减少预测标签和理想标签间的差异。

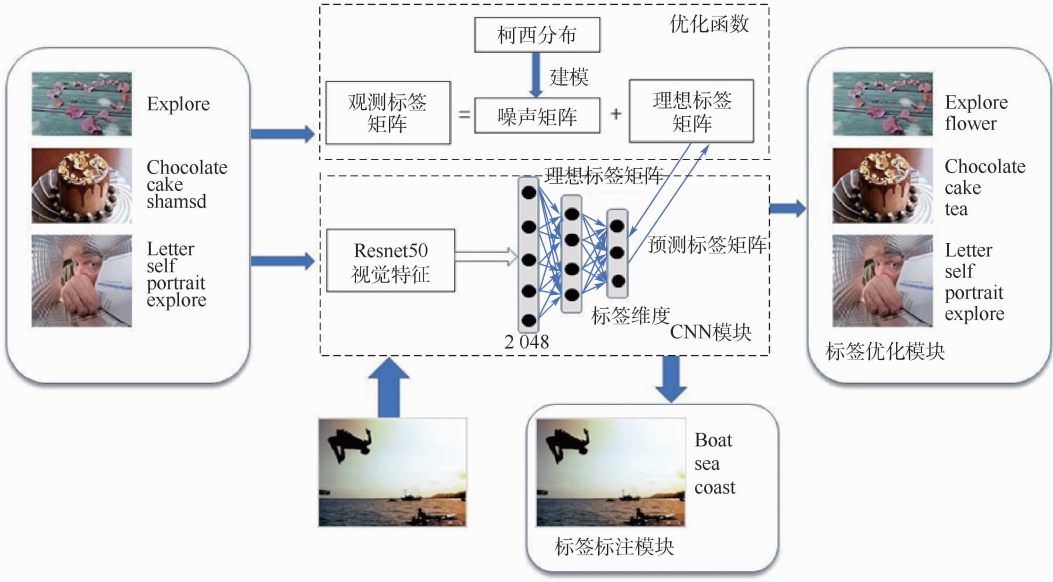


图1 基于噪声柯西分布的社交网络图像标注和重标注模型图

Fig.1 Illustration of the model for social image tag refinement and annotation based on noise Cauchy distribution

2 柯西分布模型

在本节中,将介绍基于噪声柯西分布的弱监督非负低秩深度学习模型(见图1),着重阐述本文研究目的,并对优化函数进行推导。

2.1 研究目的

对于图像标签标注问题,最重要的是要揭示视觉内容和语义标签的内在联系。而对于社交网络图像而言,用户通常会提供一些标签用来标注图像。而这些标签形成的语义空间是可以由真实标签空间中的显式标签子集来近似,而且用户倾向于选择语义相关的标签来对图像进行标记。因此,社交网络图像的标签-图像关联矩阵本质是低秩的。此外,标注过程中不常出现错误,所以噪声矩阵是稀疏的,这是符合常理的。因此,本文采用非负低秩模型来解决社交图像标注中的标签噪声

优化问题。本文选择对各种噪声更加鲁棒的柯西概率分布拟合噪声,由此建立基于噪声柯西分布的弱监督非负低秩模型。图像的视觉特征对图像-标签学习至关重要,因此本文选择 Resnet50 网络来提取深层视觉特征。另外,尽管观测标签矩阵仍然存在很多不相关以及错误的标签,图像的视觉特征和标签仍然是紧密联系的,所以本文采用 CNN 框架来学习图像和标签之间的映射关系。

2.2 模型构建

定义 A 为矩阵, A_{ij} 表示矩阵 A 的第 (i,j) 元素, 矩阵 A 的核范数为 $\|A\|_*$, 矩阵 A 的 l_1 范数为 $\|A\|_1 = \sum_{i=1}^m \sum_{j=1}^n |A_{ij}|$, 而矩阵 A 的 F 范数为 $\|A\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n A_{ij}^2$ 。对于图像标注问题,图像数据

集包括 n 幅图像和 m 个用户标签,每幅图像对应若干个标签构成二值矩阵 $\mathbf{F}$,即观测标签矩阵。定义理想标签矩阵为 $\mathbf{Y}$,特征矩阵为 $\mathbf{X}$,由数据集每幅图像对应的特征向量 $\mathbf{x}_i$ 构成。在低秩框架下,最重要的是优化噪声标签矩阵 $\mathbf{E}$ 、理想标签矩阵 $\mathbf{Y}$ 以及标签预测对应的损失函数。观测标签矩阵由理想标签矩阵和噪声标签矩阵组成:

$$\mathbf{F} = \mathbf{Y} + \mathbf{E} \quad (1)$$

根据文献[28]中的工作,用 $\text{rank}(\mathbf{Y})$ ($\mathbf{Y}$ 的秩)来衡量理想标签矩阵 $\mathbf{Y}$ 的低秩性, $S(\mathbf{E})$ (由不同的噪声优化目标决定)来衡量噪声标签矩阵的稀疏性, $\text{loss}(\mathbf{Y}, \mathbf{W}_g(\mathbf{X}))$ 来衡量预测标签的损失, $\mathbf{W}_g(\mathbf{X})$ 为 CNN 网络预测标签,最后为了防止过拟合问题,再引入正则化项 $\Omega(\Theta)$,并且引入超参数 λ_1, λ_2 以及 λ_3 来构成模型的优化函数:

$$\min_{\mathbf{Y} \geq 0, \mathbf{W}} \text{rank}(\mathbf{Y}) + \lambda_1 S(\mathbf{E}) + \lambda_2 \text{loss}(\mathbf{Y}, \mathbf{W}_g(\mathbf{X})) + \lambda_3 \Omega(\Theta) \quad (2)$$

对于上述优化函数,利用 $\|\mathbf{Y}\|_*$ 来衡量 $\mathbf{Y}$ 的秩,则 $\text{rank}(\mathbf{Y}) = \|\mathbf{Y}\|_*$,而对于 $S(\mathbf{E})$,假设噪声标签矩阵 $\mathbf{E}$ 符合柯西分布,其中 b 为分布的尺度参数:

$$p(E_{ij}) = \frac{b}{\pi} \cdot \frac{1}{b^2 + E_{ij}^2} \quad (3)$$

则对于 $S(\mathbf{E})$ 的优化为

$$\min \sum_{i=1}^m \sum_{j=1}^n \ln(b^2 + (F_{ij} - Y_{ij})^2) = \|\ln(b^2 + (\mathbf{F} - \mathbf{Y})^2)\|_1 \quad (4)$$

对于预测的标签矩阵 $\mathbf{W}_g(\mathbf{X})$ 采用 $\text{loss}(x, y) = \frac{1}{2}(x - y)^2$ 的损失函数进行误差分析,则最后优化目标为

$$\min_{\mathbf{Y} \geq 0, \mathbf{W}} \|\mathbf{Y}\|_* + \lambda_1 \|\ln(b^2 + (\mathbf{F} - \mathbf{Y})^2)\|_1 + \frac{\lambda_2}{2} \|\mathbf{Y} - \mathbf{W}_g(\mathbf{X})\|_F^2 + \lambda_3 \Omega(\Theta) \quad (5)$$

根据文献[3,28],对上述优化问题进行优化,引入2个辅助因子 $\mathbf{Y}_1, \mathbf{Y}_2$ 分离优化问题, $\mathbf{Y}_1$ 用来分析低秩性, $\mathbf{Y}_2$ 辅助 $\mathbf{E}$ 分析稀疏性。得到以下优化问题:

$$\begin{aligned} \min_{\mathbf{Y} \geq 0, \mathbf{W}} & \|\mathbf{Y}_1\|_* + \lambda_1 \|\ln(b^2 + (\mathbf{F} - \mathbf{Y}_2)^2)\|_1 + \\ & \frac{\lambda_2}{2} \|\mathbf{Y} - \mathbf{W}_g(\mathbf{X})\|_F^2 + \lambda_3 \Omega(\Theta) \\ \text{s. t.} & \mathbf{Y}_1 = \mathbf{Y}, \mathbf{Y}_2 = \mathbf{Y}, \mathbf{Y}_2 \geq 0 \end{aligned} \quad (6)$$

再根据非精确增广拉格朗日方法^[28],对应于式(6)中的 $\mathbf{Y}_1$ 和 $\mathbf{Y}_2$,引入 $\mathbf{Z}_1$ 和 $\mathbf{Z}_2$ 到增广拉格朗日方程为

$$\zeta(\mathbf{Y}, \mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Z}_1, \mathbf{Z}_2, \eta) = \|\mathbf{Y}_1\|_* +$$

$$\begin{aligned} & \lambda_1 \|\ln(b^2 + (\mathbf{F} - \mathbf{Y}_2)^2)\|_1 + \frac{\lambda_2}{2} \|\mathbf{Y} - \mathbf{W}_g(\mathbf{X})\|_F^2 + \\ & \frac{\eta}{2} \left(\left\| \mathbf{Y} - \mathbf{Y}_1 + \frac{1}{\eta} \mathbf{Z}_1 \right\|_F^2 + \left\| \mathbf{Y} - \mathbf{Y}_2 + \frac{1}{\eta} \mathbf{Z}_2 \right\|_F^2 \right) - \\ & \frac{1}{2\eta} (\|\mathbf{Z}_1\|_F^2 + \|\mathbf{Z}_2\|_F^2) \end{aligned} \quad (7)$$

式中: ζ 为增广拉格朗日方程符号; η 为梯度下降的学习率。

通过对上述优化函数求解偏导,得到 $\mathbf{Y}, \mathbf{Y}_1, \mathbf{Y}_2$ 的偏导:

$$\arg \min_{\mathbf{Y}} \frac{\lambda_2}{2} \|\mathbf{Y} - \mathbf{W}_g(\mathbf{X})\|_F^2 + \frac{\eta}{2} \left(\left\| \mathbf{Y} - \mathbf{Y}_1 + \frac{1}{\eta} \mathbf{Z}_1 \right\|_F^2 + \left\| \mathbf{Y} - \mathbf{Y}_2 + \frac{1}{\eta} \mathbf{Z}_2 \right\|_F^2 \right)$$

$$\arg \min_{\mathbf{Y}_1} \|\mathbf{Y}_1\|_* + \frac{\eta}{2} \left\| \mathbf{Y} - \mathbf{Y}_1 + \frac{1}{\eta} \mathbf{Z}_1 \right\|_F^2$$

$$\arg \min_{\mathbf{Y}_2} \lambda_1 \|\ln(b^2 + (\mathbf{F} - \mathbf{Y}_2)^2)\|_1 +$$

$$\frac{\eta}{2} \left\| \mathbf{Y} - \mathbf{Y}_2 + \frac{1}{\eta} \mathbf{Z}_2 \right\|_F^2$$

得到 $\mathbf{Y}, \mathbf{Y}_1$ 和 $\mathbf{Y}_2$ 的更新为

$$\mathbf{Y} = \frac{1}{2\eta + \lambda_2} (\lambda_2 + \mathbf{W}_g(\mathbf{X}) - \mathbf{Z}_1 - \mathbf{Z}_2 + \eta \mathbf{Y}_1 + \eta \mathbf{Y}_2) \quad (8)$$

$$\mathbf{Y}_1 = \Gamma_{\frac{1}{\eta}} \left(\mathbf{Y} + \frac{1}{\eta} \mathbf{Z}_1 \right) \quad (9)$$

$$\mathbf{Y}_2 = \mathbf{F} - \text{soft} \left(\mathbf{F} - \frac{1}{\eta} \mathbf{Z}_2 - \mathbf{Y}, \frac{\lambda_1}{\eta}, \sqrt{1 - b^2} \right) \quad (10)$$

式(9)中:定义矩阵 $\mathbf{A} = \mathbf{U}\mathbf{A}\mathbf{V}^T$, $\mathbf{A}$ 为矩阵 $\mathbf{A}$ SVD 分解的奇异值矩阵。传统软阈值算法定义为: $\text{soft}(\mathbf{A}, \delta) = \text{sign}(\mathbf{A}) \max(|\mathbf{A}| - \delta, 0)$, 则定义计算软阈值算子奇异值的方法为 $\Gamma_{\delta} = \mathbf{U}_{\text{soft}}(\mathbf{A}, \delta) \mathbf{V}^T$, δ 为阈值偏差值。式(10)中为更改阈值基准值重定义 $\text{soft}(\mathbf{A}, \delta, \tau) = \text{sign}(\mathbf{A}) \max(|\mathbf{A}| - \delta, \tau)$, τ 表示基准值。

综上所述,首先提出了优化目标函数(式(2)),并对3部分优化目标引入不同的损失函数。根据文献[28],通过非精确增广拉格朗日方法来求解目标函数(式(6)),并对目标函数中的变量求解偏导数,得到变量 $\mathbf{Y}, \mathbf{Y}_1$ 和 $\mathbf{Y}_2$ 的迭代过程,而神经网络参数 $\mathbf{W}$ 则根据 CNN 网络进行更新。根据文献[28],非精确增广拉格朗日方法可以有效地保证该算法的收敛性。

本文通过输入预测标签矩阵 $\mathbf{W}_g(\mathbf{X})$,经过上述的优化过程,可以计算得到理想的标签矩阵 $\mathbf{Y}$ 。再将视觉特征 $\mathbf{X}$ 和理想的标签矩阵 $\mathbf{Y}$ 输入到

CNN 网络中,学习图像-标签关系,并得到新的观测标签矩阵 $W_g(X)$,不断迭代上述过程直至收敛。再利用训练得到图像-标签关系模型,对已有图像重标注,并对新图像标注。下面介绍本文 CDNL 模型的算法流程,CDNL 模型对数据集 MIRFlickr 采用 4 层网络结构,对 NUS-WIDE 采用 3 层网络结构,但这不会影响优化模块的结果。

算法 1 CDNL 算法。

输入:深层视觉特征 X 以及观测标签矩阵 F 。

预训练:将 X 和 F 输入 CNN 网络,设置 Sigmoid 函数为激活函数,二分类交叉熵函数(Binary Cross Entropy)为损失函数,学习率为 0.000 1,指数衰减率为(0.9, 0.999),预先训练 CNN。

准备:载入预训练 CNN,随机正态初始化 Y_1 、 Y_2 、 Z_1 、 Z_2 。

训练:repeat:

$\eta=0.000\ 1,\rho=1.1$

根据式(8)~式(10)更新 Y 、 Y_1 和 Y_2

$Z_1=Z_1+\eta(Y-Y_1)$

$Z_2=Z_2+\eta(Y-Y_2)$

$\eta=\eta\rho$

CNN 模型更新 $W_g(X)$ (参数同上)

Until 模型收敛

输出:理想标签矩阵 Y ,模型参数 W ,评价数据。

使用:利用训练的模型,来重标注训练集,并对测试集中新图片进行标注。

3 实 验

本节将会详细介绍本文方法的实验流程,并且对实验结果进行分析。

实验主要基于 2 个公开的社交网络图像数据集:MIRFlickr 和 NUS-WIDE,这 2 个数据集都在社交网络图像理解和检索任务上得到广泛的应用。MIRFlickr^[29]数据集有 25 000 幅图像,386 个用户标签,去掉出现少于 50 次的标签,得到 457 个实验标签,另外有 18 个真实有效标注用于评价性能。NUS-WIDE^[30]数据集有 269 648 幅图像和 5 018 个用户标签,挑选出现频率最高的 1 000 个标签作为实验标签,160 000 幅图像作为实验数据集。另外有 81 个真实有效标注用于评价。实验中所用数据集的信息如表 1 所示。

为了验证模型的稳定性,从数据集中随机抽取图像作为训练集,并用剩下的图像作为测试集,并且重复实验 5 次。对于 MIRFlickr 数据集每次抽取 12 500 幅图像作为训练集,剩下 12 500 幅图

表 1 MIRFlickr 和 NUS-WIDE 实验数据
Table 1 Experimental data of MIRFlickr and NUS-WIDE

参数	MIRFlickr	NUS-WIDE
图片数量	25 000	269 627
标签数量	457	1 000
真实标签数量	18	81
每次训练集图片数量	12 500	160 000

像作为测试集。对于 NUS-WIDE 数据集每次抽取 160 000 幅图像作为训练集,剩余图像作为测试集。

由于传统方法只能提取图像的底层视觉信息。为了得到图像更深的视觉内容,实验采用 Resne150 残差网络来提取 2 048 维的图像特征。为了使预测标签更接近于理想标签,训练开始时会将深层视觉特征 X 以及观测标签矩阵 F 输入 CNN 来预先训练网络。这样能够更有效地训练 CDNL 模型。进而,不断优化理想标签矩阵 Y ,和网络模型参数 W 。最后利用已训练好的模型来重标注训练集,并标注测试集新图像。

本文对于目标函数中的超参数 λ_1 、 λ_2 以及 λ_3 ,从{0.001,0.01,0.05,0.1,0.5,1,5,10,100,1 000}中依次选择实验,对比实验结果。其中 MIRFlickr 数据集上 λ_1 为 0.1、 λ_2 为 0.2 以及 λ_3 为 0.000 1,NUS-WIDE 数据集上 λ_1 为 0.7、 λ_2 为 0.5 以及 λ_3 为 0.000 1 时最优。

本文采用 MicroAUC、MacroAUC 以及平均精度均值(mAP)这 3 个指标来对模型进行评价。

将 CDNL 模型和前人的工作进行对比,包括 LSCCA^[7]、CCA-CNN^[31]、TCCA^[18]、DMF^[32]、MPMF^[33]、DNMF^[34]、WDMF^[3]、WDNL^[5]、DCE^[23]等模型,其中 LSCCA 和 TCCA 属于传统方法,CCA-CNN 属于利用深度学习方法学习图像-标签关系,DMF、MPMF 以及 DNMF 属于矩阵分解方法,而 WDMF、WDNL 和 DCE 均在标签矩阵分解的基础上学习了图像视觉空间和标签语义空间的关系。

本文提出的 CDNL 模型将从图像重标注和新图像标注 2 个方面进行评价。从图 2 可以看出,CDNL 模型在 MIRFlickr 数据集上效果显著,远高于当前主流方法。从图 3 可以看出,CDNL 模型在 NUS-WIDE 数据集上效果也有提升,仅仅稍差于 DCE。而从表 2 和表 3 可以发现,本文提出的 CDNL 模型在标签重标注和新图像标注上的平均 MicroAUC 和 MacroAUC 均提升明显,相较于其他方法在 MIRFlickr 数据集上有显著优势,在 NUS-WIDE 数据集上也有进步。其中相较于 LSCCA、TCCA 等传统方法,矩阵分解(如 DMF、MPMF、DNMF)有效地提升了标签准确率,但是由于没有

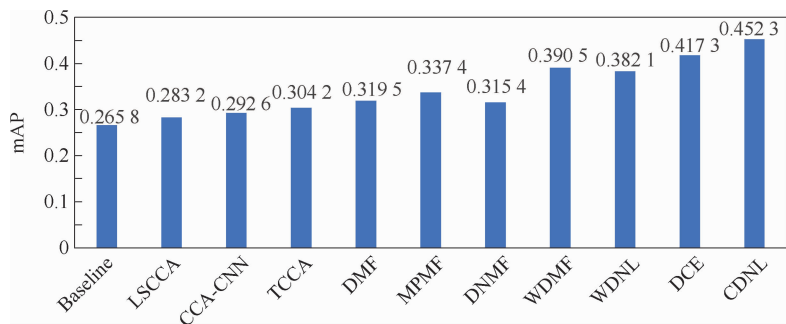


图 2 MIRFlickr 数据集的重标注 mAP

Fig. 2 Experimental results(mAP) on the MIRFlickr for image tag refinement

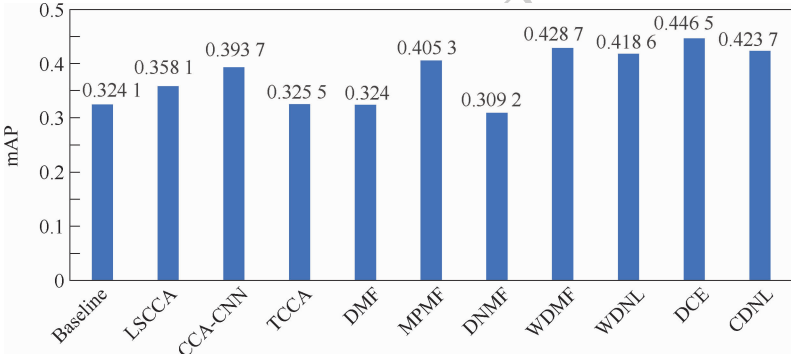


图 3 NUS-WIDE 数据集的重标注 mAP

Fig. 3 Experimental results(mAP) on the NUS-WIDE for image tag refinement

表 2 在 MIRFlickr 和 NUS-WIDE 数据集进行标签重标注的实验结果 (平均 MicroAUC/MacroAUC ± 标准偏差值)

方法	MIRFlickr		NUS-WIDE	
	MicroAUC	MacroAUC	MicroAUC	MacroAUC
Baseline	0.558	0.587	0.623	0.754
LSCCA	0.594 ± 0.006	0.586 ± 0.004	0.732 ± 0.002	0.632 ± 0.003
CCA-CNN	0.644 ± 0.006	0.631 ± 0.003	0.675 ± 0.005	0.743 ± 0.007
TCCA	0.643 ± 0.006	0.632 ± 0.004	0.768 ± 0.007	0.675 ± 0.008
DMF	0.639 ± 0.002	0.628 ± 0.002	0.751 ± 0.005	0.739 ± 0.004
MPMF	0.634 ± 0.004	0.607 ± 0.002	0.782 ± 0.002	0.681 ± 0.005
DNMF	0.624 ± 0.005	0.621 ± 0.006	0.759 ± 0.009	0.665 ± 0.003
WDMF	0.704 ± 0.005	0.678 ± 0.006	0.805 ± 0.004	0.775 ± 0.006
WDNL	0.685 ± 0.003	0.671 ± 0.003	0.789 ± 0.006	0.762 ± 0.006
DCE	0.732 ± 0.003	0.718 ± 0.004	0.825 ± 0.004	0.797 ± 0.003
本文	0.745 ± 0.004	0.775 ± 0.006	0.774 ± 0.009	0.831 ± 0.005

表 3 在 MIRFlickr 和 NUS-WIDE 数据集进行新图像标签标注的实验结果 (平均 MicroAUC/MacroAUC ± 标准偏差值)

方法	MIRFlickr		NUS-WIDE	
	MicroAUC	MacroAUC	MicroAUC	MacroAUC
LSCCA	0.585 ± 0.006	0.562 ± 0.004	0.681 ± 0.002	0.599 ± 0.003
CCA-CNN	0.642 ± 0.005	0.627 ± 0.002	0.617 ± 0.004	0.641 ± 0.003
TCCA	0.610 ± 0.005	0.597 ± 0.006	0.727 ± 0.009	0.625 ± 0.006
DMF	0.635 ± 0.002	0.623 ± 0.003	0.737 ± 0.004	0.632 ± 0.004
MPMF	0.617 ± 0.004	0.596 ± 0.002	0.742 ± 0.002	0.635 ± 0.005
DNMF	0.619 ± 0.005	0.601 ± 0.006	0.699 ± 0.009	0.618 ± 0.003
WDMF	0.661 ± 0.007	0.646 ± 0.004	0.768 ± 0.005	0.675 ± 0.007
WDNL	0.665 ± 0.004	0.652 ± 0.005	0.758 ± 0.004	0.671 ± 0.007
DCE	0.693 ± 0.005	0.667 ± 0.004	0.787 ± 0.006	0.746 ± 0.004
本文	0.715 ± 0.006	0.735 ± 0.005	0.761 ± 0.008	0.794 ± 0.007

考虑图像特征和标签语义的关系,效果明显不如 WDMF、WDNL 和 DCE,而本文方法在引入噪声柯西分布拟合后,对比 WDMF 和 WDNL 方法在 2 个数据集上的效果均有所提升。提升幅度有大有小,这是由于数据集噪声分布是未知的,高斯分布或柯西分布均没有办法完全拟合噪声分布,但是柯西分布相对高斯分布依然有一定优势。由此可以看出,柯西分布对噪声的拟合、Resnet50 对图像深层语义内容的提取以及 CNN 的训练都是卓有成效的,CDNL 模型对比其他主流方法有很大的提升。

而对于噪声柯西分布的具体形式,本实验也做了相关对比工作。在柯西分布中,尺度参数 b 代表了分布中最大值一半处的宽度,并代表着噪声分布的状态,通过调整 b 值分析数据集中噪声的分布特征和状态,在 2 个数据集上的实验结果如表 4 所示。可以看到,MIRFlickr 数据集更加拟合尺度参数 b 为 0.8 的柯西分布,而 NUS-WIDE 数据集更加拟合尺度参数 b 为 0.6 的柯西分布。

从表 4 的实验数据可以看出,提出的模型在 MIRFlickr 数据集上效果远好于前人的工作,并且在 b 为 0.8 的效果最为优异,说明 MIRFlickr 数据集的噪声非常符合柯西分布,且基于噪声柯西分布的深度学习模型是十分有效的。而在 NUS-WIDE 数据集上,本文模型的结果稍逊于 DCE 方法,但是也有所提升,并在 b 为 0.6 时得到最好效果,说明 NUS-WIDE 噪声分布不是十分拟合柯西分布,但是由于柯西分布对大多数噪声的适用性,仍然很好地提升了模型的鲁棒性。

表 4 不同尺度参数 b 对数据集 mAP 性能的影响

Table 4 Experimental results(mAP) on the MIRFlickr and NUS-WIDE for different scale parameter b

b	mAP	
	MIRFlickr	NUS-WIDE
0.2	0.417	0.401
0.4	0.423	0.415
0.6	0.425	0.423
0.8	0.452	0.403

4 结 束 语

本文提出了一种基于噪声柯西分布的弱监督社交网络图像标签修正和标注方法,并利用深度网络学习视觉特征和标签语义之间的关系。本文提出的框架不仅可以对训练图像的用户标签进行修正,同时最后的深度学习模块还可以用来对新

图像进行标签标注。在 2 个公开数据集上的实验结果验证了本文方法的有效性,并且利用不同的尺度参数来对噪声进行拟合,初步得到了 2 个数据集的噪声分布情况。从结果上来说,CDNL 模型对 MIRFlickr 数据集噪声拟合非常好,表现好于主流方法,而对 NUS-WIDE 数据集噪声拟合程度稍差于 MIRFlickr,但也取得非常不错的效果。

在后续的工作中,可以利用更多的分布模型来更好地拟合噪声,尝试利用更为先进的网络模型,来揭示图片深层特征和标签语义的内在关系。

参考文献 (References)

[1] KENNEDY L S, CHANG S F, KOZINTSEV I V. To search or to label? Predicting the performance of search-based automatic image classifiers[C] // Proceedings of the 8th ACM International Conference on Multimedia Information Retrieval. New York: ACM Press, 2006: 249-258.

[2] CHEN M, ZHENG A, WEINBERGER K. Fast image tagging [C] // International Conference on Machine Learning. New York: ACM Press, 2013: 1274-1282.

[3] LI Z, TANG J. Weakly supervised deep matrix factorization for social image understanding[J]. IEEE Transactions on Image Processing, 2016, 26(1): 276-288.

[4] ZHU G, YAN S, MA Y. Image tag refinement towards low-rank, content-tag prior and error sparsity[C] // Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press, 2010: 461-470.

[5] LI Z, TANG J. Weakly-supervised deep nonnegative low-rank model for social image tag refinement and assignment[C] // Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017: 4154-4160.

[6] LI X, SNOEK C G M, WORRING M. Learning social tag relevance by neighbor voting[J]. IEEE Transactions on Multimedia, 2009, 11(7): 1310-1322.

[7] TANG J, SHU X, QI G J, et al. Tri-clustered tensor completion for social-aware image tag refinement[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(8): 1662-1674.

[8] MA H C, LIU I, KING I, et al. Probabilistic factor models for web site recommendation[C] // Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM Press, 2011: 265-274.

[9] PARK S, SERPEDIN E, QARAQE K. Gaussian assumption: The least favorable but the most useful lecture notes [J]. IEEE Signal Processing Magazine, 2013, 30(3): 183-186.

[10] ABDI H, WILLIAMS L J. Principal component analysis[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433-459.

[11] LIU D, HUA X S, WANG M, et al. Image retagging[C] // Proceedings of the 18th ACM International Conference on Multimedia. New York: ACM Press, 2010: 491-500.

- [12] 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.
- [13] DOLAN-GAVITT B,LEEK T,ZHIVICH M,et al. Virtuoso:Narrowing the semantic gap in virtual machine introspection[C]// 2011 IEEE Symposium on Security and Privacy. Piscataway: IEEE Press,2011:297-312.
- [14] BARNARD K,DUYGULU P,FORSYTH D,et al. Matching words and pictures[J]. Journal of Machine Learning Research, 2003,3(2):1107-1135.
- [15] MAKADIA A,PAVLOVIC V,KUMAR S. Baselines for image annotation[J]. International Journal of Computer Vision,2010,90(1):88-105.
- [16] WU L,JIN R,JAIN A K. Tag completion for image retrieval [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2012,35(3):716-727.
- [17] ZHAO R,GROSKY W I. Narrowing the semantic gap-improved text-based web document retrieval using visual features[J]. IEEE Transactions on Multimedia,2002,4(2):189-200.
- [18] FENG Z,FENG S,JIN R,et al. Image tag completion by noisy matrix recovery[C]// Proceedings of the European Conference on Computer Vision. Berlin:Springer,2014:424-438.
- [19] BENGIO Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning,2009,2(1):1-127.
- [20] MURTHY V N,MAJI S,MANMATHA R. Automatic image annotation using deep learning representations[C]// Proceedings of the 5th ACM International Conference on Multimedia Information Retrieval. New York:ACM Press,2015:603-606.
- [21] ZHANG J,WU Q,ZHANG J,et al. Kill two birds with one stone:Weakly-supervised neural network for image annotation and tag refinement[C]// Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence,2018:234-257.
- [22] LI Z,TANG J,ZHANG L,et al. Weakly-supervised semantic guided hashing for social image retrieval[J]. International Journal of Computer Vision,2020,128:2265-2278.
- [23] LI Z,TANG J,MEI T. Deep collaborative embedding for social image understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,41(9):2070-2083.
- [24] BRIAN C,BO W,ALIREZA Z,et al. General partial label learning via dual bipartite graph autoencoder[EB/OL].[2020-07-12]. <https://arxiv.org/abs/2001.01290>.
- [25] YOU R,GUO Z,CUI L,et al. Cross-modality attention with semantic graph embedding for multi-label classification[C]// Proceedings of the Thirty-Forth AAAI Conference on Artificial Intelligence,2020:12709-12716.
- [26] LI Z,TANG J. Weakly supervised deep metric learning for community-contributed image retrieval[J]. IEEE Transactions on Multimedia,2015,17(11):1989-1999.
- [27] DU X,LIU Q,LI Z,et al. Cauchy matrix factorization for tag-based social image retrieval[J]. IEEE Access,2019,7:132302-132310.
- [28] LIN Z,CHEN M,MA Y. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices[EB/OL].[2020-08-20]. <https://arxiv.org/abs/1009.5055>.
- [29] HUISKES M J,LEW M S. The MIR flickr retrieval evaluation [C]// Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval. New York:ACM Press,2008:39-43.
- [30] TANG J,SHU X,LI Z,et al. Generalized deep transfer networks for knowledge propagation in heterogeneous domains[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM),2016,12(4s):1-22.
- [31] MURTHY V N,MAJI S,MANMATHA R. Automatic image annotation using deep learning representations[C]// Proceedings of the 5th ACM International Conference on Multimedia Information Retrieval. New York:ACM Press,2015:603-606.
- [32] GONG Y,JIA Y,LEUNG T,et al. Deep convolutional ranking for multi label image annotation[EB/OL].[2020-08-20]. <https://arxiv.org/abs/1312.4894>
- [33] VERMA Y,JAWAHAR C V. Image annotation using metric learning in semantic neighbourhoods[C]// Proceedings of the European Conference on Computer Vision. Berlin:Springer, 2012:836-849.
- [34] TRIGEORGIS G,BOUSMALIS K,ZAFEIRIOU S,et al. A deep matrix factorization method for learning attribute representations [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2016,39(3):417-429.

作者简介:

练连荣 男,硕士研究生。主要研究方向:社交媒体多标签分类。

项欣光 男,博士,副教授,硕士生导师。主要研究方向:视频处理、压缩与通信、智能媒体分析、图像处理。

Social image tag refinement and annotation based on noise Cauchy distribution

LIAN Lianrong, XIANG Xinguang*

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: With the rapid development of social networks, images with social tags have increased explosively. However, these tags are usually inaccurate and irrelevant which will make it harder for the relevant multimedia tasks. Although label noise is chaotic and disordered, it still conforms to a certain probability distribution. Most of the current methods use Gaussian distribution to fit the noise, but Gaussian distribution is very sensitive to large noise. Thus we use the Cauchy distribution to fit the noise, which is robust to various noises. In this paper, we propose a weakly-supervised Non-negative Low-rank deep learning model based on Cauchy Distribution (CDNL), which builds the noise model by Cauchy distribution to obtain the ideal label and uses deep neural network to reveal the intrinsic connection between the visual features of the image and the ideal labels. The proposed method can not only correct wrong labels and add missing labels, but also tag new images. Experiments are conducted on two public social network image datasets. Compared with some of the latest related work, the results show the effectiveness of the proposed method.

Keywords: social tag; Cauchy distribution; deep neural network; image annotation; matrix factorization

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

DOI: 10.13700/j.bh.1001-5965.2020.0446

结合颅骨形态特征与神经网络的民族判别

全下
文载

杰¹, 赵俊莉^{1,*}, 郑鑫¹, 热孜万古丽·夏米西丁², 李奕¹, 周明全³

青岛大学 数据科学与软件工程学院, 青岛 266071; 2. 新疆师范大学 计算机科学技术学院, 乌鲁木齐 830054;

3. 教育部虚拟现实应用工程研究中心, 北京 100875)

摘 要: 针对颅骨民族判别问题, 提出结合颅骨形态特征与神经网络的判别方法, 可以推进法医人类学的发展, 加快探索民族发展历程。首先, 根据颅骨形态学相关研究, 提取36个维吾尔族和汉族颅骨数据的几何特征; 其次, 采用反向传播神经网络(BPNN)对特征向量进行民族判别, 并通过Adam算法对网络进行优化, 避免陷入局部最优值, 添加正则化项保证算法稳定性; 最后, 分别采用2种网络结构进行对比实验, 输入层、隐藏层和输出层的神经元个数分别为36、6、2和36、12、2, 并设置不同初始学习率进行对比实验。结果表明: 隐藏层神经元个数为12、学习率为0.0001时, 分类精度最高, 测试阶段平均准确率最高为97.5%。为了验证所提方法的普适性, 生成116例国外颅骨数据进行实验, 测试阶段平均准确率为90.96%。相比较于支持向量机(SVM)、决策树、KNN、Fisher等机器学习方法, 所提方法学习能力更强且分类精度有明显提升。

关 键 词: 颅骨形态特征; 反向传播神经网络(BPNN); 颅骨民族判别; 机器学习; Adam 算法

中图分类号: TP183

文献标志码: A

文章编号: 1001-5965(2021)03-0641-09

中国是一个多民族国家, 不同的民族具有独特民族文化和风俗, 由于生存环境、遗传等因素使得各民族相貌特征也存在一定的差异。颅骨是法医人类学、考古学、面貌复原等领域的重要研究对象, 由于颅骨本身骨质坚硬、不易受损的特性, 可以更好地保存特征, 从而通过颅骨识别性别、种族或者民族逐渐成为当前研究热点。在法医人类学领域, 许多刑事案件中尸体受损严重, 很难确认身份, 通过颅骨判别其所属民族可以帮助识别死者身份, 加快案件侦破过程。在考古学领域, 颅骨是重要的出土文物, 颅骨的民族判别有助于了解历史发展进程、挖掘民族文化背景。

国内的人类学家在关于不同民族特征方面已有许多研究, 20世纪80年代, 张振标提出了对藏族^[1]、广西壮族^[2]、海南岛黎族体质特征^[3]的研究, 90年代, 艾琼华^[4]、郑连斌^[5]等分别对维吾尔族体质及宁夏回族体质进行了研究。近些年, 关于不同地区汉族体质特征的研究^[6-9]也逐渐增多。根据相关研究比较发现, 少数民族与汉族面部差异较大, 在鼻子、眼睛部位差异显著^[10]。

目前针对民族判别问题的研究主要集中在二维人脸图像上, 李志洁等^[11]利用C5.0算法对藏族、维吾尔族、壮族3个民族图像的几何特征进行了民族判别, 平均准确率达到90.95%。由于颅

收稿日期: 2020-08-24; 录用日期: 2020-09-05; 网络出版时间: 2020-10-10 17:47

网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201010.1540.002.html

基金项目: 国家自然科学基金(61702293); 全国统计科学研究项目(2020355); 教育部虚拟现实应用工程研究中心基金项目(MEOBNUVRA201601); 中国博士后科学基金(2017M622137); 山东省重点研发计划重大科技创新工程(2019JZZY020101)

* 通信作者: E-mail: zhaojl@yeah.net

引用格式: 孙慧杰, 赵俊莉, 郑鑫, 等. 结合颅骨形态特征与神经网络的民族判别[J]. 北京航空航天大学学报, 2021, 47(3): 641-649. SUN H J, ZHAO J L, ZHENG X, et al. Ethnic identification by combining features of skull morphology with neural network [J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 641-649 (in Chinese).

骨相较于二维人脸图像更为复杂,且更难获取,使得国内外对于不同民族(同一人种)的颅骨数据及分类问题相关研究甚少。本文对采集的汉族和维吾尔族三维颅骨形态特征进行研究,提出结合颅骨几何形态特征与神经网络的方法进行民族判别,主要贡献与创新点如下:

1) 在分析不同民族颅骨特征差异的基础上,提出了颅骨边界环状信息的特征表示方法,选取36维的特征向量作为颅骨民族分类特征。

2) 提出将颅骨边界环状信息作为输入,采用反向传播神经网络(Back-Propagation Neural Network, BPNN)的方法对颅骨进行民族判别,并对网络进行优化。

3) 在400套颅骨数据集上采用2种不同的网络结构进行了实验,并与支持向量机(Support Vector Machine, SVM)、KNN(K-Nearest Neighbor)、决策树、Fisher判别等分类方法进行对比实验,评估本文方法的性能。并生成116例国外颅骨数据,同汉族、维吾尔族数据进行三分类实验,验证了本文方法的普适性。

1 相关工作

颅骨的民族判别问题与颅骨的性别鉴定、种族判别具有一定的相似性,因此可以借鉴相关的研究。传统的性别、种族判别方法主要采用观察法,根据经验知识主观识别,或者提取颅骨特征,通过建立多元判别方程半自动进行分类。周文莲和吴新智^[12]对比中国人和欧洲人颅骨发现犬齿窝、梨状孔上外侧部骨表面膨隆更容易出现在现代欧洲人头骨,而颧骨缘突的出现率几乎相同。Weinberg等^[13]对70例8个月左右的黑种人和白种人胎儿颅骨种族差异研究发现,白种人胎儿的枕骨下部窄于黑种人胎儿,且犁骨较长、前鼻棘突出、鼻骨下骨嵴粗壮、颞鳞边缘呈半圆形。文献[14-17]通过提取不同的特征变量建立判别方程分别进行种族判别,种族判别准确率都在80%以上,其中Iscan和Steyn^[15]对非洲南部的黑种人和白种人建立的判别方程平均准确率最高达到了96.8%。国内^[18,21]及国外^[19,20,22-24]多位学者采用判别分析的方法对不同的颅骨特征进行性别判别。2016年,刘玉勇^[18]对包括鼻宽、乳突间宽等12项特征指标建立判别方程,性别判别识别率可达95%。

近年来,随着计算机技术的快速发展,统计学理论与计算机技术结合越来越密切,机器学习、深度学习的方法开始被广泛应用在分类问题当中,

观察测量提取特征也逐渐发展为根据统计学算法由计算机自动提取特征,使得颅骨性别、种族识别问题趋于自动化。在性别识别问题中,Luo等^[25]和杨稳等^[26]分别利用主成分分析(PCA)建立统计形状模型对获得的颅骨特征向量进行颅骨性别鉴定,识别率均在90%以上。Luo等^[27]则是采用稀疏主成分分析(SPCA)的方法提取颅骨特征找到4个最优区域,最终对性别识别率可达97%左右。杨稳等^[28]将卷积神经网络和最小二乘法相结合进行颅骨性别鉴定,平均正确率可达94.4%。2019年,Yang等^[29]将测量得到的颅矢状弧、颅矢状弦、顶矢状弧、顶矢状弦、枕矢状弧、枕矢状弦6个颅骨特征作为神经网络的输入进行性别鉴定实验,实验测试的准确率达到96.764%。在种族识别问题中,Hefner等^[30]根据随机森林算法对多特征进行种族识别,平均准确率为89.6%。Oakley等^[31]采用与Luo等^[25]同样的特征向量进行种族识别,平均准确率为79%。相较于传统的分类方法,机器学习、深度学习的方法对于噪声鲁棒性强、泛化能力更好,且无需严格的前提假设条件,操作相对简单快捷,更具有应用价值,因而本文将深度学习的方法用于对颅骨民族分类问题进行研究。

2 本文方法

由于颅骨的特殊性,颅骨数据集获取困难,且颅骨数据模型为三维网格数据,模型复杂,包含41 049个点,其中包含大量冗余数据,直接将整个模型送入神经网络进行训练计算量庞大,需要大量训练时间且神经网络模型设计也非常复杂。因而本文首先根据先验知识提取颅骨几何特征,将高维颅骨数据转化为低维特征向量,再利用改进的BPNN方法进行颅骨民族判别。所提方法可以在充分利用颅骨先验知识的基础上,降低计算量、减少训练时间、提高判别速度。本文整体颅骨民族鉴定流程如图1所示。

2.1 颅骨数据获取与预处理

本文所使用的实验数据包括400例无病理颅骨CT原始数据,其中200例汉族头部CT数据来自北京师范大学教育部虚拟现实应用工程研究中心的颅面CT数据库;200例维吾尔族头部CT数据来自新疆托克逊县人民医院放射中心采集的颅面CT数据,所采集的CT图像分辨率均为 512×512 ,颜色深度同为16位。首先,对采集的CT数据提取颅骨轮廓边界^[32],然后利用Marching Cube算法^[33]进行三维模型重建得到三维颅骨和面貌模型,为避免姿态、大小、坐标系的影响,将所

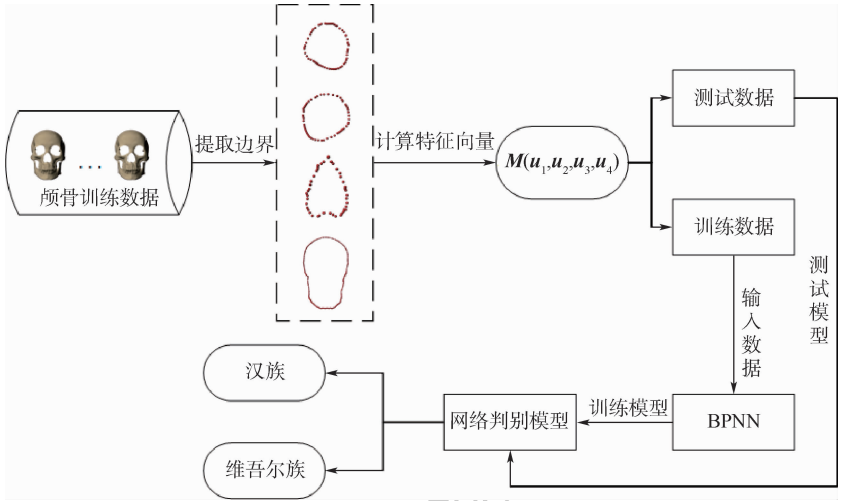


图 1 颅骨民族鉴定流程
Fig. 1 Flowchart of skull ethnic identification

有样本模型调整到同一的法兰克福坐标系下并进行归一化^[34],在此基础上,使用非刚性的数据配准方法进行配准^[35],配准后每个颅骨模型均有 41 059 个点。

2.2 颅骨形态几何特征提取

针对颅骨数据获取困难、数据较少的问题,本文结合形态学研究,提取颅骨的形态几何特征,作为深度学习的输入数据,可以充分利用颅骨民族差异的先验知识,克服数据不足的问题,提高判别的准确率。

2.2.1 颅骨形态特征分析

根据形态学家研究发现^[4],维吾尔族人作为少数民族与汉族人在面部及头部形态上有着明显差异,具体差异如表 1 所示。

通过表 1 可以发现,维吾尔族与汉族面貌形态差异主要集中在鼻子及眼睛、整体颅骨头部轮廓部位,由于人类面貌形态与颅骨是密切相关的,本文颅骨样本选用整体颅骨头部轮廓、鼻子、左眼及右眼的边界信息作为颅骨形态几何特征。

表 1 维吾尔族与汉族面部及头部主要形态差异
Table 1 Main morphological differences of face and head between Uygur and Han

主要特征	维吾尔族	汉族
眼睛	眼裂开度中等、内外平行、眼为楔形结构	眼裂开度中等、倾斜程度水平
鼻子	鼻根中等偏高、鼻尖大多向前、鼻长明显长于其他民族、属狭鼻型	鼻根中等、鼻基部水平、属中鼻型
肤色	肤色呈黄、白或棕褐	黄色居多
头型	特圆头型和超圆头型	特圆头型和圆头型

2.2.2 颅骨几何特征提取与表示

依据颅骨形态特征,在经过配准和预处理后的颅骨模型上,提取数据样本中颅骨头部轮廓、鼻子、左眼及右眼边界中 x 、 y 、 z 三个轴向的最大值、最小值及最大值与最小值间的差值共 36 个特征进行维吾尔族和汉族的民族分类研究。

首先提取颅骨头部轮廓、鼻子、左眼及右眼的边界信息。由于颅骨模型采用三角网格表示,边界只在一个三角形中,可以通过统计边在三角面片中出现的次数提取出边界,具体计算每个颅骨环状边界信息的算法如下所示。

算法 1 颅骨边界信息提取算法。

输入:颅骨样本三角网格数据。

输出:颅骨 k 个环状边界 $D(d_1, d_2, \dots, d_k)$ 。

FOR $i = 1$ to x do

计算颅骨样本中每条边 b 在三角面中出现的次数 j ;

IF $j = 1$ b 为边界边;

ELSE b 不为边界边;

所有边界边存入矩阵 S ;

END

将所有边界边划分到 k 个环状边界信息中

FOR $i = 1$ to k do

WHILE $S \neq \emptyset$ do

将矩阵 S 中第一条边作为环状边界 d_i 的起始边 l_i ;基于起始边 l_i 找到 d_i 中所有的边;

更新 $S: S = S - (S - d_i)$;

END

END

将颅骨头部轮廓、鼻子和眼睛的边界信息可视化后如图 2、图 3 所示。

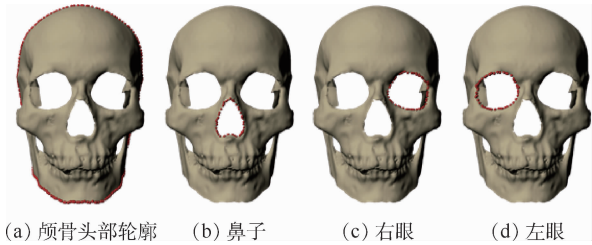


图2 颅骨头部轮廓、鼻子及眼睛环状边界信息
Fig.2 Skull contour, nose and eye ring boundary information



图3 颅骨头部轮廓边界特征示意图
Fig.3 Schematic diagram of skull contour boundary feature

不同民族颅骨头部轮廓、眼睛、鼻子大小形状具有差异性,因此接下来分别计算边界信息中表示大小与形状的信息:包括 x, y, z 三个轴向的最大值、最小值及最大值与最小值间的差值。假设边界向量 B 由 n 个点组成,即

$$B = \{(x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3)\} \quad (1)$$

遍历边界中所有的点,得到 3 个坐标轴向上的最大值和最小值,即

$$\begin{cases} x_{\max} = \max(x_1, x_2, \dots, x_n) \\ y_{\max} = \max(y_1, y_2, \dots, y_n) \\ z_{\max} = \max(z_1, z_2, \dots, z_n) \end{cases} \quad (2)$$

$$\begin{cases} x_{\min} = \min(x_1, x_2, \dots, x_n) \\ y_{\min} = \min(y_1, y_2, \dots, y_n) \\ z_{\min} = \min(z_1, z_2, \dots, z_n) \end{cases} \quad (3)$$

从而可计算出每个轴向上最大值和最小值间的差值:

$$\begin{cases} d_x = x_{\max} - x_{\min} \\ d_y = y_{\max} - y_{\min} \\ d_z = z_{\max} - z_{\min} \end{cases} \quad (4)$$

综上,通过每个颅骨样本的颅骨头部轮廓、鼻子及双眼的边界信息可计算得到特征向量 $M(u_1, u_2, u_3, u_4), u_i = (x_{\max}, x_{\min}, y_{\max}, y_{\min}, z_{\max}, z_{\min}, d_x, d_y, d_z) (i = 1, 2, 3, 4)$, 将特征向量 M , 共 36 个特征作为 BPNN 的输入进行民族分类。

2.3 基于改进的 BPNN 的颅骨民族判别

BPNN 是最基础的神经网络,其通过前向传播输出结果,反向传播误差的方式进行学习^[36],不断调整网络的权值和阈值,从而最小化误差,使

网络尽可能输出正确的预测结果。

2.3.1 网络结构

BPNN 通常包含输入层、隐藏层、输出层 3 层结构,输入层神经元个数通常与输入特征向量维数相同。根据算法提取的每个颅骨训练样本的几何特征向量 M 为 36 维,则设置输入层 s 神经元个数为 36,为进行对比实验,隐藏层 h 神经元个数为 6 和 12,由于输出结果为二分类,即汉族和维吾尔族,所以输出层 o 设置为 2 个神经元。因此本文采用 $[36:6:2]$ 和 $[36:12:2]$ 两种结构进行实验,网络结构示意图如图 4 所示。

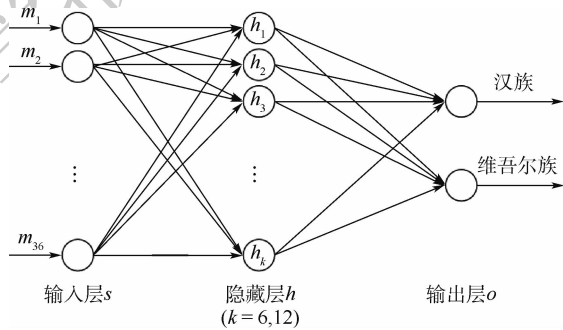


图4 BPNN 网络结构
Fig.4 BPNN network structure

2.3.2 激活函数

为减轻在反向传播过程中梯度消失的现象,加快收敛速度,提高神经网络的稀疏表达能力,隐藏层激活函数选用 Relu 函数,其表达式如下:

$$f(x) = \max(0, x) \quad (5)$$

以隐藏层神经元个数为 6 的网络为例,根据前向传递算法,特征向量 M 通过输入层传递到隐藏层,在隐藏层进行加权求和,然后采用激活函数进行非线性化从而得到隐藏层的输出:

$$v_h = f\left(\sum_{q=1}^{36} w_{sh} x_q\right) \quad (6)$$

式中: w_{sh} 为输入层和隐藏层间的权重值; x_q 为输入特征数值。为了便于分类,输出层激活函数采用 Sigmoid 函数, Sigmoid 函数可以将输出层的输出限制到 $(0, 1)$ 之间:

$$\sigma(a) = \frac{1}{1 + e^{-a}} \quad (7)$$

最终得到训练样本属于汉族和维吾尔族两类的概率,即

$$v_o = \sigma\left(\sum_{j=1}^6 w_{ho} v_j\right) \quad (8)$$

式中: w_{ho} 为隐藏层与输出层间的权重值; v_j 为隐藏层的输出; $v_o \in (0, 1)$ 。

2.3.3 损失函数

交叉熵损失函数在参数更新的过程中不依赖

于激活函数的偏导,当输出层激活函数为 Sigmoid 函数时,相较于均方误差损失函数可持续保持较快的收敛速度,因此采用交叉熵损失函数为

$$C_0 = \frac{1}{N} \sum_{n=1}^N - [y_n \ln p_n + (1 - y_n) \ln(1 - p_n)] \quad (9)$$

式中: N 为样本个数; y_n 为原样本标签值,汉族为 1,维吾尔族为 0; p_n 为预测样本; n 为汉族的概率。在反向传播过程中,假设更新步数为 t ,则权重更新过程如下:

$$\Delta w_{ho} = \frac{\partial C_0}{\partial w_{ho}} = \frac{1}{N} \left[-\frac{y_n}{p_n} + \frac{1 - y_n}{1 - p_n} \right] \frac{\partial \sigma(a)}{\partial w_{ho}} \quad (10)$$

$$\Delta w_{ho} = \frac{1}{N} [\sigma(a) - y_n] v_h \quad (11)$$

$$w_{ho}(t+1) = w_{ho}(t) + \Delta w_{ho} \quad (12)$$

2.3.4 Adam 优化算法

标准的 BPNN 所采用的随机梯度下降算法过度依赖于当前批次的训练数据,导致权重更新不稳定、收敛速度慢且容易陷入局部最优值,本文采用 Adam 算法^[37]对网络进行优化。Adam 是一种自适应优化算法,它将带动量的梯度下降法与 RMSprop(Root Mean Square prop)算法结合加速梯度下降。假设 r 时刻,目标函数 C_0 对参数 θ 的导数为 g_r ,通过计算梯度的一阶矩的估计 $\mu_r = E(g_r)$ 和二阶矩的估计 $\delta_r = E(g_r^2)$,对 μ_r 及 δ_r 矫正: $\hat{\mu}_r = \frac{\mu_r}{1 - \beta_1}$, $\hat{\delta}_r = \frac{\delta_r}{1 - \beta_2}$, β_1 和 β_2 分别为一阶矩衰减系数和二阶矩衰减系数,所以:

$$\theta_{r+1} = \theta_r - \frac{\eta}{\sqrt{\hat{\delta}_r} + \varepsilon} \hat{\mu}_r \quad (13)$$

式中: η 为学习率; ε 为极小常数项,用于维持数值稳定。从而为不同的参数设计独立的自适应性学习率,实现更高效的训练。

2.3.5 加入正则化项 L2

在训练数据较少、特征较多且差异不大的情况下,训练过程中容易产生过拟合的问题,为了避免这个问题,提高网络模型泛化能力,本文通过对代价函数增加 L2 正则化项,增大训练误差而减小测试误差,从而提高对测试集的预测准确率。在增加 L2 正则项后,网络的目标函数变为

$$\min: C_0 + \lambda \|w\|_2^2 \quad (14)$$

式中: λ 为正则系数; w 为权重向量; $\|w\|_2$ 为所有权重值平方和的平方根^[29]。

3 实验结果与分析

本文首先对 400 例颅骨模型进行汉族和维吾

尔族分类实验,其中 200 例维吾尔族,200 例汉族,并将其随机划分,70% 为训练集,30% 为测试集。提取颅骨头部轮廓、鼻子、眼睛共 36 维特征向量作为输入,分别采用 $[36:6:2]$ 和 $[36:12:2]$ 两种网络结构进行实验分析,隐藏层激活函数采用 Relu 函数,输出层激活函数采用 Sigmoid 函数,损失函数选用交叉熵损失。同时为了验证本文方法的普适性,新生成 116 例国外颅骨数据进行分类实验。

3.1 不同初始学习率下的实验结果与分析

在网络模型的训练过程中,网络参数的设置尤为重要,其直接影响网络模型效果的好坏,因此本文对比了不同初始学习率下训练过程中 2 种网络结构损失的变化情况。分别设置初始学习率为 0.01,0.001,0.0001,迭代次数为 100,对训练集训练。图 5 和图 6 分别显示了在 3 个不同学习率下迭代 100 次内 2 种网络结构训练损失的变化。

从图 5、图 6 中可知,不同的学习率导致网络收敛速度不同。整体来看当迭代次数为 50×10^4 时,训练损失已下降到很小的值,趋于平稳,说明

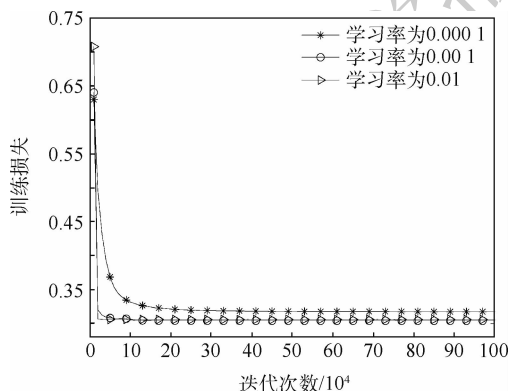


图 5 网络结构 $[36:6:2]$ 下不同学习率的训练损失变化
Fig. 5 Training loss change of network structure $[36:6:2]$ under different learning rates

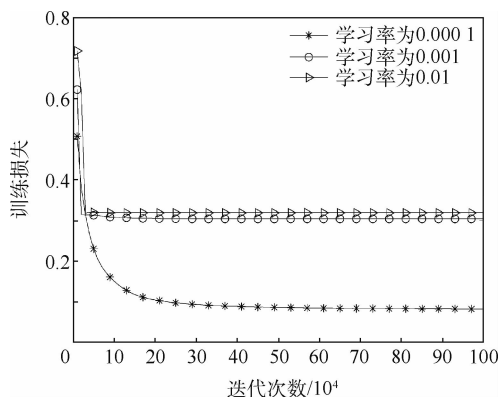


图 6 网络结构 $[36:12:2]$ 下不同学习率的训练损失变化
Fig. 6 Training loss change of network structure $[36:12:2]$ under different learning rates

模型已基本达到稳定状态。

3.2 不同网络结构下的实验结果与分析

网络结构的不同会对分类结果造成一定的影响,因此选择合适的网络结构是实验过程中非常重要的任务。图 7 为当初始学习率为 0.000 1 时,[36:6:2]和[36:12:2]两种结构在训练阶段的平均准确率。

由图 7 可以看出,当初始学习率为 0.000 1 时,网络结构[36:12:2]在训练阶段分类结果优于[36:6:2]。由此证明了网络结构在实验过程中的重要性。最后,表 2 列出了不同参数设置及网络结构在测试阶段的平均准确率。

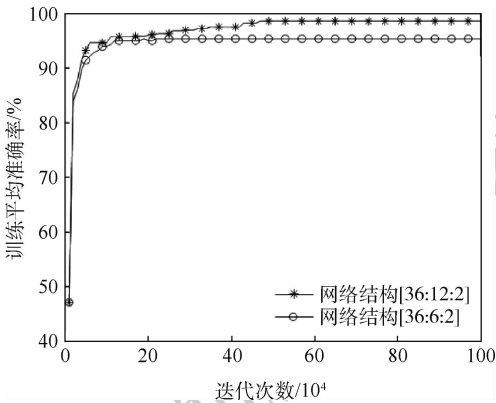


图 7 学习率为 0.000 1 时 2 种网络结构的训练平均准确率
Fig. 7 Training accuracy rate of two network structures when learning rate is 0.000 1

表 2 不同网络结构及参数下的测试平均准确率
Table 2 Test accuracy rate under different network structures and parameters

网络结构	学习率	迭代次数/ 10^4	平均准确率/%
[36:6:2]	0.01	10	95
	0.001	10	94.16
	0.000 1	10	92.5
	0.01	10	93.3
[36:12:2]	0.001	10	94
	0.000 1	10	97.5

可以看出,当初始学习率取 0.000 1,网络结构为[36:12:2]时分类效果最好,准确率达到 97.5%。因而后面的实验均采用这些参数和网络结构进行实验分析。

3.3 不同分类方法分析比较

为了比较本文所提出的改进的 BPNN 传播网络的性能,将网络结构为[36:12:2]、初始学习率为 0.000 1 时的实验结果与采用相同数据进行实验的 SVM、KNN、决策树、Fisher 判别 4 种方法的结果进行了对比。SVM 的核函数选用径向基核函数,KNN 的 K 值设为 3,最终各类方法的平均准确率如表 3 所示。

表 3 不同分类方法的平均准确率
Table 3 Identification rate of different classification methods

方法	平均准确率/%
SVM	77.5
KNN	84.1
Fisher	57.5
决策树	91.6
本文	97.5

从表 3 中可以看出,本文方法准确率明显高于传统的线性判别分析方法,相较于其他机器学习的方法也有一定提升。

为进一步评估不同分类器的性能,图 8 显示了各个分类方法的 ROC (Receiver Characteristic Operator) 曲线。可知在使用相同实验数据进行实验时,本文采用的颅骨形态几何特征与神经网络结合的方法的分类性能优于其他分类方法,说明神经网络的方法对于颅骨这种复杂数据具有更好的鲁棒性,能够取得良好的实验结果。

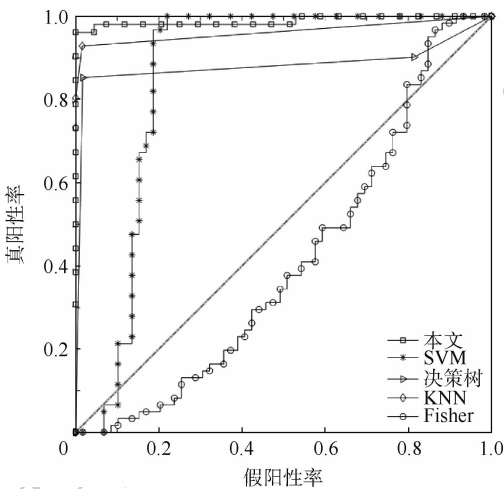


图 8 不同分类方法的 ROC 曲线
Fig. 8 ROC curves of different classification methods

3.4 扩展应用到新生成的国外颅骨数据

为了验证本文方法具有一定的普适性,新生成 116 例国外颅骨进行实验。116 例颅骨是采用 PCA 颅骨复原的方法^[38-40]对来自 Texas 3D Face Recognition Database 数据集^[41]中的人脸深度图复原所得,该数据集包含使用立体相机获取的成年人的人脸深度图及人脸图像,本文选取了表情中性、姿势标准的 116 张人脸深度图进行颅骨复原。首先,将已有颅骨对应的三维人脸模型转化为深度图;然后,用三维颅骨模型和对应人脸深度图数据作为训练样本,用国外的 116 张人脸深度图作为测试数据,利用 PCA 复原出对应颅骨三维模型;最后,得到 116 例与汉族及维吾尔族数据点数一致、姿态统一的完整颅骨模型。

在此部分对共 516 例颅骨模型提取 36 维颅骨边界特征向量送入 BPNN 进行分类实验,其中 361 例作为训练集,155 例作为测试集,由于最后输出结果分为 3 类,因此采用[36:12:3]的网络结构,输出层包含 3 个神经元,初始学习率设置为 0.0001。在迭代 100 次后,其混淆矩阵如图 9 所示。

由混淆矩阵可以看出,模型对于汉族颅骨数据的识别率较高,维吾尔族和国外颅骨数据的识别难度较大。模型在测试集上的平均准确率为 90.96%,说明将颅骨形态几何特征与神经网络结合的方法可以较好地对复杂的颅骨数据进行民族分类,且对于不同民族及种族数据具有一定的普适性。

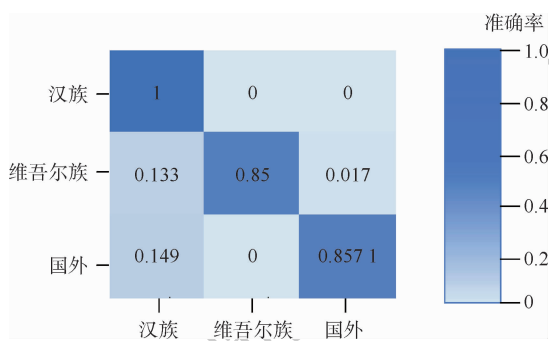


图9 汉族、维吾尔族和国外颅骨分类的混淆矩阵
Fig.9 Confusion matrix of Han, Uighur and foreign skull ethnic identification

4 结 论

1) 本文提出了基于改进的 BPNN 进行民族判别的方法,选取了维吾尔族和汉族人面貌差异较大的颅骨头部轮廓、眼睛、鼻子部位的信息作为主要特征,采用 BPNN 的方法进行判别,平均准确率最高达到 97.5%。

2) 传统观察方法过于依赖专家经验,使得识别结果不够可靠具有一定主观性,而线性分类方法在选取颅骨特征时要求严格、过程复杂、泛化能力不强。深度学习的方法对于噪声的鲁棒性强,针对结构复杂的颅骨数据可以更好地学习特征,其模型泛化能力也优于其他分类方法,因而在增加国外颅骨数据后,本文采用的结合颅骨几何特征与神经网络的颅骨民族判别方法仍取得了良好的效果。

3) 网络参数设置及网络结构的调整是深度学习过程中的重要任务,对学习特征和最终实验结果都有着很大影响,而往往这一任务非常复杂繁琐。神经网络架构搜索(Neural Architecture

Search,NAS)通过定义搜索空间,根据搜索策略评估得到最优网络结构,从而大大减少人工构建网络的工作量。

因此在未来的工作中,会继续对深度学习方法进行更深层次的探究,尝试使用 NAS 的方式自动构建网络,更快速高效地找到最优网络结构,从而提高识别率,形成更完善、更具有应用意义的自动化识别方法。

参考文献 (References)

- [1] 张振标. 藏族的体质特征[J]. 人类学学报, 1985, 4(3): 250-258.
ZHANG Z B. The physical characters of Zang(Tibetan) nationality[J]. Acta Anthropologica Sinica, 1985, 4(3): 250-258 (in Chinese).
- [2] 张振标, 张建军. 广西壮族体质特征[J]. 人类学学报, 1983, 2(3): 260-271.
ZHANG Z B, ZHANG J J. Physical characters of Zhuang nationality in Guangxi[J]. Acta Anthropologica Sinica, 1983, 2(3): 260-271 (in Chinese).
- [3] 张振标, 张建军. 海南岛黎族体质特征之研究[J]. 人类学学报, 1982, 1(1): 53-71.
ZHANG Z B, ZHANG J J. Anthropological studies on Li nationality in Hainan island[J]. Acta Anthropologica Sinica, 1982, 1(1): 53-71 (in Chinese).
- [4] 艾琼华, 肖辉, 赵建新, 等. 维吾尔族的体质特征研究[J]. 人类学学报, 1993, 12(4): 63-71.
AI Q H, XIAO H, ZHAO J X, et al. A survey on physical characteristics of Uigur nationality[J]. Acta Anthropologica Sinica, 1993, 12(4): 63-71 (in Chinese).
- [5] 郑连斌, 朱钦, 王巧玲, 等. 宁夏回族体质特征研究[J]. 人类学学报, 1997, 16(1): 11-21.
ZHENG L B, ZHU Q, WANG Q L, et al. The physical characteristic of Hui nationality in Ningxia[J]. Acta Anthropologica Sinica, 1997, 16(1): 11-21 (in Chinese).
- [6] 张兴华, 郑连斌, 宇克莉, 等. 山东寿光汉族体质特征[J]. 人类学学报, 2011, 30(2): 95-106.
ZHANG X H, ZHENG L B, YU K L, et al. A study of the physical characteristics of the Han people from Shouguang in Shandong[J]. Acta Anthropologica Sinica, 2011, 30(2): 95-106 (in Chinese).
- [7] 李咏兰, 陆舜华, 郑连斌, 等. 江西汉族体质特征[J]. 解剖学报, 2012, 43(1): 114-122.
LI Y L, LU S H, ZHENG L B, et al. Physical characteristics of the Han people in Jiangxi[J]. Acta Anatomica Sinica, 2012, 43(1): 114-122 (in Chinese).
- [8] 郑连斌, 武亚文, 张兴华, 等. 四川汉族体质特征[J]. 解剖学报, 2011, 42(5): 695-702.
ZHENG L B, WU Y W, ZHANG X H, et al. Physical characteristics of Han in Sichuan[J]. Acta Anatomica Sinica, 2011, 42(5): 695-702 (in Chinese).
- [9] 李咏兰, 陆舜华, 郑连斌, 等. 浙江汉族的体质特征[J]. 解剖学报, 2013, 44(5): 707-716.

- LI Y L, LU S H, ZHENG L B, et al. Physical characteristics of Zhejiang Han [J]. *Acta Anatomica Sinica*, 2013, 44 (5) : 707-716 (in Chinese).
- [10] 肖明. 民族形象与解剖结构—西北地区几个少数民族外貌特征的探索 [J]. *西北民族大学学报 (哲学社会科学版)*, 1980 (1) : 86-93.
- XIAO M. National image and anatomical structure - A study on the physical features of several ethnic minorities in Northwest China [J]. *Journal of Northwest University for Nationalities (Philosophy and Social Science)*, 1980 (1) : 86-93 (in Chinese).
- [11] 李志洁, 段晓东, 王存睿. 中国 6 个民族的面部几何特征聚类分析 [J]. *大连民族学院学报*, 2015, 17 (1) : 73-76.
- LI Z J, DUAN X D, WANG C R. Cluster analysis of facial geometric features for six Chinese nationalities [J]. *Journal of Dalian Nationalities University*, 2015, 17 (1) : 73-76 (in Chinese).
- [12] 周文莲, 吴新智. 现代人头骨面部几项非测量性状的观察 [J]. *人类学学报*, 2001, 20 (4) : 42-48.
- ZHOU W L, WU X Z. Observations of some non-metric Traits in the modern human skulls [J]. *Acta Anthropologica Sinica*, 2001, 20 (4) : 42-48 (in Chinese).
- [13] WEINBERG S M, PUTZ D A, MOONEY M P, et al. Evaluation of non-metric variation in the crania of black and white perinates [J]. *Forensic Science International*, 2005, 151 (2-3) : 177-185.
- [14] HOLLAND T D. Race determination of fragmentary crania by analysis of the cranial base [J]. *Journal of Forensic Science*, 1986, 31 (2) : 719-725.
- [15] ISCAN M Y, STEYN M. Craniometric determination of population affinity in South Africans [J]. *International Journal of Legal Medicine*, 1999, 112 (2) : 91-97.
- [16] HOLIDAY T W, FALSETTI A B. A new method for discriminating African-American from European-American skeletons using postcranial osteometrics reflective of body shape [J]. *Journal of Forensic Science*, 1999, 44 (5) : 926-930.
- [17] GILL G W, HUGHES S S, BENNETT S M, et al. Racial identification from the midfacial skeleton with special reference to American Indians and whites [J]. *Journal of Forensic Science*, 1988, 33 (1) : 92-99.
- [18] 刘玉勇. 华北地区汉族成人面颅骨 X 线片性别判定的研究 [J]. *中国司法鉴定*, 2016, 84 (1) : 26-31.
- LIU Y Y. The sex determination of Han nationality in North China by adult facial skull X-ray [J]. *Chinese Journal of Forensic Sciences*, 2016, 84 (1) : 26-31 (in Chinese).
- [19] KHAITAN T, KABIRAJ A, GINJUPALLY U, et al. Cephalometric analysis for gender determination using maxillary sinus index: A novel dimension in personal identification [J]. *International Journal of Dentistry*, 2017: 1-4.
- [20] FRANKLIN D, FREEDMAN L, MILNE N. Sexual dimorphism and discriminant function sexing in indigenous South African crania [J]. *Homo-Journal of Comparative Human Biology*, 2005, 55 (3) : 213-228.
- [21] 税午阳, 殷荣超, 周明全, 等. 中国汉族人颅骨数字模型的性别判别方法 [J]. *中国法医学杂志*, 2013, 28 (6) : 19-21.
- SHUI W Y, YIN R C, ZHOU M Q, et al. Sex determination from digital skull model for the Han people in China [J]. *Chinese Journal of Forensic Medicine*, 2013, 28 (6) : 19-21 (in Chinese).
- [22] AMORES-AMPUERO A. Sexual dimorphism in base of skull [J]. *Anthropologischer Anzeiger*, 2017, 74 (1) : 9-14.
- [23] SMALL C, SCHEPARTZ L, HEMINGWAY J, et al. Three-dimensionally derived interlandmark distances for sex estimation in intact and fragmentary crania [J]. *Forensic Science International*, 2018, 287 (2) : 127-135.
- [24] MAASS P, FRIEDING L J. Morphometric analysis of the neurocranium in an adult South African cadaveric sample [J]. *Journal of Forensic Sciences*, 2019, 64 (2) : 367-374.
- [25] LUO L, WANG M Y, DUAN F Q, et al. Automatic sex determination of skulls based on a statistical shape model [J]. *Computational and Mathematical Method in Medicine*, 2013: 1-6.
- [26] 杨稳, 刘晓宁, 朱菲, 等. 基于统计形状模型的颅骨自动性别识别 [J]. *计算机科学*, 2019, 46 (6) : 282-287.
- YANG W, LIU X N, ZHU F, et al. Automatic sex determination of skulls based on statistical shape model [J]. *Computer Science*, 2019, 46 (6) : 282-287 (in Chinese).
- [27] LUO L, CHANG L, LIU R, et al. Morphological investigations of skulls for sex determination based on sparse principal component analysis [C] // *Proceedings of the Chinese Conference on Biometric Recognition*. Berlin: Springer, 2013: 449-456.
- [28] 杨稳, 刘晓宁, 刘雄乐, 等. 结合改进卷积神经网络和最小二乘法的颅骨性别鉴定 [J]. *人类学学报*, 2019, 38 (2) : 265-275.
- YANG W, LIU X N, LIU X L, et al. Skull sex identification using improved convolution neural network and least squares method [J]. *Acta Anthropologica Sinica*, 2019, 38 (2) : 265-275 (in Chinese).
- [29] YANG W, LIU X N, WANG K G, et al. Sex determination of three-dimensional skull based on improved back propagation neural network [J]. *Computational and Mathematical Methods in Medicine*, 2019: 1-8.
- [30] HEFNER J T, SPRADLEY M K, ANDERSON B. Ancestry assessment using random forest modeling [J]. *Journal of Forensic Sciences*, 2014, 59 (3) : 583-589.
- [31] OAKLEY C, BAI L, LIAO I Y, et al. A novel method for race determination of human skulls [C] // *International Conference on Pattern Recognition*. Berlin: Springer, 2018: 89-102.
- [32] DUAN F, YANG Y, LI Y, et al. Skull identification via correlation measure between skull and face shape [J]. *IEEE Transactions on Information Forensics and Security*, 2014, 9 (8) : 1322-1332.
- [33] LORENSEN W E, CLINE H E. Marching cubes: A high resolution 3D surface construction algorithm [J]. *ACM SIGGRAPH Computer Graphics*, 1987, 21 (4) : 163-169.
- [34] HU Y, DUAN F, YIN B, et al. A hierarchical dense deformable model for 3D face reconstruction from skull [J]. *Multimedia Tools and Applications*, 2013, 64 (2) : 345-364.
- [35] HUANG R, ZHAO J, DUAN F, et al. Automatic craniofacial registration based on radial curves [J]. *Computers & Graphics*, 2019, 82 (5) : 264-274.

- [36] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning representations by back-propagating errors [J]. *Nature*, 1986, 323(6088):533-536.
- [37] KINGMA D, BA J. Adam: A method for stochastic optimization [C]//*Proceedings of the 3rd International Conference for Learning Representations*, 2015:1-15.
- [38] BERAR M, DESVIGNES M, BAILLY G, et al. 3D semi-landmarks based statistical face reconstruction [J]. *Journal of Computing & Information Technology*, 2006, 14(1):31-43.
- [39] SHUI W Y, ZHOU M Q, MADDOCK S, et al. A PCA-based method for determining craniofacial relationship and sexual dimorphism of facial shapes [J]. *Computers in Biology and Medicine*, 2017, 90(8):33-49.
- [40] 王琳, 赵俊莉, 段福庆, 等. 颅面复原方法综述 [J]. *计算机工程*, 2019, 45(12):8-18.
- [41] WANG L, ZHAO J L, DUAN F Q, et al. Survey on craniofacial-reconstruction method [J]. *Computer Engineering*, 2019, 45(12):8-18 (in Chinese)
- [42] GUPTA S, CASTLEMAN K R, MARKEY M K, et al. Texas 3D face recognition database [C]//*Proceedings of the Image Analysis & Interpretation*. Piscataway: IEEE Press, 2010:97-100.

作者简介:

孙慧杰 女, 硕士研究生。主要研究方向: 计算机图形学。

赵俊莉 女, 博士, 副教授, 硕士生导师。主要研究方向: 计算机图形学、计算机视觉、虚拟现实。

Ethnic identification by combining features of skull morphology with neural network

SUN Huijie¹, ZHAO Junli^{1,*}, ZHENG Xin¹, REZIWANGULI Xiamixiding², LI Yi¹, ZHOU Mingquan³

(1. School of Data Science and Software Engineering, Qingdao University, Qingdao 266071, China;

2. School of Computer Science and Technology, Xinjiang Normal University, Urumqi 830054, China;

3. Virtual Reality Research Center of Ministry of Education, Beijing 100875, China)

Abstract: China is a multi-ethnic country. It is of great significance for the skull identification to realize the skull ethnic identification through computers, which can promote the development of forensic anthropology and exploration of national development. Firstly, according to the skull morphology studies, 36 Uighur and Han geometric features of the skull data are extracted, and the Back-Propagation Neural Network (BPNN) of feature vectors is used for ethnic identification. In order to optimize the network, Adam algorithm is adopted to avoid falling into local minimum, and to ensure the stability of the algorithm with regularization terms. Two network structures are used for comparative experiments. The number of neurons in the input layer, hidden layer and output layer are 36, 6, 2 and 36, 12, 2, respectively, and different initial learning rates are set for comparative experiments. The results show that, when the number of hidden-layer neurons is 12 and the learning rate is 0.0001, the classification accuracy is the highest and the highest accuracy rate in the test stage is 97.5%. In order to verify the universality of the method in this paper, 116 foreign skull data are generated for experiments, and the accuracy rate of the test stage is 90.96%. Compared with machine learning methods such as Support Vector Machine (SVM), decision-making tree, KNN, and Fisher, the proposed method has stronger learning ability and significantly improved classification accuracy.

Keywords: features of skull morphology; Back-Propagation Neural Network (BPNN); skull ethnic identification; machine learning; Adam algorithm

Received: 2020-08-24; **Accepted:** 2020-09-05; **Published online:** 2020-10-10 17:47

URL: kns.cnki.net/kcms/detail/11.2625.V.20201010.1540.002.html

Foundation items: National Natural Science Foundation of China (61702293); National Statistical Science Research Project (2020355); the Open Research Fund of the Ministry of Education for Engineering Research Center of Virtual Reality Application (MEOB-NUEVRA201601); China Postdoctoral Science Foundation (2017M622137); Key Research and Development Plan-Major Scientific and Technological Innovation Projects of ShanDong Province (2019JZZY020101)

* **Corresponding author.** E-mail: zhaohl@yeah.net

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0447

融合语义信息的视频摘要生成



滑蕊, 吴心筱*, 赵文天

(北京理工大学 计算机学院, 北京 100081)

摘 要: 视频摘要任务旨在通过生成简短的视频片段来表示原视频的主要内容, 针对现有方法缺乏对语义信息探索的问题, 提出了一种融合语义信息的视频摘要生成模型, 学习视频特征使其包含丰富的语义信息, 进而同时生成描述原始视频内容的视频摘要和文本摘要。该模型分为 3 个模块: 帧级分数加权模块、视觉-语义嵌入模块、视频文本描述生成模块。帧级分数加权模块结合卷积网络与全连接层以获取帧级重要性分数; 视觉-语义嵌入模块将视觉特征与文本特征映射到同一空间, 以使 2 种特征相互靠近; 视频文本描述生成模块最小化视频摘要的生成描述与文本标注真值之间的距离, 以生成带有语义信息的视频摘要。测试时, 在获取视频摘要的同时, 该模型获得简短的文本摘要作为副产品, 可以帮助人们更直观地理解视频内容。在 SumMe 和 TVSum 数据集上的实验表明: 该模型通过融合语义信息, 比现有先进方法取得了更好的性能, 在这 2 个数据集上 F -score 指标分别提高了 0.5% 和 1.6%。

关 键 词: 视频摘要; 视觉-语义嵌入空间; 视频文本描述; 视频关键帧; 长短期记忆 (LSTM) 模型

中图分类号: TP391

文献标志码: A

文章编号: 1001-5965(2021)03-0650-08

随着视频拍摄、存储技术和网络传输的飞速发展, 互联网上的视频数据呈爆炸性增长^[1]。但由于生活节奏越来越快, 观众在没有确定视频是否符合他们的期望前, 不会轻易花太多时间观看完整视频, 观众更期望可以通过视频预告等形式对视频内容产生大致的了解。视频摘要任务从原始视频中提取具有代表性和多样性的简短摘要, 使观看者在不观看完整视频的情况下, 快速掌握视频的主要内容。

早期的视频摘要方法^[2-3]是基于无监督学习的, 使用诸如外观或运动特征等底层视觉信息和聚类方法来提取视频摘要。近年来, 深度神经网络^[4-8]被用于视频摘要任务中, 其主要致力于学习更具表示能力的视觉特征。这些方法将视频摘要任务建模为视频镜头子集的挑选, 通过使用长

短期记忆 (Long Short-Term Memory, LSTM) 单元来实现视频摘要的生成。

随着计算机视觉与自然语言处理的融合发展, 视频文本描述等许多跨模态任务引起了研究人员的广泛关注。所以自然而然联想到通过添加文本内容到视频摘要任务中, 利用语义信息对视觉任务进行监督从而获取视频摘要, 部分近期研究^[9-11]也开始研究视频摘要中包含的语义信息。文献[9]使用了视频主题或视频标题来提供视觉上下文信息, 以标识原始视频包含的重要语义信息。文献[10-11]结合了自然语言处理领域中常用的一些方法, 将语义信息添加到视频摘要中。但这些工作提出的框架相对简单, 没有对语义信息进行充分利用。

本文提出一种融合语义信息的视频摘要生成

收稿日期: 2020-08-24; 录用日期: 2020-09-27; 网络出版时间: 2020-10-15 10:32
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201015.1011.002.html
基金项目: 国家自然科学基金 (61673062, 62072041)
* 通信作者: E-mail: wuxinxiao@bit.edu.cn

引用格式: 滑蕊, 吴心筱, 赵文天. 融合语义信息的视频摘要生成[J]. 北京航空航天大学学报, 2021, 47(3): 650-657.

HUA R, WU X X, ZHAO W T. Video summarization by learning semantic information[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 650-657 (in Chinese).

模型,与现有的视频摘要方法不同,该模型在训练时,增加文本监督信息,从原始视频镜头中挑选出具有语义信息的镜头。首先,利用卷积网络提取视频特征并获得每帧对应的帧级重要性分数。其次,利用文本监督信息学习视觉-语义嵌入空间,将视觉特征和文本特征投影到嵌入空间中,通过计算跨域数据之间的相似性并使其靠近,从而使视觉特征可以学习更多的语义信息。本文还使用视频文本描述生成模块生成视频对应的文本摘要,并通过与文本标注真值进行比较来优化视频的帧级重要性分数。该模块在测试时仍可用,可以为测试视频生成对应的文本摘要。在测试时,现有的视频摘要方法只能获得视频摘要,而本文模型在获得具有语义信息的视频摘要的同时,还可以获得相应的文本摘要,可以更加直观地反映视频内容。

本文的主要贡献如下:

1) 提出了一种融合语义信息的视频摘要生成模型,通过学习视觉-语义嵌入空间丰富视觉特征的语义信息,以确保视频摘要最大限度地保留原始视频的语义信息。

2) 本文模型能同时生成视频摘要与文本摘要。这对于目前互联网视频的短片预告与推荐具有很强的现实意义。

3) 在2个公开视频摘要数据集 SumMe 和 TVSum 上,相比现有先进方法,本文模型 F -score 指标分别提高了 0.5% 和 1.6%。

1 相关工作

1.1 视频摘要

视频摘要方法可以分为早期的传统方法和近期的深度学习方法。传统方法一般利用机器学习,提取底层视觉特征,并利用聚类方法提取视频摘要。深度学习方法则利用神经网络获得更高级的视觉特征以指导视频摘要生成。Zhao 等^[6]对 LSTM 网络进行了改进,利用分层结构的自适应 LSTM 网络提取视频摘要。Sharghi 等^[12]则在此基础上结合行列式点过程 (Determinantal Point Process, DPP) 优化了视频镜头子集的选择问题。针对递归神经网络占用资源过大、结构复杂不适合处理长视频等问题,Rochan 等^[13]使用全卷积模型代替递归神经网络,仅通过卷积网络对视频帧进行评价挑选也可以生成优质的视频摘要。

近年来,深度学习无监督视频摘要生成因其不需要依赖人工标注,引起了研究人员的关注。文献[14-16]中认为好的视频摘要能重建原始视

频,采用生成式对抗网络 (Generative Adversarial Networks, GAN) 指导视频摘要的生成,生成器用于生成视觉特征,判别器度量摘要视频和原始视频之间的相似度,以此优化网络。在文献[17-18]中,强化学习被应用于无监督视频摘要,通过设计不同的奖励函数实现无监督视频摘要。但这些方法没有考虑视频内容的语义信息,仅通过无监督方法无法生成具有丰富语义信息的视频摘要,很难满足人们对准确反映视频内容的视频摘要的需求。

1.2 视频文本描述

视频文本描述 (Video Captioning) 旨在为视频生成对应的文本描述,最常用的模型为编码器-解码器 (encoder-decoder)。该模型使用卷积神经网络 (Convolutional Neural Network, CNN) 或循环神经网络 (Recurrent Neural Network, RNN) 对视频进行编码获得视觉特征,然后使用语言模型对其进行解码获得文本描述。本文采用 encoder-decoder 模型,训练时最大化生成人工标注的概率指导包含语义信息的视频摘要生成,测试时使用该模块生成相应的文本摘要。

2 融合语义信息的视频摘要生成模型

本文提出了一种融合语义信息的视频摘要生成模型,如图1所示,该模型可分为3个模块:帧级分数加权模块、视觉-语义嵌入模块和视频文本描述生成模块。

在本节中,首先介绍视频摘要生成网络;其次将介绍网络的训练过程;最后展示了测试过程中如何同时生成视频摘要和文本摘要。

2.1 模型介绍

2.1.1 帧级分数加权模块

帧级分数加权模块如图2所示,在此模块中,首先对原始视频进行采样,每个视频获得 k 个关键帧。通过卷积神经网络对这些关键帧进行处理,提取其深层视频特征,记作 $\mathbf{f}_{\text{vis}} = \{\mathbf{f}_{\text{vis}}^1, \mathbf{f}_{\text{vis}}^2, \dots, \mathbf{f}_{\text{vis}}^k\}$, k 为采样的关键帧个数。帧级分数加权模块具有2个功能。一是获取帧级重要性分数,将得到的特征 $\mathbf{f}_{\text{vis}}$ 通过一个全连接层来生成帧级重要性分数 $s_i (i \in (1, 2, \dots, k))$, 在图2中表示为矩形条长度,分数条越长则帧级重要性分数越高。帧级重要性分数是用于描述视频帧在整个视频中代表性的分数,视频帧越能代表原始视频内容则分数越高,它也是之后在视频摘要生成过程中的重要依据。另一个功能则是将视频特征聚合为一个紧凑的特征向量。将特征 $\mathbf{f}_{\text{vis}}$ 与其对应的帧级

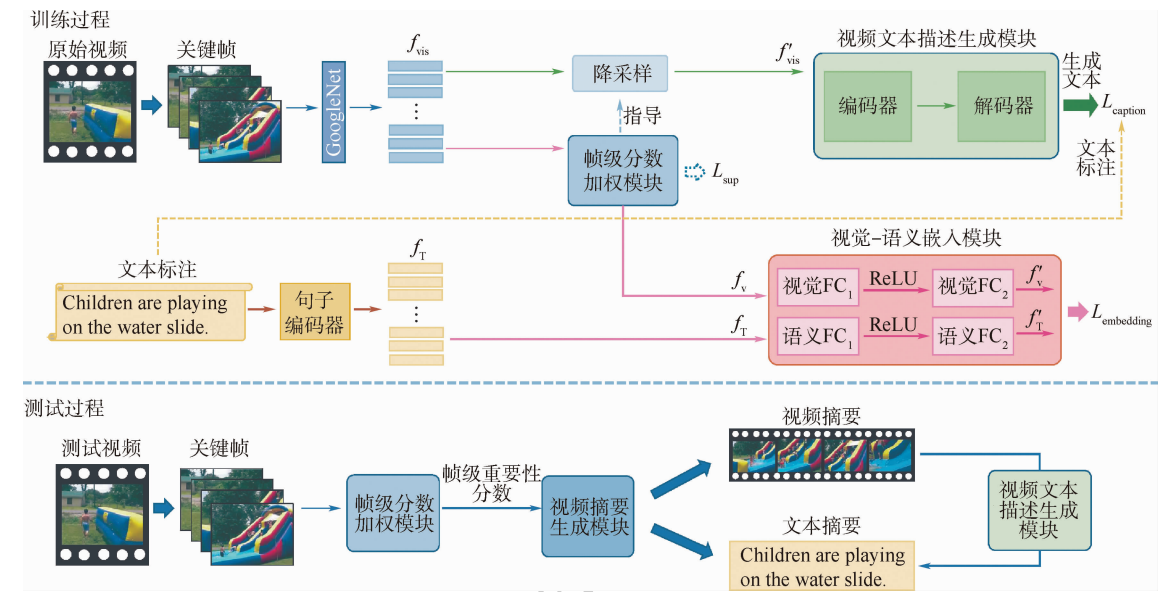


图1 融合语义信息的视频摘要生成流程

Fig. 1 Flowchart of video summarization by learning semantic information

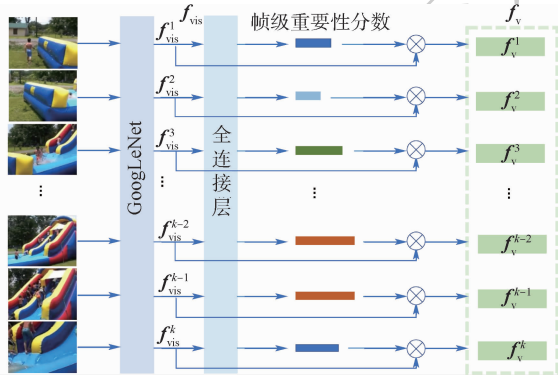


图2 帧级分数加权模块框架

Fig. 2 Framework of frame-level score weighting module

重要性分数 s_i 相乘后,得到加权向量 $f_v = \{f_v^1, f_v^2, \dots, f_v^k\}$, 定义为

$$f_v^i = s_i f_{vis}^i \quad (1)$$

2.1.2 视觉-语义嵌入模块

本文为数据集中每个视频都提供了 3 ~ 6 句简明的文本标注。使用词嵌入表示文本中出现的单词,之后利用 LSTM 网络处理词向量从而获得句子向量,该特征用 f_T 表示。获得视频的视觉特征和文本特征表示后,将视觉特征和文本特征投射到一个公共的嵌入空间中,通过使文本特征和视觉特征相互靠近,丰富视觉特征中的语义信息。该模块共有 2 个网络分支,分别负责对视觉特征与文本特征进行处理,由于视觉特征 f_v 与文本特征 f_T 维度并不相同,通过 2 个具有非线性层的 2 层全连接网络分别对 f_v 与 f_T 进行映射,使特征达到相同维度。将每个视频的视觉和文字特征表示映射到嵌入空间中,分别用 f'_v 和 f'_T 表示,以使用各种度量标准来衡量投影的视觉和文本特征之

间的相似性。本文中为了使 2 种特征彼此靠近,使用均方误差损失函数:

$$L_{embedding} = \frac{1}{k} \sum_{i=1}^k (f'_v - f'_T)^2 \quad (2)$$

2.1.3 视频文本描述生成模块

本文使用 encoder-decoder 模型生成视频文本描述生成模块,视频文本描述生成模块可以最直观地展示网络是否拥有选择具有最丰富语义信息的镜头以生成视频摘要的能力。从原始视频关键帧中提取出的视觉特征 f_{vis} ,将经过降采样操作后降维成 f'_{vis} 后送入视频文本描述生成模块,降采样操作是通过帧级分数加权模块得到每帧对应的分数后,从中挑选出分数最高的前 N 帧所对应的视觉特征,按时序顺序排列后组成 f'_{vis} 。

降采样操作必不可少,因为原始视频帧序列过长,且镜头变换丰富,包含较多的冗余信息,故将整个视频直接输入视频文本描述生成模块后编码器对视觉内容的编码能力会大大降低。降采样操作可以通过帧级重要性分数选择最具代表性的视觉特征,同时也可通过损失项的反馈调整帧级重要性分数,从而不断选择更好的关键帧用于文本生成,是十分必要的。

本文选择单层的 LSTM 网络作为编码器,并使用编码器对融合的特征进行编码以获得隐藏状态和单元状态,该状态用于初始化解码器。解码器同样是单层的 LSTM 网络,在通过编码器进行编码获得向量 z 后,解码器将其转换为单词序列 $S = \{w_1, w_2, \dots, w_n\}$, 每个 w 表示一个单词, n 为该句单词数量。视频文本描述模型生成句子的概率分布为 $P(S|z)$ 。为了使生成的文本句子逼近

给定的人工文本标注,本文将句子中单词的负对数似然之和作为损失函数 L_{caption} ,定义为

$$L_{\text{caption}} = -\log P(S|z) \quad (3)$$

2.2 网络训练

本文使用 Adam 优化器训练模型,通过使用 Adam 逐渐降低损失项,以逐步更新网络的参数。

2.2.1 预训练

由于 TVSum 和 SumMe 中包含的数据量很少,仅通过训练小样本数据无法达到文本生成的实验目的,因此利用现有的大型视频文本描述数据集对视频文本描述生成模块进行预训练。MSR-VTT 数据集^[19]包含了 10 000 个视频片段,每个视频都标注了 20 条英文句子,视频平均时长约为 10 s,视频内容采集自网络,包含各种类型的拍摄内容。实验证明,通过使用 MSR-VTT 对模型进行预训练确实可以有效缓解由样本不足引起的问题。

2.2.2 稀疏约束

为了充分利用 SumMe 和 TVSum 数据集提供的帧级重要性分数标注,本文还为网络添加了稀疏约束。稀疏约束的计算是利用在训练阶段关键帧计算获得的帧级重要性分数 s_i ,与数据集中提供的用户标注的帧级重要性分数 $\hat{s}_i$ 进行交叉熵损失计算:

$$L_{\text{sup}} = \frac{1}{k} \sum_{i=1}^k (\hat{s}_i \log s_i + (1 - \hat{s}_i) \log(1 - s_i)) \quad (4)$$

当利用稀疏约束时,训练网络的总损失函数为

$$L_{\text{total}} = \alpha L_{\text{embedding}} + \beta L_{\text{caption}} + \gamma L_{\text{sup}} \quad (5)$$

式中: α 、 β 、 γ 均为需要人工调节的超参数,当不使用稀疏约束时, γ 为 0。

2.3 视频摘要生成与文本摘要生成

测试时,模型不使用文本标注和句子编码器,只使用训练时得到的视频文本描述生成模块来生成视频摘要和文本摘要。

对于视频分割,本文使用 Potapov 等^[20]提出的 KTS 技术分割出不重叠的镜头。对于镜头重要性分数的计算,则是通过将镜头中包含的关键帧的帧级重要性分数求平均得到的。为了生成视频摘要,本文通过最大化镜头总分来选择镜头,同时确保视频摘要长度不超过视频长度的 15%,实际上,选择满足条件的镜头这一问题等价于 0/1 背包问题,也称为 NP 困难(Non-deterministic Polynomial Hard)问题。

在获得视频摘要后,本文将视频摘要中的帧级重要性分数排名前 N 的视觉特征,按时间顺序送入训练过程中获得的视频文本描述生成模块,来获得相应的文本摘要。

3 实验结果与分析

3.1 实验设置

3.1.1 数据集

本文在 SumMe^[21]和 TVSum^[22]数据集上评估模型。SumMe 由 25 个用户视频组成,涵盖了各种主题,例如假期和体育。SumMe 中的每个视频时长在 1 ~ 6 min 之间,并由 15 ~ 18 个人提供标注。TVSum 包含 50 个视频,其中包括新闻、纪录片等主题。每个视频时长从 2 ~ 10 min 不等。与 SumMe 类似,TVSum 中的每个视频都有 20 个标注,用于提供帧级重要性评分。

3.1.2 文本标注

上述 2 个数据集仅提供视频和帧级重要性分数,无法对视频长期语义进行建模。为此,为数据集 TVSum 和 SumMe 提供了文本标注,为 2 个数据集每个视频提供相互独立的 3 ~ 6 个简短的句子,以描述视频的主要内容。

3.1.3 评价指标

使用文献[23]中的方法对视频摘要进行评价,即通过度量模型生成的视频摘要与人工选择的视频摘要间的一致性来评估机器所生成摘要的性能。假设 A 为机器生成的摘要, B 为人工选择的摘要, D_A 为 A 的持续时间, D_B 为 B 的持续时间, D_{AB} 为 AB 重复部分的持续时间,则精度 P 和召回率 R 分别定义为

$$P = \frac{D_{AB}}{D_A} \quad (6)$$

$$R = \frac{D_{AB}}{D_B} \quad (7)$$

用于评估视频摘要的 F -score 定义为

$$F\text{-score} = \frac{2PR}{P + R} \times 100\% \quad (8)$$

文本摘要通过图像描述任务中常用的 3 个指标来进行评价,分别为 BLEU1、ROUGE-L 和 CIDEr。BLEU1 用于判断句子生成的准确性,ROUGE-L 用于计算句子生成的召回率,CIDEr 则体现的是生成句子与人工共识的匹配度,以上指标越高证明句子生成效果越好。

3.1.4 实验细节

使用在 ImageNet 上预训练的 GoogLeNet 作为获取视频特征的网络模型,每个视频以 2 fps(fps 为帧/s)的帧率对关键帧进行采样。每帧图像特征维度为 1 024,句子特征维度为 512,在视觉语义视频嵌入空间中将共同被映射为 256 维的向量。描述生成模型的编码器和解码器为单层 LSTM 网络,隐藏层维度设置为 512。在大型视频文本描述

数据集 MSR-VTT 上预训练视频文本描述生成模块,对其中每个视频都均匀采样 40 帧,所以框架中降采样参数 N 也设置为 40。在联合训练中,为了实现在预训练模型上的微调,学习率设置为 0.000 1。将 2 个数据集分别进行训练测试,其中每个数据集中 80% 的视频用于训练,其余用于测试。

3.1.5 对比方法

选择了 6 种最新的视频摘要模型与本文模型进行比较:vsLSTM^[4]、dppLSTM^[4]、SUM-GAN_{sup}^[5]、DR-DSN_{sup}^[17]、SASUM_{sup}^[11]、CSNet_{sup}^[24]。vsLSTM 是一种利用 LSTM 的视频摘要生成模型,也是较为基础的一种模型。dppLSTM 同样是一种利用 LSTM 的视频摘要生成模型,它通过使用 DPP 来选择内容多样化的关键帧。SUM-GAN_{sup} 是一种利用 VAE 和 GAN 的视频摘要生成模型。DR-DSN_{sup} 是一种无标签的强化学习视频摘要生成模型。SASUM_{sup} 是一个同样融合语义参与视频摘要模型。CSNet_{sup} 是另一种利用 VAE 和 GAN 的视频摘要生成模型。上述视频摘要模型均为有监督模型。

3.2 定量结果与分析

首先介绍本文方法的 2 种不同设置。

本文方法(无监督):本文提出的融合语义信息的视频摘要生成模型,没有任何稀疏性约束。

本文方法(有监督):由人工标注的视频摘要 L_{sup} 监督的融合语义信息的视频摘要生成模型。

如表 1 所示,与最先进的方法相比,本文模型具有更好的性能。本文无监督方法仅使用提供的带标注的文本描述,其结果几乎与其他带有人工标注的视频摘要的有监督方法相当。本文有监督方法在所有数据集上的表现优于所有引用的方法, F -score 指标较目前效果最好的 CSNet_{sup} 方法在 2 个数据集上分别提高了 0.5% 和 1.6%,这证明融合语义信息的视频摘要确实能够生成更高质量的视频摘要。

表 1 与 6 个最新方法之间的 F -score 比较

Table 1 Performance comparison (F -score) between our frameworks and six state-of-the-art methods

实验方法	F -score/%	
	SumMe	TVSum
vsLSTM ^[4]	37.6	54.2
dppLSTM ^[4]	38.6	54.7
SUM-GAN _{sup} ^[5]	41.7	56.3
DR-DSN _{sup} ^[17]	42.1	58.1
SASUM _{sup} ^[11]	45.3	58.2
CSNet _{sup} ^[24]	48.6	58.5
本文方法(无监督)	45.5	57.3
本文方法(有监督)	49.1	60.1

接下来分析本文方法 2 种不同设置产生的实验结果。本文有监督方法的性能优于本文无监督方法 2.8% ~ 3.6%。显然,这是因为本文有监督方法使用数据集提供的帧级重要性分数作为监督,因此相比于本文无监督方法生成的摘要,本文有监督方法生成的摘要与人工选择的视频摘要更具有一致性。

本文有监督方法在 SumMe 和 TVSum 上的表现分别为 49.1% 和 60.1%,之所以在 2 个数据集上的性能具有较大差异,是因为 SumMe 的视频内容变化缓慢且场景中的对象很少,而在 TVSum 中场景是多变的。丰富多样的镜头可能会在视频摘要生成过程中占据更大的优势,如果整个视频都是过于相似的镜头,则会对镜头挑选的边界更加模糊,从而导致生成的视频摘要评分较低。

表 2 展示了在本文有监督方法中,不同评估标准下在 SumMe 和 TVSum 数据集上生成视频摘要的性能。可以看出文本摘要可以一定程度上描述视频内容,之后也会在 3.4.2 节中定性展示文本生成的效果。不过由于视频摘要任务数据集数据量过小,文本摘要性能受到了很大影响,容易生成结构较为单一的短句。

表 2 不同数据集生成的文本摘要评测

Table 2 Evaluation of text summaries generated by different datasets

数据集	BLEU-1/%	ROUGE-L/%	CIDEr/%
SumMe	28.3	27.6	9.6
TVSum	32.8	29.9	12.7

3.3 消融实验

为了验证视频摘要模型中的视觉-语义嵌入模块和视频文本描述生成模块,本文在 TVSum 数据集上进行消融实验,通过分别从模型中排除不同模块来展示每个模块对网络性能的贡献。此外,如果不使用这 2 个模块中的任何一个,则该问题将归结为最基本的视频汇总问题,即损失函数只能使用 L_{sup} 。为了保持实验参数的一致性,本文将采用有监督方法来进行消融实验,以确保 L_{sup} 可用。从表 3 数据中可以看出,应用不同模块时 F -score 分数将逐渐增加,当 2 个模块同时应用时达到实验的最优结果。这证明本文提出的 2 个模

表 3 TVSum 数据集上的消融实验结果

Table 3 Results of ablation experiment on TVSum

实验编号	嵌入空间	描述生成	F -score/%
1	×	×	50.4
2	✓	×	51.5
3	×	✓	58.3
4	✓	✓	60.1

块确实能够指导具有语义信息的关键帧获得更高的帧级重要性分数,从而生成更优质的视频摘要。

3.4 定性结果与分析

3.4.1 视频摘要示例

选择用 TVSum 数据集中一则“宠物美容”相关新闻视频来展示视频摘要的生成情况。图 3(a)展示了原始视频内容,可以看到原始视频除宠物美容相关镜头外,还包含新闻片头片尾、采访路人等画面。图 3 还分别展示了本文无监督方法、本文有监督方法、本文有监督方法但不含视频文本描述生成模块、本文有监督方法但不含视觉-语义嵌入模块 4 种设置下,从原始视频中挑选的视频摘要情况。图中浅色条形图表示人工标注的帧级重要性分数,深色条形图表示该设置下挑选的视频摘要。

从图 3(b)~(e)可以看出,每种设置都能够很好地生成针对整个视频的视频摘要,挑选的视频摘要覆盖了人工标注的峰值,能够反映出原始视频是一则与宠物相关的视频,这表明本文模型

可以从视频中提取出语义丰富的视频镜头,生成优质的视频摘要。

同时对比图 3(b)与图 3(c),发现本文有监督方法相较于无监督方法表现更好。由于稀疏约束的加入,相比于无监督方法只注重于挑选具有代表性的视频,本文有监督方法提取的视频镜头覆盖范围更广、彼此间隔更分明,生成的视频摘要富含语义信息的同时还具有镜头多样性。

在图 3(c)~(e)几种有监督设置中,利用全部模块的设置表现最佳。在缺失部分模块的情况下,虽然摘要的代表性与多样性依然较好,但还是容易挑选与语义信息不相关的镜头。如图 3(d)与图 3(e)中,2 种设置下的模型都将新闻的片头画面挑选为视频摘要,而应用全部模块的有监督方法挑选的摘要几乎不含这些无关镜头,这也正与 3.3 节中的消融实验结果相对应。

关于视频摘要的定性结果展示视频可前往 <https://github.com/huarui1996/vsc> 进行观看,后续更多相关数据也会在此开源。

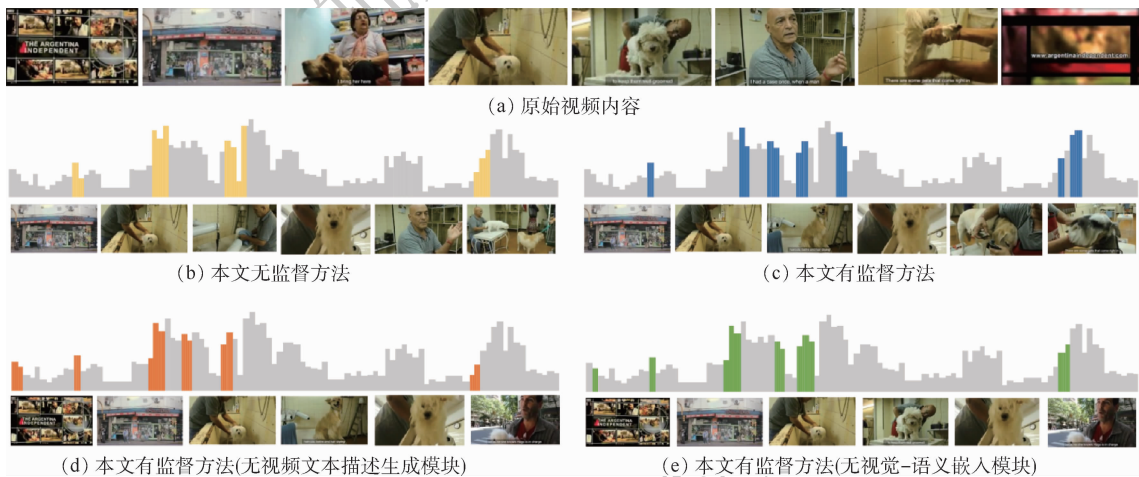


图 3 TVSum 数据集中生成视频摘要的示例

Fig. 3 Examples of video summarization in TVSum

3.4.2 视频文本摘要对应

图 4 展示了在 TVSum 数据集中,对名为 Yi4Ij-

2NM7U4 的视频同时生成的文本摘要和视频摘要的内容对应,同时还展示了 3 个不同用户对原始视频添加的文本标注。可以看出,虽然本文模型生成句子的结构较为简单,但是可以描述出视频主要内容。

4 结 论

1) 本文提出了一种融合语义信息的视频摘要生成模型,该模型通过帧级分数加权模块、视觉-语义嵌入模块和视频文本描述生成模块 3 个模块,使得视觉特征包含丰富的语义信息,可以同时生成具有语义信息的视频摘要以及相应的文本摘要。

2) 实验结果表明,由本文模型生成视频摘要能够展示出原始视频中重要的语义内容,同时文



用户1:a man is cooking
用户2:a man cooking in the kitchen
用户3:a man is cooking vegetables and hot dog
文本摘要:a person is doing a cooking show

图 4 TVSum 数据集中生成文本摘要的示例

Fig. 4 Examples of text summarization in TVSum

本摘要可以简明地对视频内容进行描述。

在未来的工作中,将利用多模态信息实现视频摘要生成任务,并根据不同用户的需求生成不同主题的视频摘要。

参考文献 (References)

[1] 刘波. 视频摘要研究综述[J]. 南京信息工程大学学报, 2020, 12(3): 274-278.
LIU B. Survey of video summary[J]. Journal of Nanjing University of Information Science & Technology, 2020, 12(3): 274-278 (in Chinese).

[2] RAV-ACHA A, PRITCH Y, PELEG S. Making a long video short: Dynamic video synopsis[C] // Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2006: 435-441.

[3] ZHAO B, XING E P. Quasi real-time summarization for consumer videos[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 2513-2520.

[4] ZHANG K, CHAO W, SHA F, et al. Video summarization with long short-term memory[C] // European Conference on Computer Vision. Berlin: Springer, 2016: 766-782.

[5] MAHASSENI B, LAM M, TODOROVIC S, et al. Unsupervised video summarization with adversarial LSTM networks[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017: 2982-2991.

[6] ZHAO B, LI X, LU X, et al. HSA-RNN: Hierarchical structure-adaptive RNN for video summarization[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 7405-7414.

[7] 冀中, 江俊杰. 基于解码器注意力机制的视频摘要[J]. 天津大学学报, 2018, 51(10): 1023-1030.
JI Z, JIANG J J. Video summarization based on decoder attention mechanism[J]. Transactions of Tianjin University, 2018, 51(10): 1023-1030 (in Chinese).

[8] 李依依, 王继龙. 自注意力机制的视频摘要模型[J]. 计算机辅助设计与图形学学报, 2020, 32(4): 652-659.
LI Y Y, WANG J L. Self-attention based video summarization[J]. Journal of Computer-Aided Design & Computer Graphics, 2020, 32(4): 652-659 (in Chinese).

[9] CHEN Y, TAO L, WANG X, et al. Weakly supervised video summarization by hierarchical reinforcement learning[C] // Proceedings of the ACM Multimedia Asia. New York: ACM Press, 2019: 1-6.

[10] CHOI J, OH T, KWEON I S, et al. Contextually customized video summaries via natural language[C] // Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2018: 1718-1726.

[11] WEI H, NI B, YAN Y, et al. Video summarization via semantic attended networks[C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 216-223.

[12] SHARGHI A, BORJI A, LI C, et al. Improving sequential determinantal point processes for supervised video summarization

[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 533-550.

[13] ROCHAN M, YE L, WANG Y, et al. Video summarization using fully convolutional sequence networks[C] // European Conference on Computer Vision. Berlin: Springer, 2018: 358-374.

[14] ZHANG Y, KAMPFFMEYER M, LIANG X, et al. Query-conditioned three-player adversarial network for video summarization[C] // British Machine Vision Conference. Berlin: Springer, 2018.

[15] YUAN L, TAY F E, LI P, et al. Cycle-sum: Cycle-consistent adversarial LSTM networks for unsupervised video summarization[C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 9143-9150.

[16] ZHANG Y, KAMPFFMEYER M, ZHAO X, et al. DTR-GAN: Dilated temporal relational adversarial network for video summarization[C] // Proceedings of the ACM Turing Celebration Conference-China. New York: ACM Press, 2019: 1-6.

[17] ZHOU K, QIAO Y, XIANG T. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward[C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2018: 7582-7589.

[18] WANG L, ZHU Y, PAN H. Unsupervised reinforcement learning for video summarization reward function[C] // Proceedings of the 2019 International Conference on Image, Video and Signal Processing. New York: ACM Press, 2019: 40-44.

[19] XU J, MEI T, YAO T, et al. Msr-VTT: A large video description dataset for bridging video and language[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 5288-5296.

[20] POTAPOV D, DOUZE M, HARCHAOU I Z, et al. Category-specific video summarization[C] // European Conference on Computer Vision. Berlin: Springer, 2014: 540-555.

[21] GONG B, CHAO W L, GRAUMAN K, et al. Diverse sequential subset selection for supervised video summarization[C] // Advances in Neural Information Processing Systems. New York: Curran Associates, 2014: 2069-2077.

[22] SONG Y, VALLMITJANA J, STENT A, et al. TVSum: Summarizing web videos using titles[C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015: 5179-5187.

[23] GYGLI M, GRABNER H, RIEMENSCHNEIDER H, et al. Creating summaries from user videos[C] // European Conference on Computer Vision. Berlin: Springer, 2014: 505-520.

[24] JUNG Y, CHO D, KIM D, et al. Discriminative feature learning for unsupervised video summarization[C] // Proceedings of the AAAI Conference on Artificial Intelligence, 2019: 8537-8544.

作者简介:
滑蕊 女, 硕士研究生。主要研究方向: 视频摘要。

吴心筱 女, 博士, 副教授, 博士生导师。主要研究方向: 视觉与语言、视频内容理解、机器学习。

赵文天 男, 博士研究生。主要研究方向: 图像和视频描述生成。

Video summarization by learning semantic information

HUA Rui, WU Xinxiao*, ZHAO Wentian

(School of Computer Science & Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Video summarization aims to generate short and compact summary to represent original video. However, the existing methods focus more on representativeness and diversity of representation, but less on semantic information. In order to fully exploit semantic information of video content, we propose a novel video summarization model that learns a visual-semantic embedding space, so that the video features contain rich semantic information. It can generate video summaries and text summaries that describe the original video simultaneously. The model is mainly divided into three modules: frame-level score weighting module that combines convolutional layers and fully connected layers; visual-semantic embedding module that embeds the video and text in a common embedding space and make them lose to each other to achieve the purpose of mutual promotion of two features; video caption generation module that generates video summary with semantic information by minimizing the distance between the generated description of the video summary and the manually annotated text of the original video. During the test, while obtaining the video summary, we obtain a short text summary as a by-product, which can help people understand the video content more intuitively. Experiments on SumMe and TVSum datasets show that the proposed model achieves better performance than the existing advanced methods by fusing semantic information, and improves *F*-score by 0.5% and 1.6%, respectively.

Keywords: video summarization; visual-semantic embedding space; video captioning; video key frame; Long Short-Term Memory (LSTM) model

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0449

基于无监督学习的多模态可变形配准



马腾宇, 李孜, 刘日升*, 樊鑫, 罗钟铉

(大连理工大学 国际信息与软件学院, 大连 116000)

摘 要: 针对医学图像配准问题,传统方法提出通过解决优化问题进行配准,但计算成本高、运行时间长。深度学习方法提出使用网络学习配准参数,从而进行配准并在单模态图像上取得高效性能。但在多模态图像配准时,不同模态图像的强度分布未知且复杂,大多已有方法严重依赖标签数据,现有方法不能完全解决此问题。提出一种基于无监督学习的深度多模态可变形图像配准框架。该框架由基于损失映射量的特征学习和基于最大后验概率的变形场学习组成,借助空间转换函数和可微分的互信息损失函数实现无监督训练。在 MRI T1、MRI T2 以及 CT 的 3D 多模态图像配准任务上,将所提方法与现有先进的多模态配准方法进行比较。此外,还在最新的 COVID-19 的 CT 数据上展示了所提方法的配准性能。大量结果表明:所提方法与其他方法相比,在配准精度上具有竞争优势,并且大大减少了计算时间。

关 键 词: 深度学习; 无监督; 医学图像配准; 多模态; 计算机视觉

中图分类号: TP183

文献标志码: A **文章编号:** 1001-5965(2021)03-0658-07

医学图像配准^[1]旨在寻求两幅医学图像之间的空间变换,从而对齐 2 幅图像中处于相同解剖位置的点,其在许多医学成像任务中至关重要。图像配准可以将同一对象在不同时间、不同状态、不同的拍摄视角甚至是来自不同成像设备的图片进行比较与融合。其中,多模态配准用于对齐不同模态的图像,涉及多种医学图像模式,如 MRI、CT 和 PET 等。这些图像在不同的空间分辨率下均显示出独特的组织特征,例如,MRI T1 图像能够很好地区分大脑中不同的健康组织,而 MRI T2 图像则适合于突出异常结构。在临床实践中,通常需要融合多种图像模式以用于诊断或介入目的,多模态配准显得尤为重要。

传统的配准方法^[2-5]通过定义相似性度量函数和正则化约束,将配准问题转换为迭代优化问题,通过不断地优化使配准后的图像与固定图像之间的相似性最大,从而得到最优配准图像。然

而,基于优化的传统方法通常需要在高维空间上进行迭代的梯度运算,导致运算成本昂贵。医学影像配准在应用中由于缺乏标签数据在训练与评估中面临较大问题,与此同时由于不同模态图像强度分布未知且复杂导致其极具挑战。

最近,基于深度学习来进行可变形图像配准的方法得到广泛发展,无监督学习^[6]、弱监督学习^[7]发展迅速,此外通过学习图像内在特征^[8-11]学习配准参数也已经取得高效性能,这些方法采用卷积神经网络求解用于配准的网络模型,直接估计输入图像对之间的目标位移场。与传统的配准方法相比,基于深度学习的图像配准技术极大缩短了运行时间。这些方法在单模态配准问题上实现了快速配准和可比较的配准精度,如 Voxel-Morph 方法^[6]作为无监督单模态图像配准框架,在脑部数据集上取得了优越的性能。国内外学者基于深度学习方法对多模态配准问题进行了大量

收稿日期: 2020-08-24; 录用日期: 2020-09-11; 网络出版时间: 2020-10-30 11:53
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201030.1134.003.html
基金项目: 国家自然科学基金(61922019,61672125,61733002,61772105); 兴辽英才计划(XLYC1807088)
* 通信作者. E-mail: rslu@dlut.edu.cn

引用格式: 马腾宇, 李孜, 刘日升, 等. 基于无监督学习的多模态可变形配准[J]. 北京航空航天大学学报, 2021, 47(3): 658-664.
MA T Y, LI Z, LIU R S, et al. Multimodal deformable registration based on unsupervised learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 658-664 (in Chinese).

研究,但由于多模态配准局限性,该领域仍有较大的发展潜力。对多模态图像配准来说,由于不同模态图像的强度分布未知且复杂,大多已有多模态图像配准方法严重依赖标签数据,例如真实的变形场或解剖标志,基于弱监督学习实现多模态配准的研究工作^[7],通过从解剖标签中包含的更高级别的对应信息中推断体素级别的转换,但获取真实解剖标签仍有较大困难。相关研究^[8]探索了无监督的深度学习方法,基于图像转换方法将图像嵌入到低维的潜在空间中并在其中学习变形场,该方法在 2D 的多模态配准问题上取得了令人满意的性能。此外,部分研究^[9]通过潜在空间特征编码的无监督域自适应技术打破不同模态之间的配准障碍,但是此方法在多模态实验上表现仍不够优异。

为应对以上挑战,本文提出一种完全无监督学习的深度多模态可变形图像配准方法,该方法在训练过程中不需要任何真实的变形场或对齐的多模态图像对。首先通过对不同模态的图像进行基于损失映射量的特征学习,然后基于最大后验概率对变形场进行估计。值得一提的是,本文借助空间转换器^[12]和可微分的互信息损失函数最大化图像间的相似性,同时在配准字段上施加平滑约束,从而实现无监督学习。和以往方法相比,本文的主要贡献如下:

- 1) 提出了一种基于无监督学习的深度多模态可变形图像配准框架,该框架不需要任何对齐的图像对或解剖学标记数据。
- 2) 所提出的框架针对多模态配准问题首先分别学习不同模态下的图像信息,提取不同分辨率下的图像特征,并通过构建更具辨别力的特征损失映射量,然后对其进行处理,利用最大后验概率求解变形场。
- 3) 本文方法具有普适性,所提出的框架和损失函数可以转移到各种不同模态的医学图像应用中。

在脑部 MRI T1、MRI T2 以及 CT-MRI 上的 3D 多模态图像配准任务上,将所提出的方法与现有先进的多模态配准方法进行了比较。此外,还在最新的 COVID-19 的 3D 肺部 CT 的单模态配准任务上展示了本文方法的配准性能。大量的实验结果证明,相比于现有最好的配准方法,本文方法能够在更短的运行时间内得到具有竞争力的配准精度,并且在多模态与单模态医学图像配准中表现同样优异。

1 相关知识

不同图像配准允许图像变换具有不同的自由度,刚性变换和仿射变换自由度低,通常用作全局对齐的初始变换。可变形图像配准是一种非线性配准过程,旨在图像之间建立密集的非线性空间对应关系,允许更高的转换自由度。本文关注可变形的多模态配准问题。令来自不同模态的移动图像 M 与固定图像 F 作为输入, φ 作为配准字段,也称作变形场。典型的可变形配准问题可以表示为

$$\hat{\varphi} = \arg \min_{\varphi} L_{\text{sim}}(F, M(\varphi)) + L_{\text{reg}}(\varphi) \tag{1}$$

式中: $\hat{\varphi}$ 为最优变形场; $M(\varphi)$ 为通过移动图像 M 与变形场进行空间转换后得到的图像; L_{sim} 为相似性度量函数; L_{reg} 为正则化函数,约束产生空间上平滑的变形场。

在大多数可变形图像配准设置中,仿射变换已考虑在内,图像之间的不匹配来源仅包括非线性的。在本文中,遵循这一假设,所有实验数据均已经过仿射变换。

2 基于深度学习的医学图像配准

2.1 方法流程

本文提出了一种完全无监督的深度多模态可变形图像配准方法,整体流程如图 1 所示。首先,在特征提取的过程中分别对 2 种模态图像进行多尺度提取特征;其次,进行损失映射量计算,通过概率变形场生成模块对损失映射量的处理,可以由粗到细地得到变形场;最后,通过空间转换网络得到最终的配准图像,并使用评估方法对配准图像进行评估,利用损失函数完成无监督训练。

2.2 网络设计

2.2.1 特征提取

本文提出的特征提取方法是基于金字塔模型进行的多尺度特征提取,借鉴了文献[13]中所提的设计思想。值得一提的是,为了对不同模态的图像提取到更多信息,本文利用不同的网络学习不同模态下的特征。本文中多模态实验在 CT 和 MRI 图像上进行。通过特征提取模块可以形成多层不同分辨率下的特征图像组成的特征金字塔,其中金字塔底部是输入图像,使用卷积进行下采样,在每次卷积之后进行激活,最后分别输出由小到大的金字塔形特征图像。在本文中共使用 2 个尺度的特征图像。通过提取特征金字塔的方式,可以得到关于原图像更多的特征信息,并且对

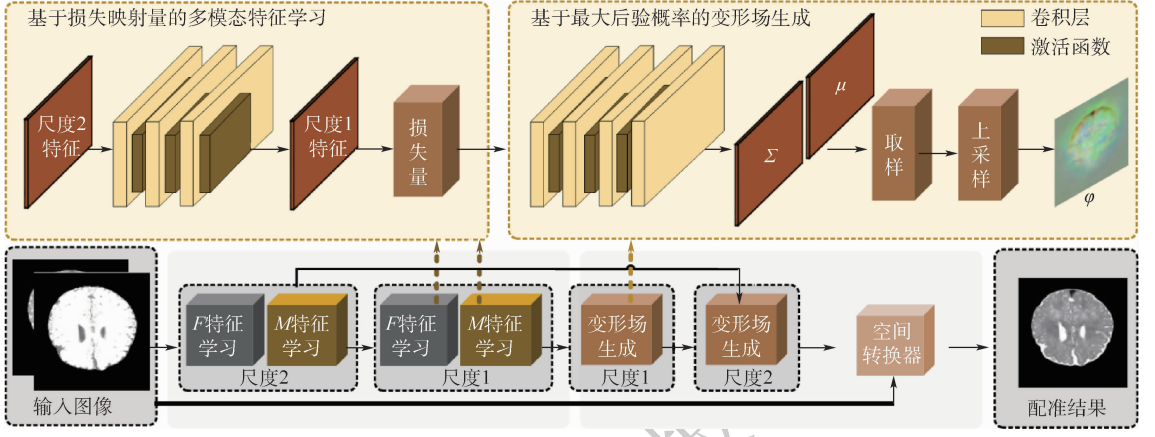


图1 由基于匹配量的特征学习和基于最大后验概率的变形场学习组成的无监督多模态配准框架

Fig.1 Unsupervised multimodal registration framework composed of feature learning based on matching amount and deformation field learning based on maximum posterior probability

不同低分辨率的图像操作使得整个网络更轻便快速。针对不同的模态图像使用不同特征提取器,这也是本文方法在实现多模态配准中的显著优点。

2.2.2 损失映射量计算

由于不同模态图像包含信息以及在成像方式上的巨大差别,因此本文在特征学习过程中引入更具辨识力的损失映射量,用于衡量不同模态特征的不匹配度,并在变形场生成中发挥作用,该模块借鉴光流领域中 TV-L1 能量模型^[14]。具体来说,损失映射函数为 L1 范数,通过计算移动图像与固定图像在特征金字塔中相同层次的特征图像的 L1 范数来表示特征匹配中的误差,该模块定义如下:

$$L1_cost = \|fixed_i - moving_i\|_1 \quad (2)$$

式中: i 为处在特征金字塔的层次;fixed 为固定图像特征;moving 为移动图像特征。式(2)即定义为 fixed 图像特征和 moving 图像特征之间的数据残差项得到对应层次特征关联之间的匹配成本。

2.2.3 基于概率的变形场生成

变形场生成模块对提取到的多个层次的特征损失映射量进行处理,得到不同分辨率的变形场。由于不同模态图像保存的空间位置信息以及保存图像方式的不同,在配准过程中预测变形场有更大的挑战,本文提出将最大后验概率应用到多模态图像配准中以获得最佳变形场。

变形场生成网络构成由卷积、激活函数交替组成,可对多模态图像特征进行融合处理。随后对于不同层次的特征进行转换,通过求解最大后验概率得到所求当前尺度的变形场。和以往直接求得变形场相比,本文提出的变形场估计网络求解由变形场平均值和方差组成的近似后验概率参数,经过采样后输出当前层次的变形场。在预测得到变形场的过程中,设 z 为所求的变形场,对变

形场概率分布进行建模:

$$p(z) = N(z;0;\Sigma_z) \quad (3)$$

通过让 $\Sigma_z^{-1} = \Lambda_z = aL$ 来鼓励 z 的空间平滑度。 Λ_z 为精度矩阵;而 a 为控制变形场 z 的比例参数; L 为定义在体素网格上的邻域图的拉普拉斯算子。对于待配准的图像 (x,y) ,将 x 作为扭曲图像 y 的嘈杂观测值, $\mathbf{I}$ 为单位矩阵; σ^2 反映附加噪声图像的方差,表示如下:

$$p(x|z;y) = N(x;y \circ \varphi_z, \sigma^2 \mathbf{I}) \quad (4)$$

所提出的方法旨在最大化后验概率 $p(z|x;y)$ 。具体来说,通过最大后验概率估计获得当前图像对 (x,y) 的最佳变形场 φ_z 。由于直接求后验概率 $p(z|x,y)$ 并不容易,使用变分方法,并引入由 ψ 参数化的近似后验概率 $q_\psi(z|x;y)$,在不断优化的过程中,将散度损失 KL 差异最小化:

$$\min_{\psi} KL[q_\psi(z|x;y) \| p(z|x;y)] \quad (5)$$

本文将近似后验概率 $q_\psi(z|x;y)$ 建模为多元正态分布,通过 ψ 参数化神经网络,输出 $\mu_{z|x,y}$ 、 $\Sigma_{z|x,y}$,并通过优化变分下界来学习参数 ψ 。

2.3 损失函数

损失函数衡量对模型输出的预测值与真实值之间相似性匹配的程度,以及变形场的光滑程度。通过不断优化损失函数从而指导模型的学习过程。

互信息是信息论中的重要概念,描述了 2 个系统之间的相关性,或互相包含信息的多少。在 2 幅图像中,通过熵以及联合熵的概念反映图像间信息相互包含的程度。互信息越大则表明 2 幅图像间的相关性越大。对于移动图像 M 、固定图像 F 来说,其互信息表示为

$$MI(M,F) = \sum_{f \in F} \sum_{m \in M} p(m,f) \left(\log_2 \frac{p(m,f)}{p(m)p(f)} \right) \quad (6)$$

式中: $p(m, f)$ 为 M 和 F 的联合概率分布函数; $p(m)p(f)$ 为 M 和 F 的边缘概率分布函数。

当 2 幅图像相似度越高或重合部分越大时,其相关性越大,互信息越大。对于多模态医学图像来说,不同模态图像在成像过程中存在显著差异,因此使用互信息评估两幅不同模态图像配准程度是有效且恰当的方式。通过互信息进行描述提取特征的匹配性^[15-16]也是当前研究下的热点。

在本文的设计中,通过最大后验概率预测得到变形场,在预测得到变形场的过程中设 z 为所求的变形场,在此过程中涉及到 KL 散度损失,如式(5)所示,对其进行化简,将此部分损失定义如下:

$$\text{Loss}_{\text{KL}} = -E_q[\log_2 p(x|z;y)] + \text{KL}[q_\psi(z|x;y) \| p(z)] \tag{7}$$

式中: E_q 表示期望。最小化 M 和 F 图像之间的差异将鼓励 $M(\varphi)$ 近似 F ,但可能导致不连续。所以在其空间梯度上使用扩散正则化器保证平滑:

$$L_{\text{smooth}}(\varphi) = \sum_{p \in \Omega} \|\nabla \varphi(p)\|^2 \tag{8}$$

式中: Ω 代表几何空间; $\nabla \varphi(p)$ 为在点 p 处求梯度。

对于多模态图像配准,本文使用互信息损失、正则化项、KL 散度设计损失函数,多模态配准使用的损失函数如下所示:

$$L(F, M, \varphi) = -\text{MI}(F, M(\varphi)) + \text{Loss}_{\text{KL}} + \lambda \sum_{p \in \Omega} \|\nabla \varphi(p)\|^2 \tag{9}$$

式中: λ 为正则化参数。

3 实验结果分析

本文使用 Tensorflow 框架设计整体网络。在多模态图像配准实验中,经历 2 次特征提取后,特征图像分辨率分别变为原图像的 1/2 和 1/4。使用 Adam 优化算法,并将学习率设为 0.000 1。由于实验环境内存大小的限制,在本文的所有实验中仅将批处理大小设置为 1。

本文多模态医学图像配准实验在脑部 MRI T1、MRI T2 扫描数据、头部 CT 扫描数据。将所有配准实验中的数据分为 3 个部分,即训练数据、验证数据、测试数据,比例为 0.75:0.05:0.2。对于头部 MRI 数据,首先将其进行预处理,进行仿射空间归一化和脑部提取,将所有的处理后数据大小设置为(160,160,160)。

此外,还在最新的肺部 CT 扫描数据进行实验,将 COVID-19 数据作为模板图像,验证本文方法在单模态图像配准中的可行性。对于肺部 CT

数据,将切片数据转换为三维数据,并将图像大小设置为(512,512,16)。

3.1 数据集

本文使用的实验数据包括以下几个数据集:ADNI^[17]:用于阿尔茨海默病早期诊断的方法,由阿尔茨海默病神经影像学倡议。

NSCLC^[18-19]:美国癌症影像档案馆 NSCLC-Radiomics 馆藏的 402 幅 CT 扫描划定的左右胸部体积分割,胸部分割中包含肺部实质等数据。

CQ500^[20]:收集了包含头部 CT 扫描以及来自各个中心临床报告的数据集。

COVID-19^[21]:来自于 COVID-19-CT-Seg-Benchmark,为新型冠状病毒肺炎分割数据。

3.2 评估指标

3.2.1 Dice

Dice 评分是一种集合相似度度量函数,通常用于计算 2 个样本的相似度。对于医学图像配准来说,如果移动图像通过和变形场拟合得到配准图像,那么配准图像与固定图像之间的医学解剖结构应该可以最大限度的重合,并可量化 2 幅图像之间的体积重叠,因此通过计算配准后图像与固定图像之间的 Dice 值来进行评估最终的配准效果:

$$\text{Dice}(F, M(\varphi)) = 2 \frac{|F \cap M(\varphi)|}{|F| + |M(\varphi)|} \tag{10}$$

Dice 得分为 1 时表示结构完全重叠,Dice 得分为 0 表示无重叠,在评估中 Dice 得分越大越好。

3.2.2 平均幅度

为了量化变形场的平滑性,定性地显示了变形场的雅可比行列式,同时计算了雅可比行列式梯度的平均幅度(Grad Det-Jac),即通过雅可比行列式梯度的平均幅度来定量表示变形场的平滑性,幅度越小,表示变形场变形越平滑。

3.2.3 均方误差

均方误差(Root Mean Square Error, RMSE)是一种通过对 2 幅图像中每个像素点对应关系的度量来计算图像相似度的方法。RMSE 越小说明模型选择和拟合更好,数据预测也越成功。RMSE 定义如下:

$$\text{RMSE}(F, M) = \sqrt{\frac{1}{|\Omega|} \sum_{p \in \Omega} (F(p) - M(p))^2} \tag{11}$$

式中: $F(p)$ 和 $M(p)$ 为固定图像与移动图像在 p 点处的值。

3.3 比较方法

医学影像研究界已经开发了基于深度学习的方法,并在包括图像配准在内的许多应用中实现

了最先进的技术。选取性能表现优异的深度学习方法 VoxelMorph^[6] 作为比较模型。同时,选取 2 个 ANTs^[22] 传统方法中的对称归一化方法 SyN^[23]、Elastix^[24] 作为本文的对比方法。

3.4 实验结果

本文共进行 3 个方面的实验,接下来将详细介绍每种实验配置以及结果:

- 1) 同一主体 MRI T1 和 MRI T2 的多模态配准。
- 2) 将 CT 脑部图像作为 atlas 模板,使用任意 3D MRI T2 图像匹配到模板图像。
- 3) 将 COVID-19 图像作为模板,使用任意 CT 肺部图像匹配到模板图像。

3.4.1 MRI T1-MRI T2 配准

在磁共振成像中,MRI T1 观察解剖结构较好。很多病灶的 MRI T2 信号要强于周围的正常组织,故 MRI T2 显示组织病变较好。将 MRI T1 与 MRI T2 图像相配准可以很好地在较好的解剖结构中观察病变组织。将同一主体的 MRI T1、MRI T2 图像分别设置为固定图像与移动图像,表 1 展示了分别将 MRI T1、MRI T2 作为固定图像时的 RMSE 配准结果,affined 为不经过配准,直接使用评价指标度量的结果(粗体代表最优结果),图 2 为可视化效果。

表 1 MRI T1-MRI T2 RMSE 配准结果

Table 1 MRI T1-MRI T2 RMSE registration results

方法	配准方向	
	MRI T2 - > MRI T1	MRI T1 - > MRI T2
affined	17.549(3.116)	17.549(3.116)
VoxelMorph	16.844(3.276)	17.545(2.982)
Elastix	17.731(3.140)	17.717(3.111)
ANTs (SyN)	16.932(3.141)	17.206(3.078)
本文	16.725(3.303)	16.915(3.223)

3.4.2 CT-MRI 配准

在多模态 CT-MRI 配准实验中,选择 CT 头部数据作为 atlas 图像,MRI 图像作为移动图像。由于没有医学图像分割标签,使用 RMSE 作为评价

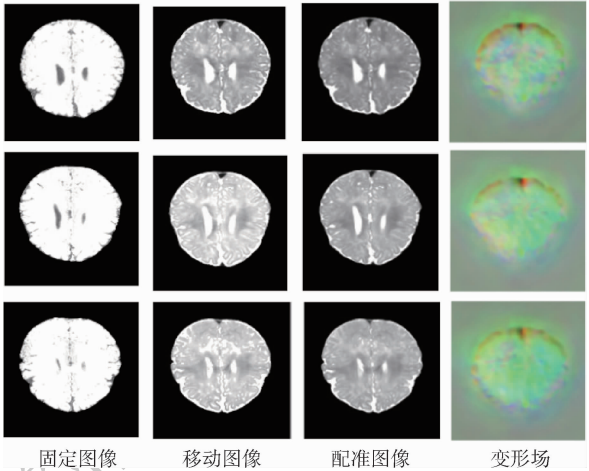


图 2 MRI T2 -> MRI T1 配准结果可视化

Fig. 2 Visualization of MRI T2 -> MRI T1 registration results

指标。此实验在测试时,使 MRI T2 数据向模板图像进行配准。

表 2 显示了将 CT 脑部数据作为模板图像进行实验的结果(粗体代表最优结果),包括均值和标准差,进一步证明本文提出方法在多模态实验表现优异。

表 2 CT-MRI 配准结果

Table 2 CT-MRI registration results

方法	RMSE
affined	29.932(0.819)
VoxelMorph	28.196(4.374)
Elastix	27.252(0.961)
ANTs (SyN)	29.401(1.021)
本文	25.886(0.374)

3.4.3 CT-CT 配准

为验证本文框架的普适性,在 COVID-19 数据上进行了单模态配准的实验。为了便于观察是否感染,将 COVID-19 作为 atlas,未患病或疑似病人图像配准到 atlas 以便帮助医生确认是否感染病毒。训练阶段使用移动图像向 atlas 进行配准,表 3 展示了 CT 配准结果,NCC 代表图像的互相关程度(粗体代表最优结果)。

表 3 CT-CT 配准结果

Table 3 CT-CT registration results

方法	Dice	RMSE	NCC	Grad Det-Jac/ 10^{-3}
affined	0.464(0.091)	32.714(1.694)	0.180(0.008)	
VoxelMorph	0.636(0.109)	21.086(2.313)	0.224(0.016)	4.497(1.058)
Elastix	0.522(0.123)	33.874(2.191)	0.112(0.021)	
ANTs (SyN)	0.611(0.233)	31.382(4.169)	0.133(0.063)	
本文-L1_cost	0.654(0.077)	30.135(3.336)	0.285(0.014)	0.866(0.004)
本文-Prob	0.678(0.094)	28.617(3.232)	0.283(0.012)	0.812(0.005)
本文	0.708(0.086)	27.461(3.162)	0.287(0.014)	0.812(0.004)

实验结果表明,本文的框架在多个评价指标上取得领先性能。尤其是在 Dice 指标上,本文方法取得十分优异的性能。原因可能是因为在胸部扫描数据上,部分像素点在变形场中需要有较大的空间位置变换,本文方法利用多个分辨率图像,可以保留更多的空间信息,从而优于比较方法。其中本文-Prob 方法直接进行变形场预测而不经概率进行变形场学习,本文-L1_cost 方法为去掉损失映射量模块的方法,通过与 2 组消融实验比较验证本文所提模块的有效性。

表 4 为 MRI T2 配准到 MRI T1 图像的配准时间比较(粗体代表最优结果)。结果显示,本文方法在达到较高精度的同时比传统方法的时间更低,这是由于本文提出在多个低分辨率特征上进行操作,使得整个网络更加轻便,运行速度较其他深度学习方法更有优势。

表 4 MRI T2 ->MRI T1 配准运行时间
Table 4 MRI T2 ->MRI T1 registration running time

方法	时间/s
VoxelMorph	0.406
Elastix	6.595
SyN	20.038
本文	0.281

4 结 论

1) 本文提出了一种基于深度学习的深度无监督多模态图像配准框架,不需要任何的标签数据。该框架由基于损失映射量的特征学习和基于最大后验概率的变形场学习组成。

2) 相较于监督方法,本文方法借助空间转换函数不需要 Ground Truth (GT),同时所提出的框架和损失函数可以转移到各种不同模态的医学图像应用中,具有普适性。

3) 大量的实验结果证明,相比于其他配准方法,本文方法能够在更短的时间内取得具有竞争力的配准精度。

参考文献 (References)

[1] MAINTZ J B A, VIERGEVER M A. A survey of medical image registration[J]. Medical Image Analysis, 1998, 2(1): 1-36.
[2] ASHBURNER J. A fast diffeomorphic image registration algorithm[J]. NeuroImage, 2007, 38(1): 95-113.
[3] GUIZAR-SICAÍROS M, THURMAN S T, FIENUP J R. Efficient subpixel image registration algorithms[J]. Optics Letters, 2008, 33(2): 156-158.
[4] MAES F, VANDERMEULEN D, SUTENS P. Medical image registration using mutual information[J]. Proceedings of the

IEEE, 2003, 10(91): 1699-1722.
[5] D' AGOSTINO E, MAES F, VANDERMEULEN D, et al. A viscous fluid model for multimodal non-rigid image registration using mutual information[J]. Medical Image Analysis, 2003, 7(4): 565-575.
[6] BALAKRISHNAN G, ZHAO A, SABUNCU M R, et al. An unsupervised learning model for deformable medical image registration[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 9252-9260.
[7] HU Y, MODAT M, GIBSON E, et al. Weakly-supervised convolutional neural networks for multimodal image registration[J]. Medical Image Analysis, 2018, 49: 1-13.
[8] QIN C, SHI B, LIAO R, et al. Unsupervised deformable registration for multi-modal images via disentangled representations[C]//International Conference on Information Processing in Medical Imaging. Berlin: Springer, 2019: 249-261.
[9] MAHAPATRA D, GE Z. Training data independent image registration using generative adversarial networks and domain adaptation[EB/OL]. [2020-07-21]. <https://doi.org/10.1016/j.patcog.2019.107109>.
[10] ZHANG Y X, LIU R S, LI Z, et al. Coupling principled refinement with bi-directional deep estimation for robust deformable 3D medical image registration[C]//Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging. Piscataway: IEEE Press, 2020: 86-90.
[11] LIU R S, LI Z, ZHANG Y X, et al. Bi-level probabilistic feature learning for deformable image registration[C]//International Joint Conference on Artificial Intelligence, 2020: 723-730.
[12] JADERBERG M, SIMONYAN K, KAVUKCUOGLU K, et al. Spatial transformer networks[C]//Conference and Workshop on Neural Information Processing Systems. New York: Curran Associates, 2015: 2017-2025.
[13] SUN D, YANG X, LIU M, et al. PWC-NET: CNNs for optical flow using pyramid, warping, and cost volume[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 8934-8943.
[14] ZACH C, POCK T, BISCHOF H. A duality based approach for real TV-L1 optical flow[J]. Pattern Recognition, 2007, 40: 214-223.
[15] 徐峻岭, 周毓明, 陈林, 等. 基于互信息的无监督特征选择[J]. 计算机研究与发展, 2012, 49(2): 158-168.
XU J L, ZHOU Y M, CHEN L, et al. An unsupervised feature selection approach based on mutual information[J]. Journal of Computer Research and Development, 2012, 49(2): 158-168 (in Chinese).
[16] MAHAPATRA D, ANTONY B, SEDAI S. Deformable medical image registration using generative adversarial networks[C]//Proceedings of the 2018 IEEE 15th International Symposium on Biomedical Imaging. Piscataway: IEEE Press, 2018: 1449-1453.
[17] MUELLER S G, WEINER M W, THAL L J, et al. Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's disease neuroimaging initiative (ADNI)[J]. Alzheimer's & Dementia, 2005, 1(1): 55-66.
[18] KISER K J, AHMED S, STIEB S M, et al. Data from the thoracic volume and pleural effusion segmentations in diseased lungs

for benchmarking chest ct processing pipelines [EB/OL]. [2020-08-22]. https://doi.org/10.7937/ncia.2020.6c7y-gq39.

[19] AERTS H J W L, WEE L, VELAZQUEZ E R, et al. Data from NSCLC-radiomics[EB/OL]. [2020-08-22]. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI.

[20] CHILAMKURTHY S, GHOSH R, TANAMALA S, et al. Deep learning algorithms for detection of critical findings in head CT scans: A retrospective study [J]. Lancet, 2018, 392 (10162): 2388-2396.

[21] MA J, GE C, WANG Y, et al. COVID-19 CT lung and infection segmentation dataset(Version1.0) [EB/OL]. [2020-08-22]. http://doi.org/10.5281/zenodo.3757476.

[22] AVANTS B B, TUSTISON N, GANG S. Advanced normalization tools (ANTs) [J]. Insight, 2009, 2 (365): 1-35.

[23] AVANTS B B, EPSTEIN C L, GROSSMAN M, et al. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain [J]. Medical Image Analysis, 2008, 12 (1): 26-41.

[24] KLEIN S, STARING M, MURPHY K, et al. Elastix: A toolbox for intensity based medical image registration[J]. IEEE Transactions on Medical Imaging, 2010, 29 (1): 196-205.

作者简介:
马腾宇 男, 硕士研究生。主要研究方向: 基于深度学习的多模态医学图像配准。
刘日升 男, 博士, 教授, 博士生导师。主要研究方向: 深度学习、计算机视觉、最优化方法。

Multimodal deformable registration based on unsupervised learning

MA Tengyu, LI Zi, LIU Risheng*, FAN Xin, LUO Zhongxuan

(International School of Information Science & Engineering, Dalian University of Technology, Dalian 116000, China)

Abstract: Multimodal deformable registration is designed to solve dense spatial transformations and is used to align images of two different modalities. It is a key issue in many medical image analysis applications. Multimodal image registration based on traditional methods aims to solve the optimization problem of each pair of images, and usually achieves excellent registration performance, but the calculation cost is high and the running time is long. The deep learning method greatly reduces the running time by learning the network used to perform registration. These learning-based methods are very effective for single-modality registration. However, the intensity distribution of different modal images is unknown and complex. Most existing methods rely heavily on label data. Faced with these challenges, this paper proposes a deep multimodal registration framework based on unsupervised learning. Specifically, the framework consists of feature learning based on matching amount and deformation field learning based on maximum posterior probability, and realizes unsupervised training by means of spatial conversion function and differentiable mutual information loss function. In the 3D image registration tasks of MRI T1, MRI T2 and CT, the proposed method is compared with the existing advanced multi-modal registration methods. In addition, the registration performance of the proposed method is demonstrated on the latest COVID-19 CT data. A large number of results show that the proposed method has a competitive advantage in registration accuracy compared with other methods, and greatly reduces the calculation time.

Keywords: deep learning; unsupervised; medical image registration; multimodal; computer vision

http://bhxb.buaa.edu.cn jbuua@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2020.0450

一种结合全局和局部相似性的小样本分割方法



刘宇轩, 孟凡满*, 李宏亮, 杨嘉莹, 吴庆波, 许林峰

(电子科技大学 信息与通信工程学院, 成都 610000)

摘 要: 针对小样本分割中如何提取支持图像和查询图像共性信息的问题, 提出一种新的小样本分割模型, 同时结合了全局相似性和局部相似性, 实现了更具泛化能力的小样本分割。具体地, 根据支持图像和查询图像全局特征和局部特征之间的相似性, 提出了一种新型注意力谱生成器, 进而实现查询图像的注意力谱生成和区域分割。所提注意力谱生成器包含 2 个级联模块: 全局引导器和局部引导器。在全局引导器中, 提出了一种新的基于指数函数的全局相似性度量, 对查询图像特征和支持图像的全局特征进行关系建模, 输出前景增强的查询图像特征。在局部引导器中, 通过引入局部关系矩阵对支持图像特征和查询图像特征之间的局部相似性进行建模, 得到与类别无关的注意力谱。在 Pascal-5ⁱ 数据集上做了大量的实验, 在 1-shot 设定下 mIoU 达到了 59.9%, 5-shot 设定下 mIoU 达到了 61.9%, 均优于现有方法。

关 键 词: 小样本语义分割; 全局相似性测度; 局部相似性测度; 知识迁移; 度量学习中图分类号: G202; U285.49; G623.58

文献标志码: A **文章编号:** 1001-5965(2021)03-0665-10

小样本语义分割旨在使用少量有标注的图像对新类目标进行分割, 其关键问题之一是如何提取支持图像(有标注的图像)和查询图像(待分割的无标注图像)的相似关系, 实现先验知识的迁移。现有的小样本分割方法^[1-5]常被建模为两支卷积神经网络, 分别是支持分支和查询分支。支持分支提取支持图像的分割先验, 查询分支则用于提取查询图像在嵌入空间中的特征, 根据支持分支的分割先验, 构建两者之间的相似关系, 实现对查询图像的分割。

然而, 现有的方法通过单独考虑图像的全局线索或者局部线索构建支持图像和查询图像的相似关系。例如, 一些方法^[1-2, 6-8]只考虑了全局线索, 而忽略了蕴含细节信息的局部线索, 使得模型难以实现细节的准确分割。另一方面, 一些方法如 LTM^[3]只考虑了局部线索, 忽略了携带有丰富语义信息的全局线索, 导致模型的通用性和泛化

性受限。综合两者的优劣, 本文将全局线索和局部线索相结合, 克服彼此的缺点, 使得两者相互促进, 获得更优的分割结果。

基于上述分析, 在建模支持图像分割先验和查询图像特征之间的关系时, 本文同时考虑了支持图像的全局线索和局部线索。具体地, 本文提出了一个新的注意力谱生成器, 根据支持图像和查询图像之间的关系, 更准确地输出查询图像的注意力谱。注意力谱生成器包含 2 个级联的模块: 全局引导器和局部引导器。全局引导器对特征间的全局相似性进行构建, 局部引导器建模特征间的局部相似性。本文提出一种新的基于指数函数的全局相似性度量方式, 可以对查询图像特征中与支持图像全局分割先验相似的部分进行增强, 同时对不相似的部分进行削弱, 更有效地对全局相似性进行度量。受到 LTM^[3]的启发, 基于局部关系矩阵, 使用局部引导器对局部相似性进行

收稿日期: 2020-08-24; 录用日期: 2020-09-11; 网络出版时间: 2020-10-12 13:29
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20201011.1352.001.html
基金项目: 国家自然科学基金(61871087); 四川省科技厅自然科学基金(2018JY0141)
* 通信作者。E-mail: fmmeng@uestc.edu.cn

引用格式: 刘宇轩, 孟凡满, 李宏亮, 等. 一种结合全局和局部相似性的小样本分割方法[J]. 北京航空航天大学学报, 2021, 47(3): 665-674. LIU Y X, MENG F M, LI H L, et al. A few shot segmentation method combining global and local similarity[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 665-674 (in Chinese).

建模,从而得到与类别无关的注意力谱。将全局引导器和局部引导器进行级联,实现了全局相似性和局部相似性的融合。在 Pascal-5ⁱ 数据集上对所提方法进行实验验证,所提方法在 1-shot 设定下 mIoU 达到 59.9%,5-shot 设定下 mIoU 达到了 61.9%,均优于现有方法。

本文所作出的贡献如下:

- 1) 提出了一种新的小样本分割模型,在进行关系建模时,同时考虑了支持图像的全局线索和局部线索,通过同时使用特征之间的全局相似性和局部相似性实现更准确的分割。
- 2) 在进行全局相似性建模时,提出了一种新的基于指数函数的全局相似性度量方式,能够更有效地建模全局相似性。
- 3) 在 Pascal-5ⁱ 数据集上进行了实验验证。所提方法优于最新的对比方法(在 1-shot 设定下 mIoU 达到了 59.9%,5-shot 设定下 mIoU 达到了 61.9%)。

1 相关工作

1.1 小样本学习

小样本学习旨在使用少量样本,学习可以迁移到未知类别的知识,进而实现未知类目标识别的泛化。现有的小样本学习方法可以分为 3 类:基于模型微调的方法^[9]、基于图模型的方法^[10]、基于度量学习的方法^[11-12]。基于模型微调的方法使用少量新类样本,对模型的参数进行微调。基于图模型的方法建立支持样本和查询样本之间的图模型,对新类样本进行识别。基于度量学习的方法度量支持样本和查询图像之间的距离,继而识别新类样本。目前,小样本学习已在图像分类、图像分割等领域得到了广泛的关注。

1.2 语义分割

语义分割对图像中每个像素点所属类别进行预测。基于全卷积网络(FCN)^[13],研究人员提出了一系列语义分割方法^[14-18],并验证了其有效性。为了得到含有丰富语义信息的特征表达,全卷积网络大量使用池化、跨步长卷积等操作,从而降低特征的分辨率,丰富特征的语义信息。不仅如此,为了提高感受野,DeepLab^[15]使用空洞卷积。然而,现有的语义分割方法不能对未知类目标分割进行泛化,因此,研究者提出小样本分割任务,旨在使用少量有标注实现未知类目标的有效分割。

1.3 小样本分割

小样本分割旨在使用少量有标注的图像对新类目标进行分割。现有的小样本分割方法根据其

思路可以分为 3 类:基于度量学习的方法^[1-3]、基于元学习的方法^[4-5]、基于原型学习的方法^[6-7]。基于度量学习的方法^[1-3]建模支持图像和查询图像之间的相似性,通过迁移支持图像分割先验实现查询图像的分割。基于元学习的方法^[4-5]利用支持分支提取的信息预测查询图像的分类器权重,进而实现查询图像的分割。基于原型学习的方法^[6-7]将支持图像的前景和背景分别映射为 2 个原型先验,然后使用距离最近的原则对查询图像进行分割。其中,基于度量学习的方法被广泛应用到小样本分割的任务中,并且取得了良好的效果。本文方法采用度量学习的思路进行建模。

1.4 国内小样本学习研究现状

文献[19]对融合小样本学习和零样本学习相关方法进行综述,主要介绍了当前小样本学习和零样本学习的一些主流方法。文献[20]将小样本学习与域迁移结合,应用于跨域人脸欺诈检测中。文献[21]利用度量学习的思路,提出了一种基于小样本学习的零器件表面缺陷检测的方法。文献[22]将小样本学习应用于点云分类中,提出了一种基于迁移学习的点云分类方法。文献[23]提出了一种基于已知类权重生成未知类权重的方法,有效地避免了小样本学习中灾难性遗忘的问题。

2 本文方法

2.1 小样本分割问题设置

对于未知类数据 D^{test} ,小样本分割网络致力于提取未知图像的前景区域。通常,分割网络在有标注的已知类数据集 D^{train} 上进行训练,其中 D^{train} 和 D^{test} 中的图像类别没有交集。具体地,小样本分割将 D^{train} 分为支持训练集和查询训练集。

对于 k -shot 设定,从支持集合的支持训练集和查询训练集中随机采样 k 张支持图像及其对应的二值掩膜进行训练。测试阶段,小样本网络则根据输入的 k 张支持图像及其对应的二值掩膜对未知的查询图像进行分割。

2.2 方法框架

本文提出一种新的小样本分割网络,致力于通过全局和局部线索对特征之间的关系建模以生成更精确的注意力谱。最终实现查询图像前景部分的增强和背景部分的削弱。

本文方法框架如图 1 所示,所提网络共包括 4 个步骤:

步骤 1 将支持图像和查询图像输入到特征提取器 CNN 中,得到支持图像特征 F_s 和查询图像特征 F_q 。

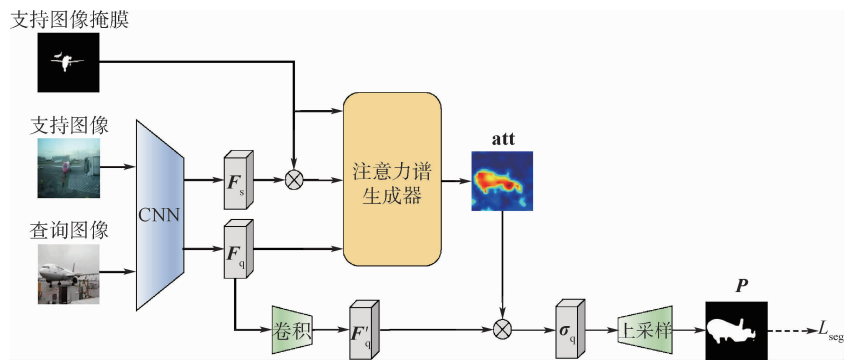


图 1 本文方法总体框架

Fig. 1 General framework of proposed method

步骤 2 为了滤除 F_s 的背景部分,接着将下采样后支持图像掩膜 M_s 和 F_s 相乘,得到的 F'_s 只包含支持图像的前景信息。

步骤 3 将 F'_s 、 F_q 以及 M_s 输入到基于全局和局部相似性度量的注意力谱生成器中,构建与类别无关的注意力谱 att ,以突出查询图像的前景区域。

步骤 4 将 att 与另一查询图像特征 F'_q 相乘,并输入到上采样网络,得到最终的预测结果 P 。

所提网络的核心是注意力谱生成器,其基于全局和局部相似性。对注意力谱 att 进行构建,其由 2 个级联的模块组成:全局引导器和局部引导器。全局引导器用于对特征之间的全局相似性进行度量。局部引导器则建模局部相似性。在全局引导器方面,提出了一种新的基于指数函数的全局相似性度量方式,可以对查询图像特征中与支持图像先验相似的部分进行增强,同时对不相似的部分进行削弱,更好地进行全局相似度量。在局部引导器方面,受到方法 LTM^[3] 的启发,基于局部关系矩阵,使用局部引导器对特征的局部相似性进行建模,从而得到最终与类别无关的注意力谱。

2.3 注意力谱生成器

本文基于全局和局部相似性对注意力谱 att 进行构建,具体如图 2 所示。具体思路是:将 F'_s 和 F_q 映射到同一个嵌入空间,在嵌入空间中对全局和局部相似性进行建模,最终生成与类别无关的

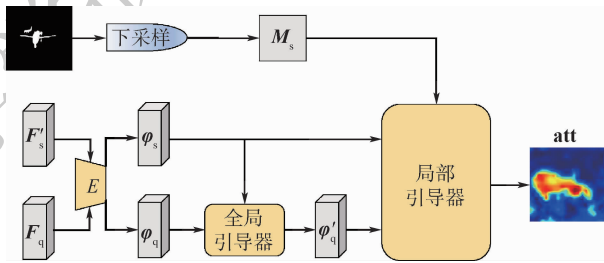


图 2 注意力谱生成器结构

Fig. 2 Structure of attention generator

注意力谱。

具体地, F'_s 和 F_q 被输入到一个嵌入模块 E 中,分别得到嵌入特征 $\varphi_s = E(F'_s)$ 和 $\varphi_q = E(F_q)$ 。然后,利用全局引导器,对 φ_s 和 φ_q 的全局相似性进行建模,实现 φ_q 与 φ_s 中全局信息相似的部分指数型增强,生成粗糙的查询图像特征 φ'_q 。接着利用局部引导器对 φ_s 和 φ'_q 中的每个空间位置之间的关系进行度量。通过将 M_s 、 φ_s 和 φ'_q 输入局部引导器中,得到最终与类别无关的注意力谱 att 。

2.4 全局引导器

现有的基于全局相似性度量的方法计算查询图像特征的每个空间位置与支持图像全局分割先验的余弦距离。然而,该方法将导致查询图像特征中前景部分和背景部分的可区分度较差的问题。针对上述问题,提出了一种新的全局引导器,如图 3 所示。

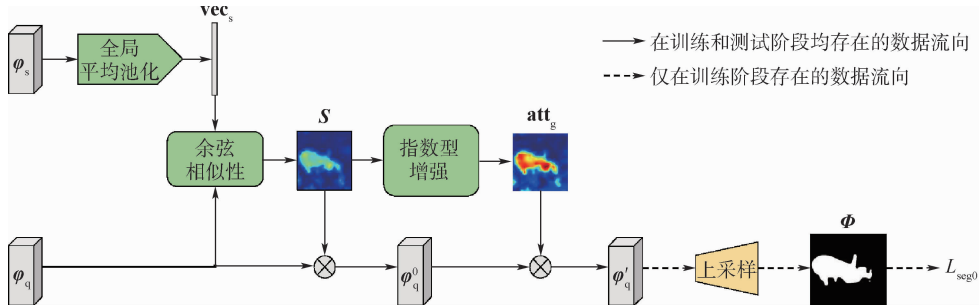


图 3 全局引导器的细节结构

Fig. 3 Detailed structure of global guider

2.4.1 全局余弦相似性

按照式(1),对 φ_s 进行全局平均池化,获得支持图像全局分割先验 vec_s 。进一步,按照式(2),计算 φ_q 每一个空间位置与 vec_s 的余弦距离,进而获得能够表征 φ_s 与 φ_q 之间全局余弦相似度的余弦相似谱 S 。 S_i 的值越高,表示当前位置 i 的局部特征 $\varphi_{q,i}$ 与 vec_s 越相似。

$$\text{vec}_s = \frac{1}{hw} \sum_{i=1}^{hw} \varphi_{s,i} \quad (1)$$

式中: h 为矩阵的高度维度大小; w 为矩阵的宽度维度大小。

$$S_i = \frac{\text{vec}_s^T \cdot \varphi_{q,i}}{\|\text{vec}_s\|_2 \cdot \|\varphi_{q,i}\|_2} \quad i = 1, 2, \dots, hw \quad (2)$$

2.4.2 基于指数函数的全局相似性度量

由于 S 的值域为 $[-1, 1]$, 本文期望对 S 值越高的部分进一步增强,对 S 值越低的部分进行削弱,从而得到一个新的可以表征距离的相似谱。基于上述分析,提出一种基于指数函数的全局相似性度量方式。具体地,基于式(3),增强了与 vec_s 相似的查询图像特征,削弱了与 vec_s 不相似的查询图像特征。

$$\text{att}_g^i = \lambda^{S_i} \quad \lambda > 1, i = 1, 2, \dots, hw \quad (3)$$

式中: λ 为超参数。

可以看出,如果 $S_i = 1$, 则表明当前位置 i 的特征 $\varphi_{q,i}$ 与 vec_s 有较高的相似性,所提方法可以得到一个新的 att_g^i 注意力值,它的值为 λ , 是大于 1 的,故所提方法对位置 i 进行了增强。相反地,若 $S_i = -1$, 则表明当前位置 i 的特征 $\varphi_{q,i}$ 与 vec_s 不相似,所提方法计算得到的 att_g^i 值为 $1/\lambda$, 若 λ 越大,则 att_g^i 越接近 0, 从而对位置 i 进行了削弱。

在实验中,将 λ 设为 10、20、40, 分别获得 $\text{att}_{g1} = 10^S$ 、 $\text{att}_{g2} = 20^S$ 、 $\text{att}_{g3} = 40^S$ 。进而对 att_{g1} 、 att_{g2} 、 att_{g3} 取平均,得到如式(4)所示的 att_g , 其表征了全局相似性度量。

$$\text{att}_g = (\text{att}_{g1} + \text{att}_{g2} + \text{att}_{g3})/3 \quad (4)$$

基于上述原理,获得了基于全局相似性度量的注意力谱 att_g , 其对查询图像特征中与支持图像全局分割先验相似的部分进行增强,同时对不相似的部分进行削弱。

因此,如图 3 所示,根据式(5),所提全局引导器可获得前景与背景可区分度更高的查询嵌入特征 φ'_q 。

$$\varphi'_q = \varphi_q \otimes S \otimes \text{att}_g \quad (5)$$

式中:“ $\otimes$ ”为逐元素相乘。

为了获得更具鲁棒性的 φ'_q , 本文在训练阶段

将 φ'_q 进行上采样,输出粗糙的分割结果 Φ 。这里使用如式(6)所示的交叉熵损失对此过程进行监督。

$$L_{\text{seg0}} = - \sum_i \sum_j [G_{i,j} \ln \Phi_{i,j} + (1 - G_{i,j}) \ln(1 - \Phi_{i,j})] \quad (6)$$

式中: $G_{i,j}$ 为查询图像标签掩膜在位置 (i, j) 的值; $\Phi_{i,j}$ 为 Φ 在位置 (i, j) 的值。

2.5 局部引导器

包含丰富细节的局部线索对于目标分割任务非常重要,而局部相似性由局部线索所刻画。受到 LTM^[3] 的启发,本文基于局部关系矩阵,使用局部引导器对每一对局部空间特征的相似性进行建模,从而得到最终的与类别无关的注意力谱。

2.5.1 局部关系矩阵

本文对支持图像和查询图像特征之间的局部相似性进行测度。所提方法考虑类似 Non-local^[24] 中提出的关系矩阵。给定 2 个特征 F_1 和 F_2 , 其尺寸为 $C \times H \times W$, 度量 F_1 和 F_2 局部相似性的局部关系矩阵 $R(F_1, F_2)$ 为

$$R(F_1, F_2)_{i,j} = \frac{F_{1,i}^T \cdot F_{2,j}}{\|F_{1,i}\|_2 \cdot \|F_{2,j}\|_2} \quad (7)$$

所得 $R(F_1, F_2)$ 为 $(H \times W) \times (H \times W)$ 尺寸的矩阵,其中的每个元素 $R(F_1, F_2)_{i,j}$ 表征了局部特征 $F_{1,i}$ 和 $F_{2,j}$ 的余弦相似性。

本文的局部引导器对 φ'_q 和 φ_s 的每一对局部特征点进行比较,根据式(7)计算局部关系矩阵 $R = R(\varphi'_q, \varphi_s)$ 。

2.5.2 基于局部关系矩阵的注意力谱构建

令 M_q 和 M_s 分别为查询图像和支持图像降采样后的掩膜。在理想情况下,式(8)成立:

$$\varphi'_q \cdot \varphi_s = M_q \cdot M_s \quad (8)$$

在式(8)中, M_q 和 M_s 的尺寸被调整为 $(H \times W) \times 1$ 和 $1 \times (H \times W)$, φ'_q 和 φ_s 分别调整尺寸为 $(H \times W) \times C$ 和 $C \times (H \times W)$ 。

局部引导器的目标是:根据局部相似性,构建与类别无关的注意力谱 att 。由于 att 对 φ'_q 中的前景部分进行增强,对背景部分进行削弱,可以将其视为查询图像的粗糙分割结果。因此,式(8)近似为

$$\varphi'_q \cdot \varphi_s = \text{att} \cdot M_s \quad (9)$$

更进一步,将式(9)的左侧替换为 R , 得到

$$R = \text{att} \cdot M_s \quad (10)$$

显然, M_s 为行满秩矩阵,其右逆 M_s^+ 可根据式(11)计算:

$$\mathbf{M}_s^{\prime-} = \mathbf{M}_s^{\text{H}} \cdot (\mathbf{M}_s \cdot \mathbf{M}_s^{\text{H}})^{-1} \tag{11}$$

式中: $\mathbf{M}_s^{\text{H}}$ 为 $\mathbf{M}_s$ 的共轭转置。

因此,基于式(12),得到最终的注意力谱 $\mathbf{att}$ 。最后,将 $\mathbf{att}$ 调整大小为 $H \times W$ 。

$$\mathbf{att} = \mathbf{R} \cdot \mathbf{M}_s^{\prime-} \tag{12}$$

2.5.3 局部引导器实现流程

局部引导器具体细节如图 4 所示。将 φ'_q 和 φ_s 的尺寸分别调整为 $(H \times W) \times C$ 和 $C \times (H \times W)$,将其输入基于式(7)的协同注意力模块中,获得局部关系矩阵 $\mathbf{R}$ 。同时,所提网络根据式(11)对重置尺寸的 $\mathbf{M}_s$ 计算右逆 $\mathbf{M}_s^{\prime-}$ 。最后,基于式(12)构建最终的注意力谱 $\mathbf{att}$ 。

在训练阶段,由于所提网络的与类无关性,将所有前景类别视为“前景”一类,背景视为一类。为了获得更具鲁棒性的 $\mathbf{att}$,所提方法将 $\mathbf{att}$ 转化为一个 $2 \times H \times W$ 的概率谱 ρ 。具体地,由于 $\mathbf{att}$ 的值域为 $[0,1]$,将其视为前景概率谱,则 $1 - \mathbf{att}$ 可视为背景概率谱。进而将 $\mathbf{att}$ 和 $1 - \mathbf{att}$ 按通道

维度拼接,获得 $\rho = \text{Concat}(\mathbf{att}, 1 - \mathbf{att})$,从而实现从注意力谱 $\mathbf{att}$ 到概率谱 ρ 之间的转化。进而使用式(13)所示的交叉熵损失对其进行优化。

$$L_a = - \sum_i \sum_j [G_{i,j} \ln \rho_{i,j} + (1 - G_{i,j}) \ln (1 - \rho_{i,j})] \tag{13}$$

式中: L_a 为该损失函数; $\rho_{i,j}$ 为 ρ 在位置 (i,j) 的值。

本文的局部引导器受到 LTM^[3] 的启发,不同之处在于:LTM^[3] 仅考虑局部相似性,并且对局部关系矩阵 $\mathbf{R}$ 的构建使用 L_2 损失函数监督。该方法的弊端有 2 点:①缺乏全局分割先验的引导;②由于 L_2 损失是逐元素计算的,缺乏结构信息。这对测试阶段新类的 $\mathbf{R}$ 的构建是不利的,使得该方法的可扩展性受限。本文方法在将查询图像特征 φ'_q 输入局部引导器时,已经滤除了大量背景,不需要对 $\mathbf{R}$ 的构建进行监督,这样使得生成的 $\mathbf{att}$ 更具鲁棒性。

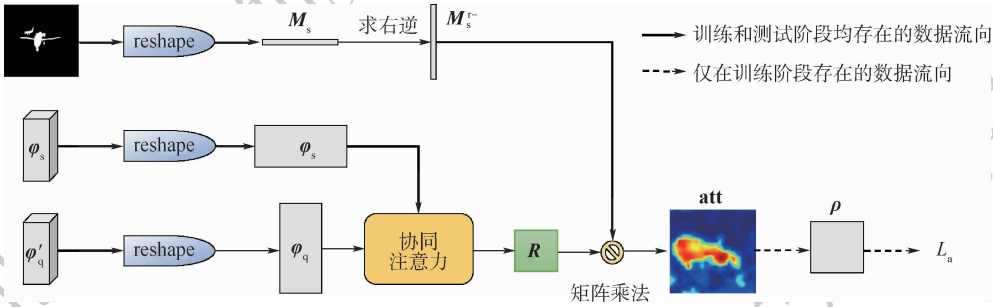


图 4 局部引导器的细节结构

Fig. 4 Detailed structure of local guider

2.6 上采样网络

获得注意力谱 $\mathbf{att}$ 以后,将 $\mathbf{att}$ 与 $\mathbf{F}'_q$ 相乘,获得最终的查询图像特征 σ'_q ,此时已经对其背景部分进行了有效的滤除。所提网络将 σ'_q 输入到上采样网络中,获得最终的预测结果 $\mathbf{P}$ 。上采样网络结构如图 5 所示,使用了空洞空间金字塔池化(ASPP)^[14] 进行多尺度融合,使得结果更具鲁棒性。使用式(14)对这个过程进行监督。

$$L_{\text{seg}} = - \sum_i \sum_j [G_{i,j} \ln P_{i,j} + (1 - G_{i,j}) \ln (1 - P_{i,j})] \tag{14}$$

式中: $P_{i,j}$ 为 $\mathbf{P}$ 在位置 (i,j) 的值。

在整个训练阶段,使用的损失函数为

$$L = L_{\text{seg}} + L_a + L_{\text{seg0}} \tag{15}$$

2.7 k-shot 分割

在 k -shot 设定下,LTM^[3] 输入 k 张支持图像,每张支持图像与查询图像建模局部关系,分别生成 k 个注意力谱 $[\mathbf{att}_1, \mathbf{att}_2, \dots, \mathbf{att}_k]$,进而将其平均化,得到最终的注意力谱 $\mathbf{att}$ 。其弊端在于平均化的过程中丢失了部分局部线索,使得注意力谱更加粗糙。为了更好地实现对全局相似性和局部相似性的结合,提出一种新的基于拼接的 k -shot 分割方案。

将 k 张支持图像 $[S_1, S_2, \dots, S_k]$ 同时输入特征提取器中,分别得到 k 个特征 $[f_s^1, f_s^2, \dots, f_s^k]$,将其特征按高度维度进行拼接,可以得到一个融合

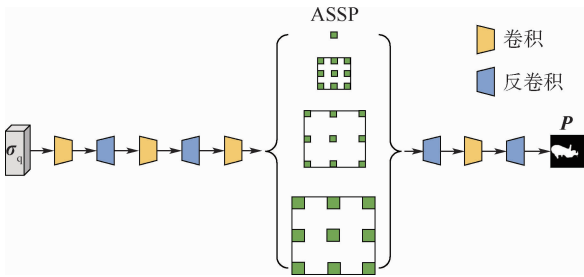


图 5 上采样网络架构

Fig. 5 Framework of up-sample network

特征 f_s 。同时,所提网络对查询图像提取特征,获得 f_q 。将 k 张支持图像分割掩膜 $[M_s^1, M_s^2, \dots, M_s^k]$ 按高度维度拼接,得到 f_s 对应的掩膜 M_s 。进而,将 f_s, f_q 和 M_s 输入所提注意力谱生成器中,先后进行2.4节所提全局相似性度量和2.5节所提局部相似性度量,获得 f_s 对应的注意力谱 $\mathbf{att}$ 。同时将 f_q 输入卷积层,得到 f'_q 。将 f'_q 与 $\mathbf{att}$ 逐元素相乘,进而输入2.6节所提上采样网络,得到 k -shot分割结果。

所提 k -shot方案优点在于建模局部相似性时,对查询图像特征与 k 个支持图像特征每个点的余弦距离进行测度,避免了局部信息的丢失,能够生成更精细的注意力谱 $\mathbf{att}$ 。

3 实验

3.1 实现细节

为了验证方法的有效性,在Pascal-5ⁱ数据集上进行了实验验证,使用Pytorch深度学习框架对模型构建,所有的实验平台为Nvidia Titan X GPU,使用Adam优化器对所提模型参数进行优化,学习率设置为 10^{-4} 。本文使用了不同的主干网络作为特征提取器验证性能,分别为在ImageNet^[25]上预训练的VGG16^[26]、ResNet50^[27]、ResNet101^[27]。在训练阶段,冻结了特征提取器的参数;在测试阶段,使用了多尺度策略和Dense-CRF^[28]。将查询图像调整大小为 320×320 、 416×416 、 224×224 ,对提取的多尺度特征进行平均化,进而进行后续操作。

3.2 Pascal-5ⁱ数据集

Pascal-5ⁱ是基于Pascal数据集和SBD数据集构建的。Pascal-5ⁱ数据集中一共有20类目标,按照现有小样本分割研究的常用设置,将其分成4个子集,每个子集图像有5类目标图像。

数据集的划分如表1所示。实验阶段,采用交叉验证法,轮流将Fold0、Fold1、Fold2、Fold3作为测试集,另外3个子集作为训练集,得到4个训练-测试集为 $[(\text{Fold1}, \text{Fold2}, \text{Fold3}), \text{Fold0}]$, $[(\text{Fold0}, \text{Fold2}, \text{Fold3}), \text{Fold1}]$, $[(\text{Fold0}, \text{Fold1}, \text{Fold3}), \text{Fold2}]$, $[(\text{Fold0}, \text{Fold1}, \text{Fold2}), \text{Fold3}]$ 。

表1 Pascal-5 ⁱ 四个子集的划分	
Table 1 Four subsets setting of Pascal-5 ⁱ	
子集	类别
Fold0	飞机、自行车、鸟、船、瓶子
Fold1	公交车、轿车、猫、椅子、牛
Fold2	餐桌、狗、马、摩托车、人
Fold3	盆栽、山羊、沙发、火车、显示器

3.3 评价指标

使用了所有前景类别的平均交并比mIoU作为评价指标,这个指标被广泛地运用于小样本分割任务评价中。mIoU值越高,模型的性能越好。

同时,也使用前景交并比和背景交并比的平均值FB-IoU作为参考指标。

3.4 对比实验

3.4.1 不同主干网络下的对比实验

由于不同主干网络对模型性能有较大的影响,为了保证公平性,分别使用不同的主干网络,和其他现有方法进行对比,mIoU对比结果如表2及表3所示,表中最佳实验结果加粗。

表2 不同主干网络下,本文与现有方法的1-shot对比实验mIoU结果

Table 2 Comparative experimental results (mIoU) of proposed method and existing methods under 1-shot setting using different backbone networks %						
主干网络	方法	mIoU				平均值
		F0	F1	F2	F3	
VGG16	OSLSM ^[4]	33.6	55.3	40.9	33.5	40.8
	Co-FCN ^[29]	36.7	50.6	44.9	32.4	41.2
	SG-One ^[1]	40.2	58.4	48.4	38.4	46.4
	PANet ^[6]	42.3	58.0	51.1	41.2	48.2
	FWB ^[8]	47.0	59.6	52.6	48.3	51.9
	本文	50.3	65.3	53.0	50.8	54.9
ResNet50	CANet ^[2]	52.5	65.9	51.3	51.9	55.4
	LTM ^[3]	54.6	65.6	56.6	51.3	57.0
	CRNet ^[30]					55.7
	本文	54.6	67.8	57.4	52.1	58.0
ResNet101	FWB ^[8]	51.3	64.5	56.7	52.2	56.2
	本文	57.5	68.7	58.7	54.5	59.9

表3 不同主干网络下,本文与现有方法的5-shot对比实验mIoU结果

Table 3 Comparative experimental results (mIoU) of proposed method and existing methods under 5-shot setting using different backbone networks %						
主干网络	方法	mIoU				平均值
		F0	F1	F2	F3	
VGG16	OSLSM ^[4]	35.9	58.1	42.7	39.1	44.0
	Co-FCN ^[29]	37.5	50.0	44.1	33.9	41.4
	SG-One ^[1]	41.9	58.6	48.6	39.4	47.1
	PANet ^[6]	51.8	64.6	59.8	46.5	55.7
	FWB ^[8]	50.9	62.9	56.5	50.1	55.1
	本文	50.3	66.3	54.7	55.3	56.7
ResNet50	CANet ^[2]	55.5	67.8	51.9	53.2	57.1
	LTM ^[3]	56.4	66.6	56.9	56.8	59.2
	CRNet ^[30]					58.8
	本文	54.8	68.1	59.9	56.2	59.8
ResNet101	FWB ^[8]	54.9	67.4	62.2	55.3	60.0
	本文	58.1	69.8	60.8	58.9	61.9

表 2 与表 3 实验结果表明,所提方法使用不同的主干网络,在 1-shot 设定和 5-shot 设定下实验 mIoU 结果均优于现有方法。

同时,本文进一步在 FB-IoU 指标下,与其他现有方法进行对比。表 4 和表 5 分别列出了在 1-shot和 5-shot 设定下,所提方法与其他现有方法的 FB-IoU 对比实验结果,表中最佳实验结果加粗。由表 4 可以看出,所提方法使用不同的主干网络,在 1-shot 设定下结果均优于现有方法。表 5 的 5-shot 实验结果也充分说明了所提方法的有效性。

表 4 不同主干网络下,本文与现有方法的 1-shot 对比实验 FB-IoU 结果

Table 4 Comparative experimental results (FB-IoU) of proposed method and existing methods under 1-shot setting using different backbone networks %

主干网络	方法	FB-IoU				平均值
		F0	F1	F2	F3	
VGG16	OSLSM ^[4]					61.3
	Co-FCN ^[29]					60.1
	SG-One ^[1]					63.1
	PANet ^[6]					66.5
	本文	68.6	77.3	65.3	68.4	69.9
ResNet50	CANet ^[2]	71.0	76.7	54.0	67.2	67.2
	CRNet ^[30]					66.8
	本文	71.5	78.7	70.6	69.1	72.5
ResNet101	本文	73.7	79.4	71.8	70.2	73.8

表 5 不同主干网络下,本文与现有方法的 5-shot 对比实验 FB-IoU 结果

Table 5 Comparative experimental results (FB-IoU) of proposed method and existing methods under 5-shot setting using different backbone networks %

主干网络	方法	FB-IoU				平均值
		F0	F1	F2	F3	
VGG16	OSLSM ^[4]					61.5
	Co-FCN ^[29]					60.2
	SG-One ^[1]					65.9
	PANet ^[6]					70.7
	本文	68.0	77.6	66.5	71.8	71.0
ResNet50	CANet ^[2]	74.2	80.3	57.0	66.8	69.6
	CRNet ^[30]					71.5
	本文	71.1	78.8	72.7	69.7	73.1
ResNet101	本文	73.5	80.0	72.6	72.9	74.8

3.4.2 全局相似性度量对比实验

为了验证所提出的全局相似性度量方式的有效性,进一步将模型中全局相似性度量方式替换为余弦距离 (SG-One^[1]中提出) 和通道维度拼接的方式 (CANet^[2]中提出)。并且为了消除局部引导器对实验结果的影响,本节实验仅仅保留全局相似性建模,即仅保留了全局引导器。本节使用

ResNet50 作为主干网络,对 3 种全局相似性度量方式进行了对比,mIoU 结果如表 6 所示,表中最佳实验结果加粗。可以看出,表 6 所示实验结果证明了本文提出的全局相似性度量方式的有效性。

表 6 全局相似性度量方式的对比实验 mIoU 结果
Table 6 Comparative experimental results (mIoU) of global similarity metric %

全局度量方式	mIoU	
	1-shot	5-shot
余弦距离	51.3	53.2
通道维度拼接	46.5	47.2
所提全局相似性度量	53.8	55.7

3.4.3 k-shot 设定下的对比实验

为验证本文 k-shot 设定方案的有效性,在 5-shot 设定下,对方法 LTM^[3]中平均化注意力谱的方式和所提 k-shot 方案进行对比,对比实验 mIoU 结果如表 7 所示,表中最佳实验结果加粗。表 7 所示实验结果证明了所提 k-shot 方案的有效性。

表 7 5-shot 设定方案对比实验 mIoU 结果
Table 7 Comparative experimental results (mIoU) under 5-shot setting %

设定方式	mIoU
平均化注意力谱	59.1
所提 k-shot 方案	59.7

3.5 消去实验

3.5.1 全局引导器和局部引导器的消去实验

本文提出了 2 个级联的模块:全局引导器和局部引导器。为了验证这 2 个模块的有效性,在 1-shot 和 5-shot 设定下,使用 ResNet50 作为主干网络,进行了消去实验。实验测得 mIoU 结果如表 8 所示,表中最佳实验结果加粗。

从表 8 可以看出,当单独使用全局引导器时,模型效果最差,这是因为仅使用全局引导器时,缺少了局部细节的刻画。单独使用局部引导器时,模型性能略有提升,其原因是局部细节的刻画更加丰富,不足之处在于缺少全局信息的引导。当同时使用全局引导器和局部引导器时,模型的性能较好,此时模型已经充分结合了全局相似性和局部相似性。

表 8 全局引导器和局部引导器的消去实验 mIoU 结果
Table 8 Ablation experimental results (mIoU) of global guider and local guider %

全局引导器	局部引导器	mIoU	
		1-shot	5-shot
√		53.8	55.7
	√	56.8	59.0
√	√	58.0	59.7

3.5.2 损失函数的消去实验

所提方法使用的损失函数 L 包含 3 项: L_{seg} 、 L_a 和 L_{seg0} , L_{seg} 为整体损失函数的核心,用于监督最终预测结果的生成。因此,实验中始终保持 L_{seg} ,对其余 2 项进行消去实验。使用 ResNet50 作为主干网络,在 1-shot 和 5-shot 设定下,进行本节的消去实验,实验的 mIoU 结果如表 9 所示,表中最佳实验结果加粗。

可知,当本文同时使用 3 项损失函数时,所提模型性能较好。

表 9 损失函数的消去实验 mIoU 结果

Table 9 Ablation experimental result (mIoU) of loss function				
L_{seg}	L_a	L_{seg0}	mIoU	
			1-shot	5-shot
✓			55.4	57.7
✓	✓		56.6	58.6
✓		✓	55.9	57.9
✓	✓	✓	58.0	59.7

3.6 可视化结果

3.6.1 分割效果较好的可视化结果

进一步列出了部分高质量的分割结果,如图 6 所示。可知,由于结合了全局和局部线索,本文方法对轮廓的刻画较为精细。

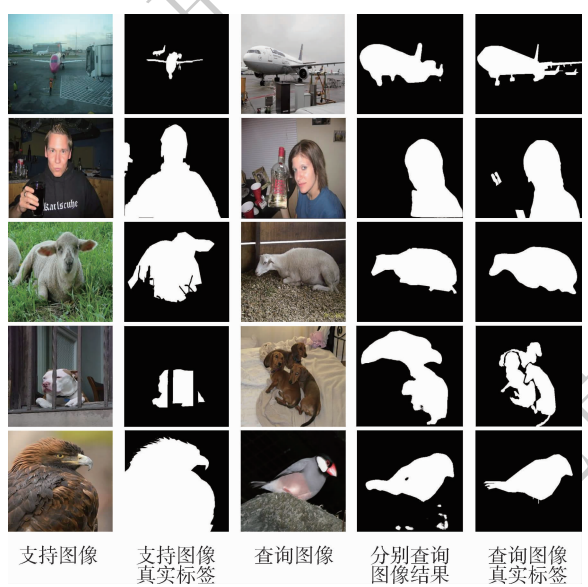


图 6 部分分割效果较好的可视化结果

Fig. 6 Some visualized high-quality segmentation results

3.6.2 分割效果较差的可视化结果

对于分割效果较差的图片,也进一步展示了部分典型代表,并且对其分割失败原因进行分析。部分分割效果较差的可视化结果如图 7 所示。

对于第 1 行可视化结果,由于查询图像的目标存在遮挡,且查询图像与支持图像外表差异较

大,所提方法对其不能很好地进行分割。对于第 2、3 行可视化结果,查询图像中含有猫、狗 2 类目标,由于猫和狗的类间差异较小,所提网络将其视为一类目标进行分割。这是小样本分割的难点,也是将来进一步研究的重点。

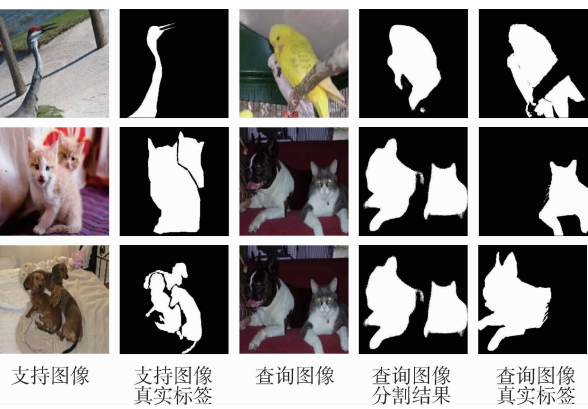


图 7 部分分割效果较差的可视化结果

Fig. 7 Some visualized low-quality segmentation results

4 结束语

本文提出了一种新的小样本分割模型,在构建注意力谱时,有效地结合全局相似性和局部相似性。该注意力谱生成器包含 2 个级联的模块:全局引导器和局部引导器。全局引导器用于建模特征间的全局相似性,局部引导器用于建模特征间的局部相似性。具体来说,在全局引导器中,提出了一种新的全局相似性度量,对查询图像特征和支持图像全局分割先验进行关系建模,更有效地输出前景增强、背景削弱的查询图像特征。进一步通过引入基于局部关系矩阵的局部引导器,所提方法对支持图像特征和查询图像局部特征之间的相似性进行建模,最后得到与类别无关的注意力谱。Pascal-5ⁱ 数据集的实验验证了所提方法的有效性。

参考文献 (References)

[1] ZHANG X, WEI Y, YANG Y, et al. Sg-One: Similarity guidance network for one-shot semantic segmentation [J]. IEEE Transactions on Cybernetics, 2020, 50 (9) : 3855-3865.

[2] ZHANG C, LIN G, LIU F, et al. CANet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019: 5217-5226.

[3] YANG Y, MENG F, LI H, et al. A new local transformation module for few-shot segmentation [C] // International Conference on Multimedia Modeling. Berlin: Springer, 2020: 76-87.

[4] SHABAN A, BANSAL S, LIU Z, et al. One-shot learning for se-

- mantic segmentation[EB/OL]. [2020-07-18]. <https://arxiv.org/abs/1709.03410>.
- [5] TIAN P, WU Z, QI L, et al. Differentiable meta-learning model for few-shot semantic segmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2020:12087-12094.
- [6] WANG K, LIEW J H, ZOU Y, et al. PANet: Few-shot image semantic segmentation with prototype alignment[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019:9197-9206.
- [7] LIU J, QIN Y. Prototype refinement network for few-shot segmentation[EB/OL]. (2020-05-09) [2020-07-18]. <https://arxiv.org/abs/2002.03579>.
- [8] NGUYEN K, TODOROVIC S. Feature weighting and boosting for few-shot segmentation[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2019:622-631.
- [9] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[J/OL]. (2017-07-18) [2020-07-18]. <https://arxiv.org/abs/1703.03400>.
- [10] KIM J, KIM T, KIM S, et al. Edge-labeling graph neural network for few-shot learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2019:11-20.
- [11] SNEEL J, SWERSKY K, ZEMEL R. Prototypical networks for few-shot learning[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2017:4077-4087.
- [12] SUNG F, YANG Y, ZHANG L, et al. Learning to compare: Relation network for few-shot learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018:1199-1208.
- [13] LONG J, SHELHAMER E, DARRELL T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2015:3431-3440.
- [14] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. [2020-07-18]. <https://arxiv.org/abs/1706.05587>.
- [15] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 40(4):834-848.
- [16] LIN G, MILAN A, SHEN C, et al. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017:1925-1934.
- [17] RONNEBERGER O, FISCHER P, BROX T. U-Net: Convolutional networks for biomedical image segmentation[C]// International Conference on Medical Image Computing and Computer-Assisted intervention. Berlin: Springer, 2015:234-241.
- [18] ZHAO H, SHI J, QI X, et al. Pyramid scene parsing network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2017:2881-2890.
- [19] 潘崇煜, 黄健, 郝建国, 等. 融合零样本学习和小样本学习的弱监督学习方法综述[J]. 系统工程与电子技术, 2020, 42(10):2246-2256.
- PAN C Y, HUANG J, HAO J G, et al. Survey of weakly supervised learning integrating zero-shot and few-shot learning[J]. Journal of Systems Engineering and Electronics, 2020, 42(10):2246-2256 (in Chinese).
- [20] 孙文赞, 金忠, 赵海涛, 等. 基于深度特征增广的跨域小样本人脸欺诈检测算法[J/OL]. 计算机科学[2020-09-27]. <http://kns.cnki.net/kcms/detail/50.1075.TP.20200911.1518.024.html>.
- SUN W Y, JIN Z, ZHAO H T, et al. Cross-domain few-shot face presentation attack detection method based on deep feature augmentation[J/OL]. Computer Science, 2020 [2020-09-27]. <http://kns.cnki.net/kcms/detail/50.1075.TP.20200911.1518.024.html> (in Chinese).
- [21] 于重重, 萨良兵, 马先钦, 等. 基于度量学习的小样本零部件表面缺陷检测[J]. 仪器仪表学报, 2020(7):214-223.
- YU C C, SA L B, MA X Q, et al. Few-shot parts surface defect detection based on the metric learning[J]. Chinese Journal of Scientific Instrument, 2020(7):214-223.
- [22] 雷相达, 王宏涛, 赵宗泽. 基于迁移学习的小样本机载激光雷达点云分类[J]. 中国激光, 2020, 47(11):1110002.
- LEI X D, WANG H T, ZHAO Z Z. Small sample airborne LiDAR point cloud classification based on transfer learning[J]. Chinese Journal of Lasers, 2020, 47(11):1110002.
- [23] 李文煜, 帅仁俊, 郭汉. 克服小样本学习中灾难性遗忘方法研究[J]. 计算机应用与软件, 2020, 37(7):136-141.
- LI W Y, SHUAI R J, GUO H. Overcoming catastrophic forgetting in few-shot learning[J]. Computer Applications and Software, 2020, 37(7):136-141 (in Chinese).
- [24] WANG X, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018:7794-7803.
- [25] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009:248-255.
- [26] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-11-18) [2020-07-18]. <https://arxiv.org/abs/1409.1556>.
- [27] 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.
- [28] KRÄHENBÜHL P, KOLTUN V. Efficient inference in fully connected CRFs with Gaussian edge potentials[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2011:109-117.
- [29] RAKELLY K, SHELHAMER E, DARRELL T, et al. Conditional networks for few-shot semantic segmentation[C]. International Conference on Learning Representations, 2018.
- [30] LIU W, ZHANG C, LIN G, et al. CRNet: Cross-reference networks for few-shot segmentation[C]// Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 4165-4173.

分割。

孟凡满 男, 博士, 副教授, 博士生导师。主要研究方向: 智能图像分析、深度学习。

作者简介:

刘宇轩 男, 硕士研究生。主要研究方向: 计算机视觉与图像

A few shot segmentation method combining global and local similarity

LIU Yuxuan, MENG Fanman*, LI Hongliang, YANG Jiaying, WU Qingbo, XU Linfeng

(School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 610000, China)

Abstract: Few shot segmentation aims at segmenting objects of novel classes with few annotated images, whose key is to extract the general information between support and query images. The existing methods utilize global features or local features to obtain the general information and validate their effectiveness. However, these two kinds of features' similarity are considered separately in the existing methods, while their mutual effect is ignored. This paper proposes a novel few shot segmentation model using both global and local similarity to achieve more generalizable few shot segmentation. Specifically, an attention generator is proposed to build the attention map of query images based on the relationship between support and query images. The proposed attention generator consists of two cascaded modules: global guider and local guider. In global guider, a novel exponential function based global similarity metric is proposed to model the relationship between query images and support images with respect to global feature, and output foreground-enhanced query image features. As for local guider, a local relationship matrix is introduced to model the local similarity between support and query image features, and obtain the class-agnostic attention map. Comprehensive experiments are performed on Pascal-5ⁱ idataset. The proposed method achieves a mean IoU value of 59.9% under 1-shot setting, and a mean IoU of 61.9% under 5-shot setting, which outperforms many state-of-the-art methods.

Keywords: few shot semantic segmentation; global similarity measure; local similarity measure; knowledge transferring; metric learning

Received: 2020-08-24; Accepted: 2020-09-11; Published online: 2020-10-12 13:29

URL: kns.cnki.net/kcms/detail/11.2625.V.20201011.1352.001.html

Foundation items: National Natural Science Foundation of China (61871087); Natural Science Foundation of Sichuan Science and Technology Department (2018JY0141)

* Corresponding author. E-mail: fmmeng@uestc.edu.cn

《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 47 卷 第 3 期 2021 年 3 月

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS (JBUA)

(Monthly, Started in 1956)

Vol.47 No.3 March 2021

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 Qinping

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

中国标准连续出版物号: ISSN 1001-5965
CN 11-2625/V

国内定价: 50.00 元 / 期

ISSN 1001-5965

