第 45 卷 第 12 期 计 算 机 工 程 2019 年 12 月
Vol. 45 No. 12 Computer Engineering December 2019

·多媒体技术及应用· 文章编号：1000-3428(2019)12-0257-06 文献标志码：A 中图分类号：TP753

基于行为主体检测的视频行为快速检测

张杰豪，陈华杰，姚勤炜，侯新雨

（杭州电子科技大学 自动化学院，杭州 310018）

摘 要：现有视频行为检测方法在生成候选区域时采用滑窗操作，处理长视频速度较慢。针对该问题，通过对静态行为主体进行定位，提出一种快速检测方法。将长视频分割为若干个视频单元，在每个单元的第 1 帧中运用 Fast R-CNN 算法进行行为主体检测，对检测到行为主体的单元划定时间区域生成行为发生候选区域，以减少行为检测网络的输入数据。在此基础上，采用 3D 卷积神经网络判别候选区域类别，对行为类区域进行边界回归，得到准确的行为时间轴定位。实验结果表明，该方法检测速度较 TURN 方法提升 2 倍以上，其 mAP 指标只降低 0.7% 。

关键词：行为检测；行为主体检测；边界回归；3D 卷积神经网络；视频单元

开放科学（资源服务）标志码（OSID）：

中文引用格式：张杰豪，陈华杰，姚勤炜，等. 基于行为主体检测的视频行为快速检测[J]. 计算机工程，2019，45(12)：257-262.

英文引用格式：ZHANG Jiehao，CHEN Huajie，YAO Qinwei，et al. Fast video action detection based on action subject detection[J]. Computer Engineering，2019，45(12)：257-262.

Fast Video Action Detection Based on Action Subject Detection

ZHANG Jiehao，CHEN Huajie，YAO Qinwei，HOU Xinyu

（School of Artificial Intelligence，Hangzhou Dianzi University，Hangzhou 310018，China）

【Abstract】The existing video action detection methods adopt the sliding window operation when generating candidate regions，which process long video speeds slowly. Aiming at this problem，a fast detection method is proposed by detecting the static action subject. First，a long video is divided into several units，and the Fast R-CNN algorithm is adopted to detect the action subject in the first frame of each unit. Then，time zones are defined in the units with action subject to generate action occurrence candidate regions，so as to reduce the input data of the action detection network. On this basis，this paper uses 3D Convolutional Neural Network（CNN）to discriminate the classification of candidate regions. Finally，the boundary regression is performed on action regions，thus obtaining an accurate action time axis positioning. Experimental results show that the detection speed of the proposed method is 2 times higher than that of the TURN method，with an mAP indicator decrease by merely 0.7% .

【Key words】action detection；action subject detection；boundary regression；3D Convolutional Neural Network（CNN）；video unit

DOI：10.19678/j.issn.1000-3428.0053184

0 概述

行为检测与识别在监控系统、视频分析等方面具有巨大的应用潜力，受到研究人员的关注[1-2]。目前，基于时序结构分析的行为识别模型已较为成熟，但将其用于行为检测任务时却不能保证时效性和通用性。

行为检测是从一段未经裁剪甚至冗长的视频中检测出包含人类行为的片段，不仅需要辨别行为类与背景类，同时也要预测行为的开始与结束时间点，即生成时间轴定位。因此，可将该任务分为 2 个阶段：首先从完整的长视频中生成可能包含行为的候选区域，然后对此段候选区域进行背景/行为二分类，输出精准的时间轴定位，进而完成行为检测。

现有的行为检测方法主要包括以 S-CNN 模型[3]为代表的方法，即通过多阶段的 3D 卷积神经网络[4-5]建模将独立片段检测与时序结构相结合，以及直接基于循环神经网络（Recurrent Neural Network，

作者简介：张杰豪（1994—），男，硕士研究生，主研方向为视频检测、模式识别、机器学习；陈华杰，教授、博士；姚勤炜、侯新雨，硕士研究生。

收稿日期：2018-11-20 **修回日期：**2018-12-27 **E-mail：**zzqyjr@foxmail.com

RNN)[6-7]的方法,如 DAPs[8]采用长短期记忆(Long Short-Term Memory,LSTM)[9]网络对视频流进行编码以生成候选时间区域[10],或通过强化学习训练得到基于 RNN 的代理,实现端对端训练并直接生成行为预测结果[11]。

现有行为检测方法的效率难以达到实际应用的要求,主要问题为视频数据的时长过长导致了长期依赖从而影响检测性能,以及不能有效地生成候选区域,在样本行为片段不密集的场景下,输入数据量过大导致检测速度过慢。

通过提高检测框架分析时序结构即分辨行为开始(结束)阶段的能力,可以生成精确的候选区域以减小输入数据量,加快检测速度。对常见的监控视频进行特性分析可以发现,长视频中通常包含大量无行为数据,行为片段占比很小(应用范围很广的 THUMOS 14 测试集[12]中时长占比 10% ~20%)。若能快速有效地初步筛选出稀疏行为片段,将极大地缩短整个行为检测任务所需时间。此外,行为片段时长通常远大于 0.5 s,这一时长也可避免多数漏检。

对目标进行静态检测的技术如 Fast R-CNN[13]等已经具有成熟的框架并且检测效果较好,因此,本文在其基础上提出一个快速行为检测方法,在行为/背景二分类操作之前,对长视频稀疏采样进行行为主体静态检测,筛选出有限的候选区域,从而解决数据量过大的问题,加快检测速度。

1 基于行为主体检测的时序行为检测

1.1 本文检测框架

本文提出的检测框架整体结构如图 1 所示,其检测流程为:将整段长视频分割为若干个视频单元,从每个视频单元抽取第 1 幅图像帧,对其采用 Fast R-CNN 进行行为主体检测,将检测到行为主体的单元划定时间区域生成行为发生候选区域;运用时序行为检测框架对候选区域进行行为/背景分类判别是否为行为;进行边界回归,得到准确的行为时间轴定位。

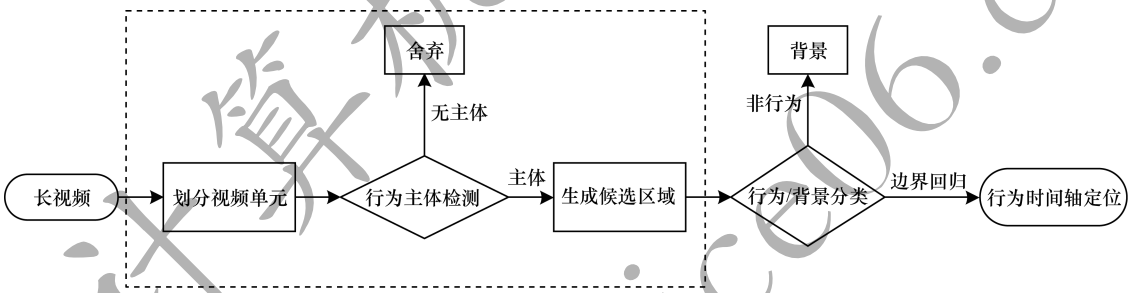

图 1 基于行为主体检测的行为快速检测框架

本文方法继承了已有方法的优点:

1)聚焦于候选区域生成。S-CNN 模型将候选区域生成阶段独立出来,本文方法也采用同样的分阶段架构。

2)对检测出行为的时间区域微调。TURN TAP 模型[15]通过对时序边界进行微调以提高检测精度。本文方法采用与 R-CNN[16]类似的边界回归方法提高最终输出的行为片段时间轴定位精度。

同时,本文相对已有方法做了如下改进:

1)已有方法的候选区域生成均采用滑窗操作,对包含重叠部分的候选片段反复提取特征,从而计算量增加,在处理长视频时极其耗费时间。因此,本文方法采用稀疏采样进行静态行为主体检测的方式,更为有效率地生成候选区域。

2)本文方法将一个多任务损失函数同时用于网络训练分类以及边界回归,以提高最终输出的行为片段时间轴定位精度。

1.2 行为主体检测

行为主体检测分为单元分割、行为主体检测以及补充上下文单元组成单元块,如图 2 所示。

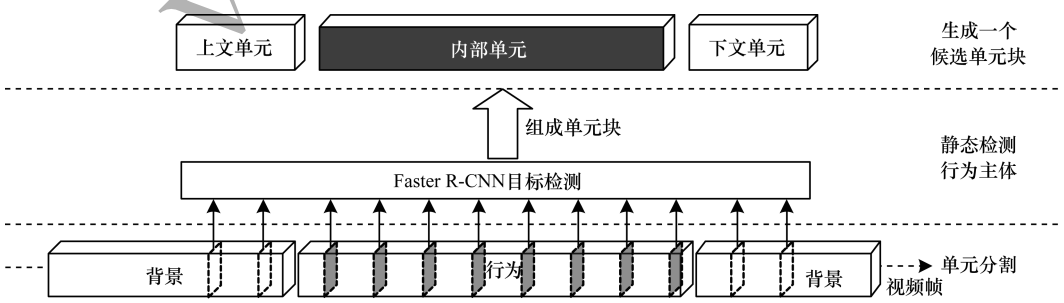

图 2 行为主体检测流程

一段时长为 t 的视频 V 包含 T 帧,则 $V=\{t_i\}_1^T$,视频帧率 $w=T/t$。将视频 V 分割为 T/n_u 个连续的视频单元,其中,n_u 表示每个单元的帧数。一个单元可以表示为 $u=\{t_i\}_{s_f}^{s_f+n_u-1}$,其中,$s_f$ 表示起始帧,s_f+n_u-1 表示结束帧,单元与单元之间无重叠部分。

将每一个单元的起始帧 $\{s_f\}_{u_1}^{u_{T/n_u}}$ 输入至文献[17]所提网络进行目标检测,输出结果为帧图像中各目标及其分类概率,对其中检测为行为主体概率大于 80% 的源视频单元生成一个单元集 $\{u_i\}$,再从中筛选出时间轴连续的视频单元组成候选行为单元集 $\{u_j\}$。对候选行为单元集 $\{u_j\}$ 中某个连续单元的组合定义为单元块 $c=\{u_j\}_{s_u}^{s_u+n_u\cdot(n_c-1)}$,其中,$s_u$ 表示起始单元的初始帧,n_c 表示单元块 c 中的单元数量。$e_u=s_u+n_u\cdot(n_c-1)$ 则是结束单元的初始帧,$\{u_j\}_{s_u}^{e_u}$ 称为 c 的内部单元。

针对行为主体检测时可能出现漏检的情况,本文设计了上下文单元。参考内部单元,类似可定义 c 的上下文单元为 $\{u_j\}_{s_u-n_{ctx}}^{s_u}$ 和 $\{u_j\}_{e_u}^{e_u+n_{ctx}}$,其中 n_{ctx} 表示作为上下文单元帧数。

1.3　行为/背景分类及边界回归

行为/背景分类及边界回归流程如图 3 所示。

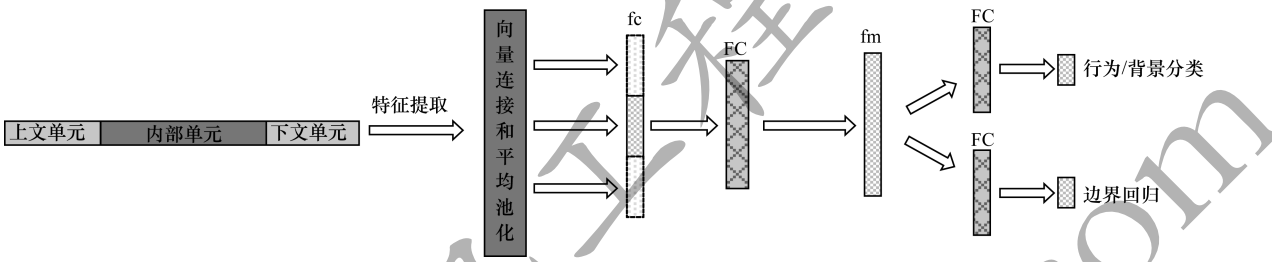

图 3　行为/背景分类及边界回归流程

单元块中的内部特征和上下文特征分别由函数 P 提取并池化生成最终特征池。c 的最终特征 f_c 与上下文特征、内部特征连接。

$$f_c=P(\{u_j\}_{s_u-n_{ctx}}^{s_u})\,\|\,P(\{u_j\}_{s_u}^{e_u})\,\|\,P(\{u_j\}_{e_u}^{e_u+n_{ctx}})$$

其中,$\|$ 表示对 P 的向量连接和平均池化。

设计一个单元回归模块,输入为单元块 c 的最终特征 f_c,2 个相关输出层分别输出输入片段是行为的置信分数和时序边界回归补偿。回归补偿 o_s 计算公式为:

$$o_s=s_u-s_{gt},o_e=e_u-e_{gt} \tag{1}$$

其中,s_u、e_u 表示 c 的起始单元初始帧和结束单元初始帧,s_{gt}、e_{gt} 是真实边界的起始帧和结束帧。

单元回归模块有 2 个方面的优势:

1)进行单元层面的边界回归而非帧层面的边界回归。以包含 n_u 帧的行为单元为单位提取的特征相比以帧为单位包含更多的时序信息。

2)由于单元划分以固定 n_u 帧为标准,可能会出现丢失一部分包含行为主体片段的情况,单元回归可以对此进行一定程度的补偿。

在训练行为/背景分类网络时,首先给每个单元块打上分类标签(是否为行为)。正样本标签的单元块需满足以下条件之一:

1)单元块与某个真实行为片段重叠。

2)单元块与某个真实行为片段的时序交除并(temporal Intersection over Union,tIoU)[18]大于 0.5。

因此,一个真实行为片段可能给多个单元块打上正标签。若单元块与任意真实行为片段的 tIoU 等于 0,则其为负样本。定义一个多任务损失函数 L 用于训练分类以及边界回归。

$$L=L_{\text{cls}}+\lambda L_{\text{reg}} \tag{2}$$

其中,L_{cls} 表示动作/背景分类的损失,是一个标准 Softmax 损失函数[19],L_{reg} 是时序边界回归的损失,λ 是一个超参数。回归损失 L_{reg} 采用 $L1$ 距离范数,计算公式如下:

$$L_{\text{reg}}=\frac{1}{N_{\text{pos}}}\sum_{i=1}^{N}l_i^*\,|(o_{s,i}-o_{s,i}^*)+(o_{e,i}-o_{e,i}^*)| \tag{3}$$

其中,l_i^* 表示标签,1 为正样本,0 为负样本,N_{pos} 表示正样本的数量。式(3)只对正样本计算回归损失。

2　实验与结果分析

图 4 给出本文方法应用于视频行为检测的一个范例,该样本视频片段标注的真实行为片段数据为 101.6 s ~ 106.5 s。本文方法首先进行行为主体检测,在 102.00 s、102.50 s、103.00 s、103.50 s、104.00 s、104.50 s 和 105.00 s 的视频帧上检测到行为主体,生成候选区域 101.0 s ~ 106.5 s;然后进行行为/背景分类得到 101.8 s ~ 105.7 s 属于行为片段;最后进行边界回归输出最终时间轴定位 101.6 s ~ 106.3 s。此预测结果与真实数据的 tIoU 大于 0.5,因此,对范例的预测正确。

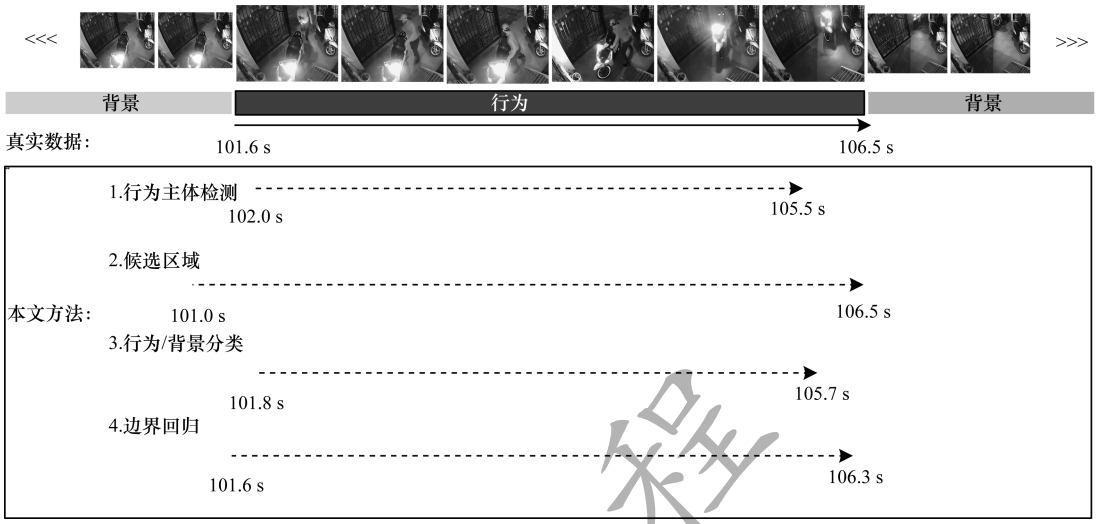

图 4 视频行为检测范例

2.1 数据集与评价指标

本文实验选用 ActivityNet[20] 和 THUMOS 14[12] 中的时序行为检测任务数据集。THUMOS 14 视频总时长超过 20 h,包含训练集、验证集、背景集和测试集。其中:训练集是 UCF101 数据集[21] 的一个子集,包含 20 类行为且都为修剪过的视频,不含背景;验证集含 20 类共 200 个视频,同样可作为训练集,标签包含视频中发生的所有行为的时间信息;测试集含 1 574 段未修剪视频,视频中包含一种或多种行为,用于测试。ActivityNet v1.3 版本数据集包含超过 648 h 的未修剪视频,共约 20 000 个视频。

根据 THUMOS 14 官方给出的评价工具集,使用插值平均精度(Interpolated Average Precision,AP)及其均值 mAP 作为性能评价指标。对于行为类 c,$AP(c)$ 由下式计算得到:

$$AP(c)=\frac{\sum_{k=1}^{n}(P(k)\times rel(k))}{\sum_{k=1}^{n}rel(k)} \tag{4}$$

其中,n 为视频总数,$P(k)$ 为视频 k 的检测精度,$rel(k)$ 是指标函数,视频 k 为正样本则为 1,反之为 0。

平均精度均值由下式得到:

$$mAP=\frac{1}{C}\sum_{c=1}^{C}AP(c) \tag{5}$$

其中,C 为测试数据的总类数,本文为 20。

重叠度 tIoU 由下式计算得到:

$$o=\frac{R_p\cap R_{gt}}{R_p\cup R_{gt}} \tag{6}$$

其中,R_p 和 R_{gt} 分别表示预测时间范围与真实时间范围。当 $o\geqslant 0.5$ 时,可认为行为检测准确。

根据 ActivityNet 官方给出的评价工具集,使用平均召回率与平均预测数量(Average Recall vs. Average Number of Proposals per Video,AR-AN)曲线作为性能评价指标。重叠度阈值同样取 $o\geqslant 0.5$。

2.2 实验设置

本文实验对比了采用不同特征时本文方法的性能:

1)C3D 特征,模型用 Sports1m 数据集预训练,将整个单元送入 C3D,并提取 fc6 特征。

2)Dense Flow 特征(下文简称 FL 特征),在单元中间取连续的 6 帧并计算对应的光流,将光流输入 BN-Inception(用 ActivityNet v1.3 预训练)中,取 global_pool 特征。

在训练过程中,每个 mini-batch 中负样本与正样本比例设置为 10∶1,学习率设置为 0.05,batch size 设置为 128 并使用 Adam 优化器。单元帧数 $n_u=w/2$,上下文单元帧数 $n_{ctx}=w$,中间层 f_m 维度为 1 000,超参数 $\lambda=2$。

2.3 实验结果

本文实验所用检测数据是从 THUMOS 14 和 ActivityNet v1.3 中的验证集中随机抽取的 100 段视频(各方法对比时采用的视频一致),实验 GPU 为 7TFLOPs 的单卡 TITAN X。在行为主体检测过程中预筛选阶段采用各模型耗费时间如表 1 所示。从表 1 可以看出,采用使用 COCO 数据集预训练的基于 Inception v2 的 Faster R-CNN 网络取得了检测精度与耗时的平衡,因此,本文实验采用这一网络进行行为主体检测。

表 1 行为主体检测各模型效率对比结果

模型	耗时/ms	COCO mAP/%	mAP 耗时比
faster_rcnn_inception_v2_coco	12	28	2.33
faster_rcnn_resnet50_coco	19	30	1.57
rfcn_resnet101_coco	20	30	1.50
faster_rcnn_resnet101_coco	23	32	1.60
faster_rcnn_inception_resnet_v2	132	37	0.28

本文方法采用 2.2 节中 2 种不同特征进行实验,其结果如表 2 所示。

表 2　不同特征实验结果对比

特征	平均召回率/%	运行速度/(frame · s^{-1})
C3D 特征	41.8	261.3
FL 特征	39.6	149.4

从表 2 可以看出,采用 C3D 特征在效率上有较大的优势,因此下文实验均采用 C3D 特征进行。

本文方法与其他方法在 THUMOS 14 数据集中的 AR-AN 曲线如图 5 所示。表 3 给出各种方法的 mAP、AR@ AN = 200、运行速度指标对比结果。

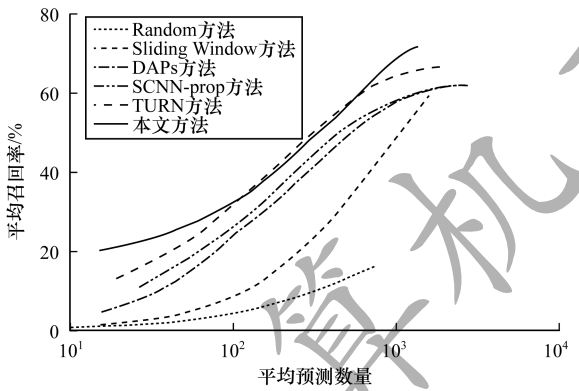

图 5　THUMOS 14 数据集 AR-AN 曲线

表 3　THUMOS 14 数据集对比实验结果

方法	mAP/%	平均召回率/%	运行速度/(frame · s^{-1})
本文方法	24.9	41.8	261.3
Sparse-prop 方法	15.3	33.3	6.5
DAPs 方法	16.3	35.7	85.0
SCNN-prop 方法	19.0	38.3	38.2
TURN 方法	25.6	43.0	82.3

从表 3 可以看出,在对比方法中,TURN 方法的指标最佳。因此,在 ActivityNet v1.3 数据集中将本文方法与 TURN 方法进行对比,结果如表 4 所示。

表 4　ActivityNet 数据集对比实验结果

方法	平均召回率/%	运行速度/(frame · s^{-1})
本文方法	39.8	243.4
TURN 方法	41.2	82.7

从上述实验结果可以看出,本文方法的检测精度与召回率均较优。与 TURN 方法相比,本文方法在检测精度相差不大的情况下,其检测速度具有明显优势。因此,本文方法对较长视频分析的速度快,在实时性敏感的应用场景上将更具应用价值。

为研究单元帧数 n_u 大小对本文方法检测效果与速度的影响,在 THUMOS 14 数据集上以 2.2 节中规定单元时长为 0.5 s 即 $n_u = w/2$ 为测试基准进行实验,结果如表 5 所示。

表 5　不同单元帧数 n_u 实验结果

n_u	平均召回率/%	运行速度/(frame · s^{-1})
$w/2$	41.8	261.3
$(w/2)+5$	41.4	283.4
$(w/2)-5$	42.1	230.7

从表 5 可以看出,单元帧数 n_u 大小对平均召回率影响较小,而对运行速度影响相对较大,在实际应用时可以根据具体场景需求进行调整。

此外,为研究数据集中行为片段占比大小对本文方法检测效果与素的的影响,本文在 THUMOS 14 数据集中进行实验测试,结果如表 6 所示。

表 6　不同行为片段占比实验结果

行为片段占比/%	平均召回率/%	运行速度/(frame · s^{-1})
14	41.8	261.3
10	41.7	311.7
20	41.1	196.4

从表 6 可以看出,行为片段占比大小对检测精度影响不大,而对运行速度影响相对较大。在行为占比已达 20% 的应用场景中,本文方法依然可以快速完成行为检测。

综上实验结果可知,本文提出的行为检测方法应用于包含行为片段的长视频时,其检测精度较高,适用场景广,对于不同行为片段占比的长视频均有明显的速度提升。

3　结束语

本文提出一种快速视频行为检测方法,通过提前定位行为主体,减少输入行为检测网络的候选数据,同时结合 3D 卷积神经网络得到最终检测结果。实验结果表明,本文方法的检测精度较高且适用场景广,能明显提高长视频的检测速度,其通过进行快速行为检测,可以得到充足的时间进行后续的行为识别工作,从而达到实时检测识别的目的。下一步将综合时序行为检测、异常行为检测和行为识别等任务,设计一个完整的检测系统。

参考文献

[1] GUPTA A, SRINIVASAN P, SHI Jianbo, et al. Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos[C]//Proceedings of CVPR'09. Washington D. C., USA: IEEE Press, 2009: 2012-2019.

[2] AGGARWAL J K, RYOO M S. Human activity analysis: a review[J]. ACM Computing Surveys, 2011, 43(3): 16.

[3] SHOU Zheng, WANG Dongang, CHANG Shih-Fu. Temporal action localization in untrimmed videos via multi-stage CNNs[C]//Proceedings of CVPR'16. Washington D. C., USA: IEEE Press, 2016: 1049-1058.

[4] PINEDA F J. Generalization of back-propagation to recurrent neural networks[J]. Physical Review Letters, 1987, 59(19): 2229-2232.

[5] WILLIAMS R J, ZIPSER D. A learning algorithm for continually running fully recurrent neural networks[J]. Neural Computation, 1989, 1(2): 270-280.

[6] JI Shuiwang, XU Wei, YANG Ming, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

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

[8] ESCORCIA V, HEILBRON F C, NIEBLES J C, et al. Daps: deep action proposals for action understanding[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 768-784.

[9] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[10] KRISHNA R, HATA K, REN F, et al. Dense-captioning events in videos[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 706-715.

[11] YEUNG S, RUSSAKOVSKY O, MORI G, et al. End-to-end learning of action detection from frame glimpses in videos[C]//Proceedings of CVPR'16. Washington D. C., USA: IEEE Press, 2016: 2678-2687.

[12] JIANG Yugang, LIU Jingen, ZAMIR A R, et al. THUMOS challenge: action recognition with a large number of classes[EB/OL]. [2018-11-01]. https://www.crcv.ucf.edu/THUMOS14/.

[13] GIRSHICK R. Fast R-CNN[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 1440-1448.

[14] LIN Fengxiao, CHEN Huajie, YAO Qinwei, et al. Target fast detection algorithm based on hybrid structure convolutional neural network[J]. Computer Engineering, 2018, 44(12): 222-227. (in Chinese)
林封笑, 陈华杰, 姚勤炜, 等. 基于混合结构卷积神经网络的目标快速检测算法[J]. 计算机工程, 2018, 44(12): 228-233.

[15] GAO Jiyang, YANG Zhenheng, CHEN Kan, et al. TURN TAP: temporal unit regression network for temporal action proposals[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 3628-3636.

[16] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich featzure hierarchies for accurate object detection and semantic segmentation[C]//Proceedings of IEEE CVPR'14. Washington D. C., USA: IEEE Press, 2014: 580-587.

[17] REN Shaoqing, HE Kaiming, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of Advances in Neural Information Processing Systems. [S. l.]: Neural Information Processing Systems, Inc., 2015: 91-99.

[18] NOWOZIN S. Optimal decisions from probabilistic models: the intersection-over-union case[C]//Proceedings of CVPR'14. Washington D. C., USA: IEEE Press, 2014: 548-555.

[19] HINTON G E, SALAKHUTDINOV R R. Replicated softmax: an undirected topic model[C]//Proceedings of Advances in Neural Information Processing Systems. [S. l.]: Neural Information Processing Systems, Inc., 2009: 1607-1614.

[20] HEILBRON F C, ESCORCIA V, GHANEM B, et al. Activitynet: a large-scale video benchmark for human activity understanding[C]//Proceedings of CVPR'15. Washington D. C., USA: IEEE Press, 2015: 961-970.

[21] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[EB/OL]. [2018-11-01]. http://crcv.ucf.edu/data/UCF101.php.

编辑 刘盛龄