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

中国图象图形学报

2021
07
VOL.26
ISSN1006-8961
CN11-3758/TB

中国多媒体大会(ChinaMM2020)专刊

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第26卷第7期（总第303期）

2021年7月16日

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

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

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

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

Journal of Image and Graphics

Title inscription: Song Jian | Monthly, Started in 1996

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

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

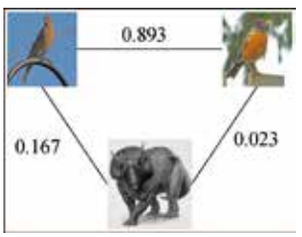
Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

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

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

2021年7月 第26卷 第7期（总第303期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次



类别语义相似性监督的小样本图像识别(第1594页)

像素聚合和特征增强的任意形状场景文本检测(第1614页)

时序特征融合的视频实例分割(第1692页)

综述

面向上行流媒体的压缩感知视频流技术前沿
刘浩，黄荣，袁浩东 …… 1545

图像处理和编码

耦合保持投影哈希跨模态检索
闵康凌，张国宾，王磊，李丹萍 …… 1558
面向视觉搜索的空间局部敏感哈希方法
黄小燕，孙彬，杨展源，朱映映，田奇 …… 1568

图像分析和识别

融合视觉关系检测的电力场景自动危险预警
高明，左红群，柏帆，田清阳，葛志峰，董兴宁，甘甜 …… 1583
类别语义相似性监督的小样本图像识别
徐鹏帮，桑基韬，路冬媛 …… 1594
局部双目视差回归的目标距离估计
张羽丰，李昱希，赵明璧，喻晓源，占云龙，林巍峣 …… 1604
像素聚合和特征增强的任意形状场景文本检测
师广琛，巫义锐 …… 1614
相位一致性指导的全参考全景图像质量评价
夏雨蒙，王永芳，王闯 …… 1625

图像理解和计算机视觉

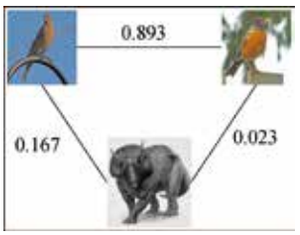
特征金字塔结构的时序行为识别网络
何嘉宇，雷军，李国辉 …… 1637
多模态多层次事件网络的谣言检测
李莎，张怀文，钱胜胜，方全，徐常胜 …… 1648
多模态零样本人体动作识别
吕露露，黄毅，高君宇，杨小汕，徐常胜 …… 1658
足球视频球员感知跟踪算法
冯思佳，宋子恺，于俊清，何云峰，管涛 …… 1668
融合时空图卷积的多人交互行为识别
成科扬，吴金霞，王文杉，荣兰，詹永照 …… 1681
时序特征融合的视频实例分割
黄泽涛，刘洋，于成龙，张加佳，王轩，漆舒汉 …… 1692

医学图像处理

多尺度深度特征提取的肝脏肿瘤CT图像分类
毛静怡，宋余庆，刘哲 …… 1704
3D多尺度深度卷积神经网络肺结节检测
孙华聪，彭延军，郭燕飞，张晓庆 …… 1716
结合多通道注意力的糖尿病性视网膜病变分级
顾婷菲，郝鹏翼，白琮，柳宁 …… 1726

Volume 26, Number 7
Published July 16, 2021

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS



Few shot image recognition based on class semantic similarity supervision(P1594)

Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement (P1614)

Video instance segmentation based on temporal feature fusion (P1692)

Review

Survey on compressive sensing video stream for uplink streaming media
Liu Hao, Huang Rong, Yuan Haodong ····· 1545

Image Processing and Coding

Structure-preserving hashing with coupled projections for cross-modal retrieval
Min Kangling, Zhang Guobin, Wang Lei, Li Danping ····· 1558

Locality-sensitive hashing approach based on semantic space for visual retrieval
Huang Xiaoyan, Sun Bin, Yang Zhanyuan, Zhu Yingying, Tian Qi ····· 1568

Image Analysis and Recognition

Visual relationship detection-based emergency early-warning description generation in electric power industry
Gao Ming, Zuo Hongqun, Bai Fan, Tian Qingyang, Ge Zhifeng, Dong Xingning, Gan Tian ···· 1583

Few shot image recognition based on class semantic similarity supervision
Xu Pengbang, Sang Jitao, Lu Dongyuan ····· 1594

Object distance estimation based on stereo regional disparity regression
Zhang Yufeng, Li Yuxi, Zhao Mingbi, Yu Xiaoyuan, Zhan Yunlong, Lin Weiyao ····· 1604

Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement
Shi Guangchen, Wu Yirui ····· 1614

Phase consistency guided full-reference panoramic image quality assessment algorithm
Xia Yumeng, Wang Yongfang, Wang Chuang ····· 1625

Image Understanding and Computer Vision

Temporal action detection based on feature pyramid hierarchies
He Jiayu, Lei Jun, Li Guohui ····· 1637

Multi-modal multi-level event network for rumor detection
Li Sha, Zhang Huaiwen, Qian Shengsheng, Fang Quan, Xu Changsheng ····· 1648

Multimodal-based zero-shot human action recognition
Lyu Lulu, Huang Yi, Gao Junyu, Yang Xiaoshan, Xu Changsheng ····· 1658

Players-aware tracking algorithm in soccer video
Feng Sijia, Song Zikai, Yu Junqing, He Yunfeng, Guan Tao ····· 1668

Multi-person interaction action recognition based on spatio-temporal graph convolution
Cheng Keyang, Wu Jinxia, Wang Wenshan, Rong Lan, Zhan Yongzhao ····· 1681

Video instance segmentation based on temporal feature fusion
Huang Zetao, Liu Yang, Yu Chenglong, Zhang Jiajia, Wang Xuan, Qi Shuhan ····· 1692

Medical Image Processing

CT image classification of liver tumors based on multi-scale and deep feature extraction
Mao Jingyi, Song Yuqing, Liu Zhe ····· 1704

3D multi-scale deep convolutional neural networks in pulmonary nodule detection
Sun Huacong, Peng Yanjun, Guo Yanfei, Zhang Xiaoqing ····· 1716

Diabetic retinopathy grading based on multi-channel attention
Gu Tingfei, Hao Pengyi, Bai Cong, Liu Ning ····· 1726

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2021)07-1637-11

论文引用格式:He J Y, Lei J and Li G H. 2021. Temporal action detection based on feature pyramid hierarchies. Journal of Image and Graphics, 26(07):1637-1647(何嘉宇,雷军,李国辉. 2021. 特征金字塔结构的时序行为识别网络. 中国图象图形学报,26(07):1637-1647)[DOI:10.11834/jig.200495]

特征金字塔结构的时序行为识别网络

何嘉宇,雷军,李国辉*

国防科技大学信息系统工程重点实验室,长沙　410072

摘　要:**目的**　时序行为识别是视频理解中最重要的任务之一,该任务需要对一段视频中的行为片段同时进行分类和回归,而视频中往往包含不同时间长度的行为片段,对持续时间较短的行为片段进行检测尤其困难。针对持续时间较短的行为片段检测问题,文中构建了3维特征金字塔层次结构以增强网络检测不同持续时长的行为片段的能力,提出了一种提案网络后接分类器的两阶段新型网络。**方法**　网络以RGB连续帧作为输入,经过特征金字塔结构产生不同分辨率和抽象程度的特征图,这些不同级别的特征图主要在网络的后两个阶段发挥作用:1)在提案阶段结合锚方法,使得不同时间长度的锚段具有与之对应的不同大小的感受野,锚段的初次预测将更加准确;2)在感兴趣区域池化阶段,不同的提案片段映射给对应级别特征图进行预测,平衡了分类和回归对特征图抽象度和分辨率的需求。**结果**　在THUMOS Challenge 2014数据集上对模型进行测试,在与没有使用光流特征的其他典型方法进行比较时,本文模型在不同交并比阈值上超过了对比方法3%以上,按类别比较时,对持续时间较短的行为片段检测准确率则普遍得到提升。消融性实验中,在交并比阈值为0.5时,带特征金字塔结构的网络则超过使用普通特征提取网络的模型1.8%。**结论**　本文提出的基于3维特征金字塔特征提取结构的双阶段时序行为模型能有效提升对持续时间较短的行为片段的检测准确率。

关键词:时序行为识别;特征金字塔;深度学习;计算机视觉;视频理解

Temporal action detection based on feature pyramid hierarchies

He Jiayu, Lei Jun, Li Guohui*

Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha　410072, China

Abstract: **Objective** Temporal action localization is one of the most important tasks in video understanding and has great application prospects in practice. With the rise of various online video applications, the number of short videos on the Internet has increased sharply, many of which contain different human behaviors. A model that can automatically locate and classify human action segments in videos is needed to detect and distinguish human behavior in short videos quickly and efficiently. However, public security departments also need real-time human behavior detection systems to help monitor and provide early warning of public safety incidents. In the task of temporal action localization, the human action segments in a video must be classified and regressed simultaneously. Accurately locating the boundaries of human behavior segments is more difficult than classifying known segments. A video always contains action segments of different temporal lengths, and detecting action segments with a short duration is especially difficult because short-duration action segment is easily ignored by the detection model or regarded as part of a closer, longer-duration segment. Existing methods have various attempts to

收稿日期:2020-08-24;**修回日期**:2020-12-28;**预印本日期**:2021-01-04

*通信作者:李国辉　guohli@nudt.edu.cn

基金项目:国家自然科学基金项目(71673293,61806215)

Supported by: National Natural Science Foundation of China (71673293, 61806215)

improve the detection accuracy of human behavior fragments with different durations. In this paper, a 3D feature pyramid hierarchy is proposed to enhance the network's ability to detect action segments of different temporal durations. **Method** A new two-stage network with a proposal network followed by a classifier named 3D feature pyramid convolutional network (3D-FPCN) is proposed. In 3D-FPCN, feature extraction is performed through the 3D feature pyramid feature extraction network built. The 3D feature pyramid feature extraction network has a bottom-up pathway and a top-down pathway. The bottom-up pathway simultaneously encodes the temporal and spatial characteristics of consecutive input frames through a series of 3D convolutional neural networks to obtain highly abstract feature maps. The top-down pathway uses a series of deconvolutional networks and lateral connection layers to fuse high-abstraction and high-resolution features, and obtain low-level feature maps. Through the feature pyramid feature extraction network, multilevel feature maps with different abstraction levels and different resolutions can be obtained. Highly abstract feature maps are used for the classification and regression of long-duration human action segments, and high-resolution feature maps are used for the regression and classification of short-duration human action segments, which can effectively improve the detection effect of the network on human behavior fragments of different durations. The whole network takes RGB frames as input and generates feature maps of different resolutions and abstract degrees via a feature pyramid structure. These feature maps of different levels mainly play a role in the latter two stages of the network. First, the anchor mechanism is used in the proposal stage. Thus, anchor segments of different temporal lengths have corresponding receptive fields of different sizes, and this is equivalent to a receptive field calibration. Second, in the region of interest pooling stage, different proposal segments are mapped to corresponding level feature maps for prediction, which makes feature prediction more targeted and balances the requirements for the abstraction and resolution of feature maps for action segments' classification and regression. **Result** Our model is evaluated on the THUMOS'14 dataset. Compared with other classic methods that do not use optical flow features, our network surpasses most of them. Specifically, when the intersection over union threshold is set to 0.5, the mean average precision (mAP) of 3D-FPCN is up to 37.4%. Compared with the classic two-stage network region convolutional 3D network (R-C3D), the mAP of our method is increased by 8.5 percentage points. The comparison results of the detection precision on different class human action segments when the intersection ratio threshold is 0.5 are shown. The detection result of 3D-FPCN for short-duration human actions segments is greatly improved compared with other methods. For example, 3D-FPCN's detection accuracy of basketball dunk and cliff diving is 10% higher than that of the same two-stage network method R-C3D, and the detection accuracy of pole vault is higher than the multi-stage segment convolutional neural network (SCNN) is about 40%. This finding proves the improvement of our model for detecting short-duration human action segments. An ablation test is also conducted in the feature pyramid feature extraction network to explore the effect of this structure on the model. When the feature pyramid structure is removed from the network, the detection accuracy of the network is approximately 2% lower than before when the intersection over union threshold is 0.5. When only the multilevel feature map generated by the feature pyramid structure is used in the first stage of the network, which is the proposal generation stage, the detection accuracy is only 0.2% higher than the model with the feature pyramid structure removed. This finding proves that the feature pyramid hierarchy can effectively enhance the detection of action with different durations, and it mainly works in the second stage of the network, which is region of interest pooling stage. **Conclusion** A two-stage temporal action localization network 3D-FPCN is proposed based on 3D feature pyramid feature extraction network. The network takes continuous RGB frames as input, which can quickly and effectively detect human action segments in short videos. Through a number of experiments, the superiority of the model is proven, and the mechanism of the 3D feature pyramid structure in the model is discussed and explored. The 3D feature pyramid structure effectively improves the model's ability to detect short-duration human action segments, but the overall mAP of the model remains low. In the next work, the model will be improved, and different feature inputs will be introduced to study the method of temporal action localization further. We hope that our work can inspire other researchers and promote the development of the field.

Key words: temporal action localization; feature pyramid network; deep learning; computer vision; video understanding

0 引 言

时序行为识别是视频内容理解中的重要任务，任务目标是从未修剪的完整视频中检测人类行为片段，将行为分类为几种动作类别之一，并精确定位行为片段的开始和结束时间，因此，时序行为识别是分类和回归共存的检测任务。该任务有非常广泛的应用前景，而实际应用中的视频往往具有背景杂乱、行为多样等特点，加之不同行为持续时长不等，边界也较为模糊，人工标注行为边界都有一定困难，因此，现有时序行为识别的各种方法在公开数据集上的准确率都不高。

面对复杂多变的待检测视频，一个好的时序行为识别模型应该能对不同持续时间长度的行为片段精准地预测边界和类别。在构建模型时，主要考虑两个问题：一是待分类和回归的片段（提案片段）如何产生，二是怎样准确分类和回归。从这个角度来说，现有方法可以分为两大类，一类是自底向上的方法，一类是自上而下的方法。自底向上的方法从更细粒度的级别如帧级别或单元级别进行预测，再通过合并的方法产生提案片段；自上而下的方法则从较高的级别进行预测，通常使用滑窗或锚方法产生提案片段，两种思想一般都使用后接一个分类回归网络对提案片段进行进一步处理，因此模型的设计最好能高效提取和利用特征。

使用 3 维卷积神经网络（convolutional neural networks，CNN）对视频进行特征的提取和利用，是解决该任务的一个非常有效的手段，但网络参数的设置、网络的层数不同，获得的特征图也不同。一般来讲，网络层数越深，获得的特征越抽象，对行为类别的预测需要高度凝练的抽象信息，对片段起止时间的预测则需要细粒度的时间信息。实际应用中，视频内行为片段时间尺度变化往往较大，单一尺度的特征图很难同时适应高抽象和细粒度的要求，因此，本文引入 3 维的特征金字塔结构，利用不同抽象程度和精细度的特征对不同时间长度的行为片段进行预测，以求达到分类和回归需求的平衡。在提案片段的产生上，基于特征金字塔网络产生的不同尺度特征图应用锚方法，产生不同时间尺度的提案片段，一方面，锚方法相较于其他产生提案片段的方法，既能利用现有特征图完成一次初步预测，又能通过锚段缩放比的调节完成对不同时间尺度的全覆盖，另一方面，锚方法与多尺度特征图结合，使得不同尺度特征图上产生的锚段在原视频尺度上感受野是不同的，间接地达到了感受野校准的目的。时序行为识别的评价指标是全类平均正确率（mean average precision，mAP）在不同交并比（intersection over union，IoU）阈值下的值，交并比即预测片段与实际片段（ground truth，GT）交集与并集的比例，比例越高说明重叠越多，预测越准确。容易推得，更高分辨率（级别较低）的特征图应当用于预测时间尺度较小的行为片段，而更抽象的特征图（级别较高）则应用于预测持续时间较长的行为片段。这是因为即使模型对行为片段的边界预测结果略有偏差，持续时间较长的行为片段也可以容忍较大的误差而满足交并比阈值要求。因此，稍微降低对持续时间较长的行为片段的边界回归精度要求，以确保使用抽象程度较高的特征图对其进行更加准确的动作类别预测。相应地，持续时间较短的行为片段对边界回归的结果更加敏感。对于持续时间较短的行为片段，本文使用更高分辨率的特征图来确保边界回归尽量准确。综上，本文提出了 3 维特征金字塔卷积神经网络（3D feature pyramid convolutional network，3D-FPCN），网络以待检测视频的 RGB 帧作为输入，输入帧在空间上尺度恒定，这些帧通过预训练的 3 维卷积网络（可以是任意的典型 3 维卷积网络）进行编码，再通过特征金字塔结构，产生 3 个不同时间尺度的特征图。然后，时序提案子网络在不同尺度特征图上应用锚方法产生提案片段。最后，分类子网络对提案片段进行分类和回归。

本文的主要贡献是：1）首次将自上而下的特征金字塔结构用于两阶段网络的时序行为识别；2）通过消融性实验研究了特征金字塔结构的有效性；3）在 THUMOS14 数据集（Jiang 等，2014）上进行测试，并与其他典型方法进行了比较，本文方法在结果上具有很强的竞争力，尤其是对于持续时间较短的行为片段进行预测上，预测的结果要好于许多常规方法。此外，提出的模型可以端到端训练，便于整体优化，实现了特征提取和行为片段检测的统一。

1 相关工作

Gaidon 等人（2011）首先提出了在未修剪的视

频中定位动作片段的任务。早期的方法是通过滑动窗口后接支持向量机(support vector machine, SVM)分类器来进行预测,但该方法边界预测的准确率很低。深度学习的方法在计算机视觉领域得到推广后,许多模型开始使用深度神经网络进行行为片段的检测。多阶段卷积神经网络(multi-stage segment convolutional neural network, SCNN)(Shou 等,2016)通过滑窗的方法产生待检测的时间片段,再通过3 个连续的3 维卷积神经网络(convolutional 3D network, C3D)(Tran 等,2015)分别对片段进行提案、分类和定位,该方法取得了不错的效果,但对每段滑窗都进行预测,特征提取时重复计算,冗余程度较高。而 R-C3D(region convolutional 3D network)(Xu 等,2017)和 TAL-Net(temporal action localization network)(Chao 等,2018)借鉴目标检测中经典结构 Faster RCNN 并进行改进,其中 TAL-Net 同时使用 RGB 帧和预提取的光流帧作为特征提取网络 I3D(two-stream inflated 3D ConvNet)(Carreira 和 Zisserman,2017)的输入,并在锚段产生的阶段使用扩张卷积来校准锚的感受野,使得模型检测的准确率相较于纯粹复刻 Faster RCNN 的方法 R-C3D 大幅提升。由于视频行为具有其独特性,一些学者通过挖掘视频中内容的特点来提高时序行为识别的效果,Long 等人(2019)使用高斯核来动态优化每一个行为的时间尺度,并通过融合高斯核进行检测,提出了高斯时间感知网络(Gaussian temporal awareness networks, GTAN),模型取得较好的检测结果,给其他方法提供了一种新的思路。Lin 等人(2019)提出了一种不同于滑窗或锚机制的待检测片段生成方法,即边界匹配网络(boundary-matching network, BMN),创新之处在于文中新定义了一种边界匹配置信度图,通过对每个位置的边界置信度判断提案片段的位置,该模型在公开数据集上取得了较好的检测准确率。Zeng 等人(2019)认为,之前方法关注点只在单个提案片段上,没有研究提案片段之间的联系,而同一段视频一般包含的是同一类别或近似类别的行为,因此同一段视频的不同提案片段包含相似的信息,能够互相影响,甚至一些背景帧也能提供关于运动类别的信息,该文创新性地引入了图卷积网络的方法学习提案片段之间的联系,通过对相邻度最高的提案片段节点进行图卷积完成对某段提案的分类和回归。除了这些基于提案进行预测的方法,一些模型采用了帧级别或单元级别进行预测的方法。而 TURN-TAP(temporal unit regression network)网络(Gao 等,2017b)将视频帧分为许多等时间长度的单元,再从单元级别进行特征提取和预测,其最大的优点是检测速度很快。卷积反卷积网络(convolutional-de-convolutional networks, CDC)(Shou 等,2017)则在 3 维卷积网络进行特征提取后,用3 次时间维反卷积操作,将特征图时间尺度扩张回原输入尺度,从而实现帧级别逐帧预测,通过设定阈值合并的方式得到预测的结果。目前,许多学者已经在该领域做出贡献,但时序行为识别各种方法在各大型数据集上的准确率都还处于一个较低的水平,亟待学者们提出新的想法和模型,提高检测的准确率。本文尝试使用特征金字塔结构的 3 维卷积网络产生不同分辨率和抽象程度的特征图,通过针对性地分配不同时间长度的提案片段给不同级别的特征图来提高预测的准确率,在时序感兴趣区域(region of interest, RoI)池化时使用扩张的提案片段进行池化,以包含上下文信息,同时防止回归时提案区域小于实际行为片段区域的现象。

2 特征金字塔结构的时序行为识别

本文提出了一个包含特征金字塔结构的 3 维卷积神经网络,是一种用于时序行为识别的新网络,目的在于增强较大范围的时间尺度上检测人类行为片段的能力。如图 1 所示,网络可以分为 3 部分:1)用于特征提取的 3 维特征金字塔网络;2)时序提案网络;3)分类和回归网络。由 3 维特征金字塔网络提取的多级特征图将由时序提案网络使用,并在分类和回归网络中进行复用,从而保证了预测的效率,避免重复提取特征。时序提案网络生成可能包含动作的不同长度的提案片段,而分类和回归网络将这些提案片段分类为特定行为类别或背景,并进一步修整提案片段的边界。

2.1 3 维特征金字塔网络

3 维特征金字塔网络以 RGB 帧序列作为输入,在多个级别上输出与输入帧成一定比例尺度大小的特征图。首先使用一个 3 维卷积神经网络同时提取时间和空间特征,并将输出特征图称为基础特征图,该过程为自下而上的过程。用于提取基础特征图的卷积神经网络可以采用任何典型的 3D 卷积神经网

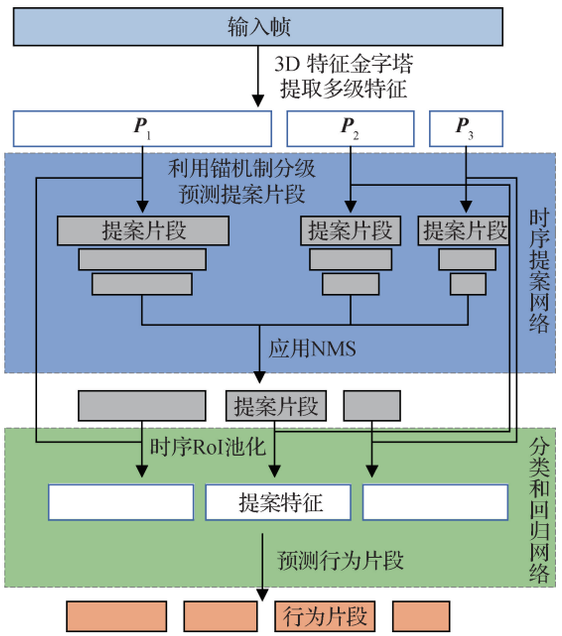


图 1　3D-FPCN 网络结构

Fig. 1　Architecture of 3D-FPCN

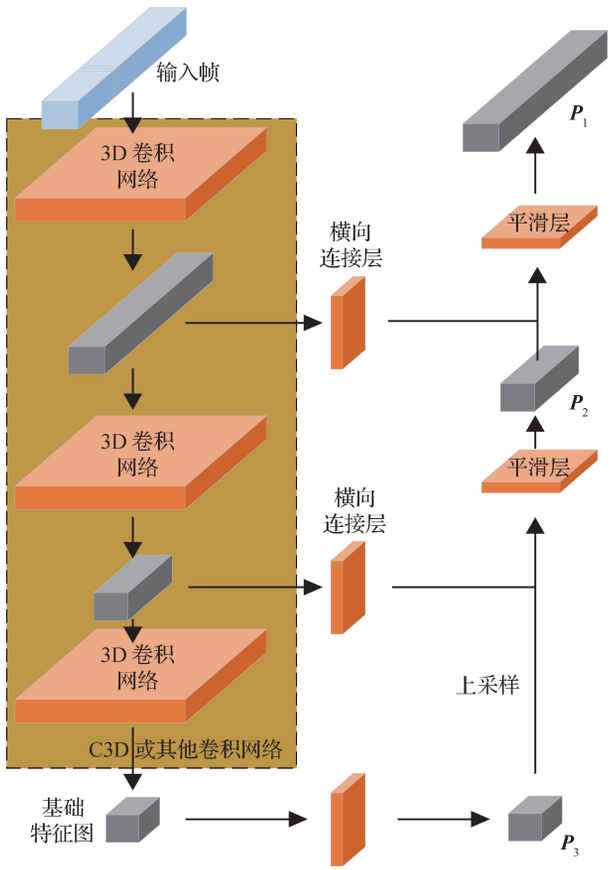


图 2　3 维特征金字塔层次结构

Fig. 2　3D feature pyramid hierarchy

络，选择传统的 C3D 网络（Tran 等，2015）。假设输入 RGB 帧的尺寸为 $3 \times L \times H \times W$，经自下而上的过程获得的基础特征图 $\boldsymbol{C}_{5b}$ 的尺寸则为 $512 \times L/8 \times H/16 \times W/16$（其中 512 是 $\boldsymbol{C}_{5b}$ 层的通道数）。使用 C3D 网络的 $\boldsymbol{C}_{1a}$ 到 $\boldsymbol{C}_{5b}$ 层构建了自下而上的通道，在基础特征图以及 $\boldsymbol{C}_{3b}$ 和 $\boldsymbol{C}_{4b}$ 之后的特征图间添加自上而下的路径和横向连接的路径，以此来构建特征金字塔结构，其在时间维度相对于输入帧的缩放比为分别为 8、4、2。自上而下的路径由多层上采样层组成，而横向连接层则是卷积核为 $1 \times 1 \times 1$ 的 3 维卷积层。自上而下的路径能够从较高的金字塔层级对时间尺度较短但语信息更强的高抽象特征图进行上采样来提取更高分辨率的特征，而横向连接可以提供语义信息上较弱，但时间维分辨率更高的特征图，即加入更细粒度的时序信息。在网络中，每一个横向连接将会合并来自同一级的特征图和来自自上而下通道的特征图，生成一幅新的特征图。

3 维特征金子塔的结构如图 2 所示。$\boldsymbol{C}_{5b}$ 后接一个 $1 \times 1 \times 1$ 卷积核的 3 维卷积层，便获得了特征图 $\boldsymbol{P}_3$，自上而下的路径始于特征图 $\boldsymbol{P}_3$。对每个特征图 $\boldsymbol{P}_n$，在时间维对其进行缩放比为 2 的上采样，接着将上采样得到的特征图与相应级别的自下而上的特征图进行合并，合并前，相应级别的自下而上的特征图会经历一个卷积核大小为 $1 \times 2 \times 2$ 的 3 维卷积层，以缩小空间尺度，调整通道数与自上而下的特征图相适应，合并的方式为逐元素相加。此后，合并的特征图通过一个卷积核为 $3 \times 3 \times 3$ 的 3 维卷积层（平滑层），以减小不同层特征图合并的混叠效应，并将特征图的空间尺度由 $H/16 \times W/16$ 降到 1×1，最终得到该层待使用的特征图 $\boldsymbol{P}_{n-1}$。这个过程将会重复直至生成时间维分辨率最精细的特征图 $\boldsymbol{P}_1$。3D-FPCN 中，共生成了 3 个不同级别的待使用特征图 $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3$，分别对应于 $\boldsymbol{C}_{5b}, \boldsymbol{C}_{4b}, \boldsymbol{C}_{3b}$ 级别的时间维度。特征图 $\boldsymbol{P}_1, \boldsymbol{P}_2, \boldsymbol{P}_3$ 将用于时序提案网络，并在分类和回归网络中重用。

2.2　时序提案网络

时序提案网络主要有 3 个功能：1）从各特征图产生锚段；2）为每个锚段指定正或负标签，即明确为前景或背景，并对锚段进行初步的边界回归；3）应用非极大值抑制（non-maximum suppression，NMS）的方法来挑选出一定数量的可能包含人类活动的时序片段。在时序提案网络中，网络生成可能包含行为片段、行为类别未知或时间长度不同的提

案片段。在特征金字塔结构产生多级特征图的基础上,使用锚方法产生不同时间尺度的待分类和回归片段。在不同级别上,定义的锚段具有不同的时间尺度,具体来讲,在每一级待使用的特征图 $\boldsymbol{P}_n$ 上,以时间维度的每个像素为中心,均匀地产生数个时间长度不同的锚段,这样每幅特征图的每个时间位置都指定了 K 个长度不同比例固定的锚段,即可以覆盖所有不同时间长度的行为片段。于是,锚段的总数将有 $K \times \left(\frac{L}{2}+\frac{L}{4}+\frac{L}{8}\right)$ 个。

特征图 $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3$ 每个时间位置上产生的512维特征向量用于预测相对于中心位置和行为片段长度 $c_i,l_i,i \in \{1,\cdots,K\}$ 的相对偏移量 $\{\delta c_i,\delta l_i\}$,同时也对锚段内容预测其属于前景还是背景的分数。对锚段的偏移和前背景预测通过在特征图 $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3$ 上各自添加卷积核大小为 $1 \times 1 \times 1$ 的卷积层来实现。使用锚方法产生待检测的片段的潜在好处是不同抽象程度的特征图在时间维相对于原始视频具有不同的缩放比,基于不同级别的特征图上的像素应用锚方法来产生不同时间长度的锚段,实际上完成锚段的感受野校准,较长的锚段在高抽象大缩放比的特征图上产生,对应的感受野相应就较大,较短的锚段在细粒度小缩放比的锚段上产生,对应的感受野就较小。在调整锚段的边界后,网络采用非极大值抑制法(NMS)来从这些锚段中挑选出更可能包含行为片段的提案片段用于训练和测试。

2.3 分类和回归网络

顾名思义,该网络的主要任务是对先前网络获得的提案片段进行分类,并进一步回归其时间维度边界。网络的关键是采用新的3维RoI池化策略来提取每个提案片段的特征。首先,不同时间长度的提案片段将指定给不同的金字塔层级,如一个时间长度为 l 的提案片段将通过式(1)指定给特征图 P_k,即

$$k_l = \lfloor k_0 + \lg(l/l_v) \rfloor$$
$$k = \begin{cases} 1 & k_l \leqslant 1 \\ k_l & \text{其他} \\ 3 & k_l \geqslant 3 \end{cases} \tag{1}$$

式中,l_v 代表输入视频的时间长度,k_0 是一个用于调整不同时间长度的提案片段分配给不同特征图的数量的常量,k_0 越大,将有越多的提案片段分配给高抽象度的特征图,k_0 越小,则更多的提案片段分配给高分辨率的特征图,为使得不同时间长度的提案片段在不同级别的特征图上分布均匀,应通过多次对比实验确定实验数据集中合适的 k_0 值。直观来看,式(1)意味着时间长度较短的提案片段将映射到较低级别的特征图上,而低级别的特征图特点是分辨率更高。然后,3维RoI池化层将从每提案片段对应的特征图中提取所需要的特征。虽然输入的提案片段具有不同的时间长度,但经过RoI池化后输出的特征向量在时间维的长度是相同的。具体来说,如果某个输入提案片段时间长度为 l,它被指定到特征图 $\boldsymbol{P}_k$,该提案片段映射到特征图 $\boldsymbol{P}_k$ 中的特征块 $\boldsymbol{C}_i$ 大小为 $l_i \times h \times w$,其中特征块的时间维长度进行了等同于自身长度的两侧扩张,以包含上下文信息,那么3维RoI池化层会将特征块 $\boldsymbol{C}_i$ 分割成 $1 \times 4 \times 4$ 的子特征块,最大池化将会应用到每个子特征块上,这样,每个具有任意时序长度的提案片段经过RoI池化后将输出固定的相同大小为 $512 \times 1 \times 4 \times 4$ 的特征向量。最后,将特征向量输入两个全连接层,获得对行为类别的预测分数以及边界偏移的预测。

2.4 网络优化

由于时序提案网络与分类和回归网络具有相似的分类和回归任务,可以使用相同的损失函数通过同时优化分类损失和回归损失来训练这两个网络。具体来说,使用交叉熵作为分类任务的损失函数,使用smooth L1损失函数作为回归任务的损失函数,其联合损失函数为

$$L_s = \frac{1}{N_c}\sum_i L_c(p_i,p_i^*) + \lambda \frac{1}{N_r}\sum_i p_i^* L_r(t_i,t_i^*) \tag{2}$$

式中,L_s 为子网络损失,L_c 和 L_r 分别是分类和回归任务的损失,N_c 和 N_r 代表在时序提案网络中用于训练的锚段的数目,值得注意的是,分类和回归子网中提案片段的数量应当等于批大小(batch size),因为在时序提案网络中对其进行了采样。λ 是损失函数中设置的权重参数,实验中被赋值为1。i 指锚段或提案片段在一批(batch)的所有锚段或提案片段中的索引号。p_i 是预测的锚段或提案片段 i 的类别概率,p_i^* 则是相对应的实际标签。$\boldsymbol{t}_i = \{\delta\hat{c}_i,\delta\hat{l}_i\}$ 是预测出的锚段或提案片段相对于其对应的行为片段的时间偏移量,$\boldsymbol{t}_i^* = \{\delta c_i,\delta l_i\}$ 是实际的时间偏移量。为了消除不同行为片段时间长度对损失大小的

影响,在计算损失时都采用相对偏移量,即

$$\begin{cases}\delta c_i = \dfrac{c_i^* - c_i}{l_i} \\ \delta l_i = \lg(l_i^*/l_i)\end{cases} \tag{3}$$

式中,c_i 和 l_i 分别指锚段或提案片段 i 的时间维度上的中心位置和时间长度;c_i^* 和 l_i^* 则指代其对应的实际行为片段的中心位置和时间长度。整个网络的目标函数为

$$L = L_p + L_c \tag{4}$$

式中,L_p 和 L_c 分别是时序提案网络、分类和回归网络的损失,在训练时可以同时进行优化,这样即可对所有网络进行联合训练。

3　实　验

本文在 THUMOS'14(Jiang 等,2014)数据集上对网络进行测试,THUMOS'14 是广泛用于动作识别和时序行为识别任务的大型数据集。在时序行为识别任务中,使用 THUMOS'14 中的验证集和测试集。THUMOS'14 共有 20 类行为,其提供的验证集和测试集是未经裁剪的视频。其中,验证集中包含 200 个视频,3 007 段行为片段,并带有标签标注了每个行为片段的起止时间;测试集中包含 213 个视频,3 358 段行为片段,也带有标签标注了行为片段的起止时间。每个视频平均时长 3 min,较长的视频也有达到 6 ～7 min,最短的视频只有几十秒,可以看出这些视频的时间长度跨度较大,使得任务有较大的挑战性。除此之外,每个视频中不一定只包含一个行为片段,有的视频包含了 3 段甚至更多行为片段,要将它们都准确检测并定位,难度也不小。

3.1　模型参数设置

对输入的视频进行分段裁剪和采样,使得每一段输入帧固定为 768 帧。模型的输入经过特征金字塔结构产生多级特征图 $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3$,为了与分类器和回归器通道数相匹配,将所有特征图的通道数固定为 512 维。多级特征图结合锚方法产生时间长度不同的锚段,设 $K = 4$,锚段在不同特征图 $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3$ 上的尺度为{8,12,16,20}、{12,14,16,18}、{10,12,14,16}。可以注意到特征图 $\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3$ 的时间维缩放比为 2,4,8,因此提案子网络中产生的所有锚段的时间尺度包含 16,24,32,40,48,56,64,72,80,96,112,128,覆盖了所有的行为片段时间长度。经初次预测后的锚段应用非极大值抑制方法进行挑选,获得 3 000 个用于训练的提案片段与 300 个用于测试的提案片段。不同时间长度的提案片段向各级特征图进行映射时,设定映射的控制参数 k_0 为 7.5。

训练时序提案网络时,需要为经非极大值抑制后得到的提案片段设定正/负标签。设定的方法遵循原则:如果提案片段与某些真实行为片段在时间维交并比(IoU)高于 0.7 或者某提案片段与某真实行为片段 IoU 在所有提案片段中最高,那么给该提案片段赋予正标签。而如果提案片段与所有的真实行为片段 IoU 低于 0.3,则给该提案片段标定负标签。既不满足正标签的条件,也不满足负标签的条件的提案片段将不会用于训练网络。同时,在实验时尝试通过随机采样的方式,将正负标签的提案片段比例维持在 1∶1。由于正标签的提案片段数量通常少于负标签的,实验时优先采样半个批大小数量的正标签的提案片段,如果正标签的提案片段数量不够,再补充负标签的提案片段。在实验中,将批大小设置为 128。训练分类和回归网络时,需要为每一个用于训练的提案片段指定行为的类别标签。如果一个提案片段与某实际行为片段的 IoU 在所有片段中最大,且 IoU 值高于 0.5,就给该提案片段指定实际行为相同的行为标签。如果提案片段与所有实际行为片段的 IoU 值都低于 0.5,则给该提案片段赋予负标签,即代表背景,不包含行为。训练分类网络和回归网络时,不限制正标签和负标签的比例。

3.2　在 THUMOS'14 上的测试

实验中使用 THUMOS'14 提供的整个验证集来训练本文网络(称作训练集)。训练集中的 200 个视频分为两部分,其中的 180 个视频用于网络优化,另外 20 个用于验证和超参数的调节。THUMOS'14 提供的测试集用于测试,其中共有 213 个视频。可以注意到大部分(99.5%)的行为片段持续时间都不超过 30 s,为了方便网络输入,实验时构造了许多的 buffer,通过在视频上滑动的方式,每 768 帧构造一个 buffer,由于视频帧率为 25 帧/s,每个 buffer 就包含 30.72 s 的视频片段,最后将这些 buffer 作为网络的输入。实验使用在 Sports-1M 上预训练的 C3D 模型参数来初始化特征金字塔的前 20 层(即原 C3D 网络的部分),其他层(包括横向连接层、平滑层和用于分类回归的全连接层)则使用标准初始化

方法。在训练模式下,固定网络前5层的权重,对其他层的权重参数以0.000 1的初始学习率进行优化。训练过程中,前3个世代(epoch)学习率保持不变,之后每3个世代学习率降为原来的1/10,权重衰减系数设为0.000 05,冲量设为0.9。网络进行端到端的训练,所有的网络模块都同时进行优化。使用单个TitanX GPU进行训练,耗时约60 h使得损失值达到稳定。对网络效果的评估指标采用不同IoU阈值下的全类平均准确率,即mAP@α, α是不同阈值。

表1展示了本文网络在THUMOS'14数据集上与其他方法的结果比较,以mAP在不同阈值下的值为评价指标(即IoU阈值从0.1～0.7,每次以0.1为间隔)。由于网络以RGB帧为输入,没有运用到光流特征,主要与同采用RGB帧输入的方法进行比较,部分较典型的同时使用了光流和RGB特征的方法也被列入。可以看出,在IoU阈值0.2以上时,本文模型准确率较高,在α=0.5时,模型的mAP达到了37.4%,超过2018年的最佳模型ETP(evolving temporal proposals)(Qiu等,2018)3.2%,与经典的两阶段方法R-C3D(Xu等,2017)相比,本文模型使准确率提升了8.5%。

表1　不同模型在THUMOS'14数据集上的检测结果

Table 1　Detection results of different models on THUMOS'14 dataset

/%

方法	IoU阈值						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
Buch等人(2017b)	–	–	37.8	–	23.0	–	–
Shou等人(2017)	–	–	40.1	29.4	23.3	13.1	7.9
Buch等人(2017a)	–	–	45.7	–	29.2	–	9.6
Shou等人(2016)	47.7	43.5	36.3	28.7	19.0	10.3	5.3
Yuan等人(2017)	51.0	45.2	36.5	27.8	17.8	–	–
Gao等人(2017a)	60.1	56.7	50.1	41.3	31.0	19.1	9.9
Gao等人(2017b)	54.0	50.9	44.1	34.9	25.6	–	–
Qiu等人(2018)	–	–	48.2	42.4	34.2	23.4	13.9
Xu等人(2017)	54.5	51.5	44.8	35.6	28.9	–	–
Zhao等人(2017)	**66.0**	**59.4**	51.9	41.0	29.8	–	–
3D-FPCN(本文)	56.4	55.9	**52.7**	**47.5**	**37.4**	**28.7**	**16.7**

注:加粗字体为每列最优值,"–"指未提供此项数据。

表2展示了IoU阈值为0.5时,模型在THUMOS'14上对不同行为进行预测的平均准确率(AP),在对大部分行为进行预测时,3D-FPCN准确率都高过现有模型。在阈值低于0.2时,结构分割网络(structured segment networks, SSN)(Zhao等,2017)方法表现较好,但阈值高于0.2时表现不足,这说明SSN的预测结果与实际行为片段交并比低的数量较多,这可能与SSN方法做的冗余预测有关,有较多数量类别预测正确但边界回归不准确的预测行为片段。比较对不同行为预测时,准确率的提升值,可以发现本文模型对于持续时间较短的行

表2　不同模型在THUMOS'14数据集上对各类行为片段的检测

Table 2　Detection results of various behavior segments on THUMOS'14 dataset(AP, α=0.5)

/%

视频	R-C3D	ETP	SCNN	3D-FPCN
棒球击球	**26.1**	22.5	14.9	22.2
篮球扣篮	54.0	30.3	20.1	**64.0**
台球	8.3	8.1	7.6	**9.2**
挺举	27.9	40.9	24.8	**50.6**
悬崖跳水	49.2	16.7	27.5	**59.5**
保龄球	30.6	16.3	15.7	**34.6**
板球射击	10.9	7.2	13.8	**30.2**
潜水	26.2	**50.9**	17.6	43.4
飞盘捕捉	**20.1**	2.3	15.3	5.3
高尔夫挥杆	16.1	**44.4**	18.2	27.6
掷锤	43.2	**71.7**	19.1	47.9
跳高	30.9	**51.2**	20.0	43.4
标枪投掷	47.0	47.3	18.2	**49.8**
跳远	57.4	**81.9**	34.8	64.1
撑竿跳高	42.7	56.5	32.1	**70.8**
铅球	19.4	**32.0**	12.1	21.6
足球点球	15.8	19.7	19.2	**28.6**
网球挥拍	16.6	**29.1**	19.3	25.9
掷铁饼	29.2	**39.1**	24.4	32.9
排球扣球	5.6	14.0	4.6	**16.9**
mAP@0.5	28.9	34.2	19.0	**37.4**

注:加粗字体为每行最优值。

为片段进行预测时，准确率提升都较高，例如篮球、挺举和高尔夫挥杆等。表 2 中的结果表明了特征金字塔结构能够提升对持续时间较短的人类行为片段的检测准确率。

3.3　消融性实验

为了了解特征金字塔在网络中发挥的作用，进行消融性实验。首先在 THUMOS'14 上评估了没有金字塔层次结构的网络的性能。此外，还与移除掉金字塔 RoI 池化层的网络进行比较。图 3 展示了比较有代表性的实验结果。

实验 1，首先需要评估不带金字塔层次结构的网络性能。实验中将锚段生成的尺度设为 2，4，5，6，8，9，10，12，14，16，只使用 C3D 生成的一幅基础特征图进行 RoI 池化和预测。由于基础特征图相对于输入帧的缩放比为 8，将池化因子也设置为 8，这样锚段可以覆盖从 16～128 的时序范围，即 0.64～5.12 s。实验 2，在第 1 个子网络使用特征金字塔结构产生多级特征图，用多级特征图预测提案片段，但在进行 RoI 池化时，只使用基础特征图。实验的结果如表 3 所示。

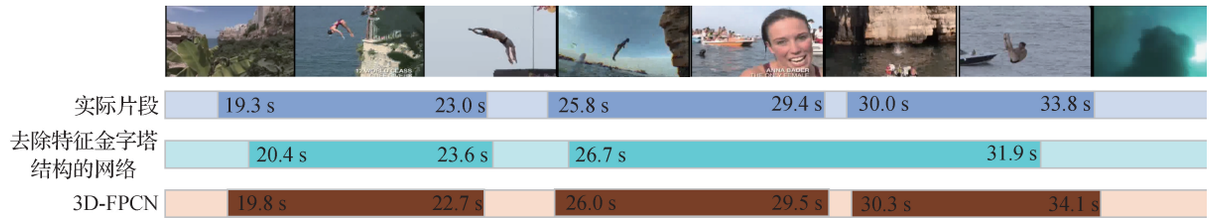


图 3　不同模型对某一视频悬崖跳水部分的检测结果，图片为视频中按顺序截取的示意图

Fig. 3　The detection results of different models on the cliff diving of a certain video, the picture is a schematic diagram of the video taken in sequence

表 3　消融性实验结果

Table 3　Results of the ablation test

/%

	提案金字塔	池化金字塔	IoU 阈值				
			0.1	0.2	0.3	0.4	0.5
移除特征金字塔	–	–	53.7	53.5	49.7	43.8	35.6
提案阶段使用特征金字塔	yes	–	54.2	53.6	50.2	44.5	35.8
3D-FPCN	yes	yes	**56.4**	**55.9**	**52.7**	**47.5**	**37.4**

注：加粗字体为每列最优值，“–”为未含此结构，“yes”为含此结构。

本文模型与移除特征金字塔层的网络相比较，mAP@0.1～0.5 提升了 1.8%～3.8%，表明特征金字塔结构有助于网络监测效果的提高。值得注意的是，特征金字塔结构在产生提案片段阶段和时序 RoI 池化阶段都发挥了作用，为理清特征金字塔结构在不同阶段的效果，需要分析消融性实验 2 的结果。可以看出，消融性实验 2 加上了特征金字塔结构，但特征金字塔结构只在第 1 阶段产生提案片段时发挥作用，在进行时序 RoI 池化时只使用了基础特征图。网络的测试结果在 mAP@0.1～0.5 相对于完全不使用特征金字塔结构的网络有所提升，但提升不明显。

特征金字塔结构在两个阶段都起作用的网络相对于只在第 1 阶段起作用的网络，mAP@0.1～0.5 平均提升了 2% 以上，说明特征金字塔结构主要在时序 RoI 池化阶段发挥了作用。消融性实验也从侧面说明了两阶段的检测网络一般要优于单一阶段的分类和回归网络。

4　结　论

本文提出了一个新的两阶段的时序行为识别网络 3D-FPCN。网络由 3 个子网组成，旨在提高网络对不同持续时长的行为片段的检测能力。为了实现这个目标，引入了 3 维特征金字塔结构，构建了 3 维特征金字塔特征提取网络，其中包含自上而下的通

道和横向连接通道。自上而下的通道可以将高抽象度的语义信息向下传递,而横向连接的通道则将本级别分辨率的语义信息进行传递,从而实现了高抽象度语义信息与高分辨率语义信息相融合的目的,产生抽象度和分辨率不同的多级别特征图。其后的两个子网络则有效地利用了多级别不同抽象度和分辨率的语义信息。在提案子网络中结合锚机制,使不同时间长度的锚段在不同级别的特征图上具有不同感受野,增强了对锚段进行初次预测的能力。在分类和回归阶子网络中则对不同时间长度的提案片段使用不同抽象程度和分辨率的特征图进行 RoI 池化,平衡了分类和回归准确度的需求。

通过网络在 THUMOS'14 数据集上的实验,展示了 3D-FPCN 的优越性。其与当前一些使用 RGB 帧作为输入的方法比较,全类平均正确率超过了多种经典方法,证明了特征金字塔结构用于两阶段网络进行时序行为识别任务的有效性。在比较不同类别的人类行为片段检测结果时,3D-FPCN 对持续时间较短的行为片段检测准确率相较于当前其他方法有显著提高,说明针对性地利用不同分辨率的特征图,能够应对持续时间较短的行为片段难以准确检测的问题。最后,通过消融性实验研究了自上而下的特征金字塔结构在网络中发挥的影响,证明了其对人类行为特征提取的有效性。针对性地利用多级特征图在网络的提案过程和池化过程均有提升预测准确率的作用,而探究发现多级特征图主要发挥作用在网络的 RoI 池化阶段。

因视频信息较复杂、行为模糊等原因,目前各方法在时序行为识别任务表现均不尽人意,3D-FPCN 虽然取得了对短时行为片段检测准确率的提升,但其整体的检测准确率还较低,只有 37.8%,无法满足实际应用的需求。下一步将使用 RGB 与光流特征相结合的方法效果对网络结构进行改进:一是加入能够利用视频背景信息的网络模块,二是加入对光流特征的使用,同时利用 RGB 和光流特征进行预测,以求进一步提升网络在时序行为识别任务上的效果。

参考文献(References)

Buch S, Escorcia V, Ghanem B and Niebles J C. 2017a. End-to-end, single-stream temporal action detection in untrimmed videos//Proceedings of the British Machine Vision Conference. London, UK: BMVA Press: #7 [DOI: 10.5244/c.31.93]

Buch S, Escorcia V, Shen C Q, Ghanem B and Niebles J C. 2017b. SST: single-stream temporal action proposals//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 2911-2920 [DOI: 10.1109/CVPR.2017.675]

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

Chao Y W, Vijayanarasimhan S, Seybold B, Ross D A, Deng J and Sukthankar R. 2018. Rethinking the faster R-CNN architecture for temporal action localization//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1130-1139 [DOI: 10.1109/CVPR.2018.00124]

Gaidon A, Harchaoui Z and Schmid C. 2011. Actom sequence models for efficient action detection//Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado, USA: Springer: IEEE: 3201-3208 [DOI: 10.1109/CVPR.2011.5995646]

Gao J Y, Yang Z H and Nevatia R. 2017a. Cascaded boundary regression for temporal action detection//Proceedings of the British Machine Vision Conference. London, UK: BMVA Press: 52.1-52.11 [DOI: 10.5244/c.31.52]

Gao J Y, Yang Z H, Sun C, Chen K and Nevatia R. 2017b. TURN TAP: temporal unit regression network for temporal action proposals//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3628-3636 [DOI: 10.1109/ICCV.2017.392]

Jiang Y G, Liu J G, Zamir A R, Toderici G, Laptev I, Shah M and Sukthankar R. 2014. Thumos challenge: action recognition with a large number of classes [EB/OL]. [2020-10-29]. http://crcv.ucf.edu/THUMOS14/

Lin T W, Liu X, Li X, Ding E R and Wen S L. 2019. BMN: boundary-matching network for temporal action proposal generation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea(South): IEEE: 3889-3898 [DOI: 10.1109/ICCV.2019.00399]

Long F C, Yao T, Qiu Z F, Tian X M, Luo J B and Mei T. 2019. Gaussian temporal awareness networks for action localization//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 344-353 [DOI: 10.1109/CVPR.2019.00043]

Qiu H N, Zheng Y B, Ye H, Lu Y, Wang F and He L. 2018. Precise temporal action localization by evolving temporal proposals//Proceedings of 2018 ACM on International Conference on Multimedia Retrieval. Yokohama, Japan: ACM: 388-396 [DOI: 10.1145/3206025.3206029]

Shou Z, Chan J, Zareian A, Miyazawa K and Chang S F. 2017. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 5734-5743 [DOI: 10.1109/CVPR.2017.155]

Shou Z, Wang D A and Chang S F. 2016. Temporal action localization in untrimmed videos via multi-stage CNNs//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 1049-1058 [DOI: 10.1109/CVPR.2016.119]

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

Xu H J, Das A and Saenko K. 2017. R-C3D: region convolutional 3d network for temporal activity detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5783-5792 [DOI: 10.1109/ICCV.2017.617]

Yuan Z H, Stroud J C, Lu T and Deng J. 2017. Temporal action localization by structured maximal sums//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 3684-3692 [DOI: 10.1109/CVPR.2017.342]

Zeng R H, Huang W B, Gan C, Tan M K, Rong Y, Zhao P L and Huang J Z. 2019. Graph convolutional networks for temporal action localization//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7094-7103 [DOI: 10.1109/ICCV.2019.00719]

Zhao Y, Xiong Y J, Wang L M, Wu Z R, Tang X O and Lin D H. 2017. Temporal action detection with structured segment networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2914-2923 [DOI: 10.1109/ICCV.2017.317]

作者简介

何嘉宇，1996 年生，男，硕士研究生，主要研究方向为深度学习、计算机视觉、虚拟现实技术。
E-mail: jyu.he@qq.com

李国辉，通信作者，男，教授，主要研究方向为计算机视觉、信息系统工程、数据挖掘、虚拟现实技术。
E-mail: guohli@nudt.edu.cn

雷军，男，讲师，主要研究方向为计算机视觉、深度学习、数据挖掘、虚拟现实技术。E-mail: leijun1987@nudt.edu.cn