计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(7):234−239 [doi: 10.15888/j.cnki.csa.007006]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于双流卷积神经网络的人体行为识别方法①

刘　云, 张　堃, 王传旭

(青岛科技大学 信息科学技术学院, 青岛 266000)
通讯作者: 张　堃, E-mail: oceanofzz@163.com

摘　要: 时序行为检测是指在一段未分割的长视频中, 检测出其中包含的若干行为片段的起止时间和类别. 针对该项任务, 提出基于双流卷积神经网络的行为检测模型. 首先使用双流卷积神经网络提取视频的特征序列, 然后使用TAG (Temporal Actionness Grouping) 生成行为提议, 为了构建高质量的行为提议, 将行为提议送入边界回归网络中修正边界, 使之更为贴近真实数据, 再将行为提议扩展为含有上下文信息的三段式特征设计, 最后使用多层感知机对行为进行识别. 实验结果表明, 本算法在 THUMOS 2014 数据集和 ActivityNet v1.3 数据集获得较好的识别率.

关键词: 行为识别; 双流卷积神经网络; 深度学习; 时序行为检测

引用格式: 刘云,张堃,王传旭.基于双流卷积神经网络的人体行为识别方法.计算机系统应用,2019,28(7):234–239. http://www.c-s-a.org.cn/1003-3254/7006.html

Human Action Recognition Algorithm Based on Two-Stream Convolutional Networks

LIU Yun, ZHANG Kun, WANG Chuan-Xu

(Information Science and Technology Academy, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Given a long, untrimmed video consisting of multiple action instances and complex background contents, temporal action detection needs not only to recognize their action categories, but also to localize the start time and end time of each instance. To this end, a temporal action detection network based on two-stream convolutional networks is proposed. First, the two-stream convolutional networks is used to extract the feature sequence of the video, and then TAG (Temporal Actionness Grouping) is used to generate the proposal. In order to construct high-quality proposals, the proposal is feed to the boundary regression network to correct the boundary and make it closer to the ground truth, then extend the proposal to a three-segment feature design with context information, and finally use a multi-layer perception to identify behavior. The experimental results show that the proposed algorithm achieves a great mAP in the THUMOS 2014 dataset and the ActivityNet v1.3 dataset.

Key words: human action recognition; two-stream convolutional networks; deep learning; temporal action localization

1　引言

随着各种摄像监控设备的快速发展, 视频和图像的数据量在不断增加. 如何分析视频图像中的信息也成为一个热门的研究内容, 视频分析中的一个重要分支就是行为识别. 人体行为识别的目标是从一个未知的视频或者是图像序列中自动分析其中正在进行的行为, 目前对于行为识别的研究热点主要是对短视频中单个行为的识别, 而在实际生活及应用中, 更多的视频数据是包含多个不同行为的复杂长视频. 这就需要使用另一种识别算法: 时序行为检测 (temporal action localization). 这种算法任务要求检测出长视频中每个行为的类别, 同时要标注出每个行为的开始时间和结

① 基金项目: 国家自然科学基金 (61472196, 61672305)
Foundation item: National Natural Science Foundation of China (61472196, 61672305)
收稿时间: 2019-01-21; 修改时间: 2019-02-21; 采用时间: 2019-03-04; csa 在线出版时间: 2019-07-01

束时间. 这种算法可以应用到许多方面, 比如自动检索和智能监控等.

时序行为检测通常可以分为两个阶段, 提议生成阶段和分类识别阶段. 提议生成阶段的主要目标是生成可能含有行为动作的视频片段, 视频片段称为行为提议, 而分类识别阶段的任务则是对提议生成阶段产生的行为提议进行识别分类, 并且进一步确定行为类别和起止时间. 尽管目前传统的行为识别已经达到较高的准确度, 但是在确定行为起止时间上仍然不尽如人意[1,2]. 因此, 如何产生高质量的行为提议, 成为该内容的一个重点研究方向[3−6]. 为了获得高质量的提议, 提议生成阶段产生的提议在持续时间上需要灵活可变, 用于应对视频片段持续时间长短不一并且差距较大的问题, 同时产生的提议应具有精确的时间边界. 最近的一些提议生成方法[3-5,7]利用不同长度的滑动窗口来生成提议, 然后使用训练好的模型来评估提议的置信度, 但是, 这种预先定义持续时间和间隔时间来产生提议的方法有一些明显的缺点: (1) 起止时间的精确度不足; (2) 固定的行为片段长度无法处理不同持续时间的行为动作, 而在不同行为动作持续时间差距较大时, 更会出现无法满足不同持续时间的要求, 而增多滑动窗口的数量又会带来大量冗余的计算.

最近的研究[7−9]将深度神经网络应用到检测框架中并且获得了较好的性能表现. S-CNN[7]提出了一个多阶段的卷积神经网络, 该算法通过使用定位网络提高了识别精度. 然而, S-CNN 使用滑动窗口产生行为提议, C3D[10]作为特征提取器最初用于单元分类器, 只能容纳 16 帧作为输入, 在应对时序行为检测任务时, 需要消耗大量的时间进行计算. 另一项研究[8]使用递归神经网络 (RNN) 来学习预测动作的起点和终点的一种策略. 这种顺序预测对于处理长视频通常非常耗时, 并且它不支持用于特征提取的逐帧 CNN 的联合训练.

本文在上述背景下, 为了克服滑动窗口的缺点, 生成高质量的行为提议, 本文提出了基于双流卷积神经网络[11]的时序行为检测模型. 该模型基于双流卷积神经网络提取的特征, 产生覆盖时间灵活可变的行为提议, 之后送入多层感知机中进行边界迭代回归, 然后将行为提议扩展为三段式的特征序列设计, 最后输入分类器中进行动作分类.

2 识别模型

本文提出一种基于双流卷积神经网络的模型, 如图 1 所示. 首先使用双流卷积神经网络提取长视频的特征序列, 然后将该特征序列作为模型的输入, 使用 Temporal Actionness Grouping (TAG)[12]方法在特征序列上灵活地生成行为提议. 利用多层感知机对每一个行为提议的起止边界进行迭代操作, 这一过程可以更为精细地处理行为提议的边界, 使之更加贴近真实的边界信息. 每一个行为提议都会使用三段式特征描述重新设计, 三段式设计将行为提议划分为开始区间、进行区间和结束区间, 按照前后顺序对应拼接相应的特征序列. 最后对包含目标动作的行为提议进行行为识别, 获得分类结果.

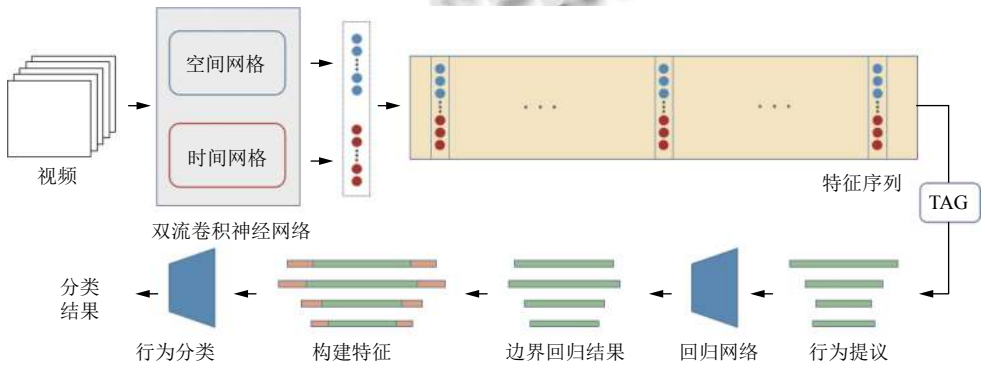

图 1 基于双流卷积神经网络的人体行为识别模型

2.1 问题描述

一个未分割的长视频可以表示为$X=\{x_n\}_{n=1}^{N}$, 其中x_n表示视频 X 中的第 n 帧. 视频 X 的动作标注由一组动作实例$\psi_g=\{\phi=(t_{s,n},t_{e,n})\}_{n=1}^{A_n}$组成, A_n是视频 X 中真实动作实例的数量, $t_{s,n}$, $t_{e,n}$分别是动作实例ϕ_n的开始时间和结束时间. 本文算法的任务就是自动定位每段行

为的起止位置并识别它们的行为属性.

2.2 特征序列提取

为了提取双流卷积神经网络特征, 将视频划分为 T 个连续等长且无重叠的单元, 则视频可以表示为 $X=\{s_t\}_{t=1}^{T}$, T 表示视频中单元的数量, 一个单元 $s_t=(x_{t_n},o_{t_n})$表示两部分的内容, x_{t_n}是视频 X 中的第 t_n个 RGB 帧, o_{t_n}是以x_{t_n}帧为中心, 附近的堆叠光流场. 为了减少计算损耗, 使用规律的帧间隔提取单元. 本任务所用数据集中的视频数据量大, 相邻的帧信息冗余度较高, 密集采样耗时且不必要, 因此使用规律的帧间隔提取单元, 在每个单元上获取特征, 可以在保证信息完整度的前提下降低计算损耗.

给定一个单元s_t, 在空间和时间网络的顶层连接输出分数以形成编码特征向量$f_{t_n}=(f_{S,t_n},f_{T,t_n})$, 其中$f_{S,t_n}$, f_{T,t_n}分别表示空间网络和时间网络的输出向量. 因此给定一个长度为 ls 的单元序列 S, 可以提取出特征序列 $F=\{f_{t_n}\}_{n=1}^{ls}$. 双流卷积特征序列将被送入 TAG 网络中生成行为提议.

2.3 行为提议

相比较于滑动窗口而言, TAG 方法能灵活的生成不同长度的动作提议, 同时并不需要大量的计算. TAG 方法使用了一个行为分类器来评估每个单元中发生动作的概率, 这个行为分类器是一个二元分类器. 该方法的基本思想是找到高动作概率的连续区域, 为了实现这个目的, 该方法重新设计了一个经典的分水岭算法, 并把它应用到了一维的动作概率值上. 该方法通过设置不同的“水位”可以得到一系列的“盆地”, 每一个盆地对应了时域范围内一段高动作概率区域.

给定一系列的盆地 G, 选用了一种类似于文献[13]的聚类方法, 这种方法试着连接小盆地变成行为提议区域. 该方案的工作流程如下: 先从一个种子盆地开始, 并且连续吸收随后的盆地, 直到盆地部分在整个持续时间内 (即从第一个盆地开始到最后一个盆地结束) 的部分下降到某个阈值 Y 以下. 通过这种方法, 可以从不同的种子盆地开始产生一组区域, 用$G'(\tau,\gamma)$来表示. 注意τ和γ并不是选择好的特定组合, 而是均匀地从 (0,1) 之间采样, 步长为 0.05. 这两个阈值的组合将会产生多组区域. 然后, 将他们结合起来, 并使用非极大值抑制的方法过滤重叠度高的区域, 设置 IoU 阈值为 0.95. 生成的行为提议将被送入多层感知机中边界回归.

2.4 边界回归

时域上进行边界回归的基本思路是利用神经网络推断行为提议的边界. 本文使用多层感知机作为回归网络, 将行为提议作为输入, 输出坐标回归偏移量, 具体计算如式 (1).

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

其中, s_{clip}, e_{clip}分别是输入的行为提议的开始和结束坐标, s_{gt}, e_{gt}分别是与之对应的真实数据的开始和结束坐标. 本文使用的坐标回归模型有两个优点: 第一, 使用单元级坐标回归, 这与双流卷积神经网络基于单元提取特征的方式相匹配, 计算消耗也比较小; 第二, 不使用坐标参数化, 直接使用起始坐标的偏移量作为回归结果. 这是因为行为提议的坐标回归在时域进行, 而空间坐标回归在空间域进行, 由于相机投影, 目标可以在图像中重新缩放, 因此需要先将边框坐标标准化为某个标准尺度. 而时域坐标可以依靠时域本身作为标准尺度, 不需要进行参数化.

在训练边界回归网络时, 需要给行为提议分配标签用以判断该行为提议中是否包含行为. 对于一个行为提议, 计算它和所有标定好的真实数据的 tIoU (temporal Intersection over Union) 重叠值, 如果其中的最大值超过了 0.5, 则将最大值对应的真实数据的边界和类别信息赋予该行为提议. 并将该行为提议视为正样本, 即含有行为, 否则视为负样本.

如图 2 所示, 本文的边界回归任务由多层感知机使用迭代的方式完成, 边界回归的输出结果作为输入再次送入多层感知机中进行计算, 重复多次以获得更为精确的结果. 该回归模型将行为提议作为输入, 输出时域上的坐标回归偏移量, 计算之后得到回归后的边界坐标值. 对于该层网络, 给定一个候选提议的边界数据输入值$p_c=[t_s,t_e]$, 输出数据$p_c^1=[t_s^1,t_e^1]$会作为输入进行第二轮的边界回归计算, 第二轮的输出为$p_c^2=[t_s^2,t_e^2]$. 迭代过程总共进行 K 次, 最后的边界结果为:

$$p_c^K=[t_s^K,t_e^K] \tag{2}$$

2.5 提议特征

为了建立如图 3 所示的提议特征φ, 对于一个行为提议, 将提议本身的范围定义为进行区间$p_c=[t_s,t_e]$, 提议φ的持续时间为$d=t_e-t_s$. 与它相关的开始区间和结束区间分别为$p_s=[t_s-d/4,t_s+d/4]$和$p_e=[t_e-d/4,t_e+$

$d/4$]. 对应选择开始、结束和进行区间三部分对应的特征序列, 将这些向量前后拼接, 即可获得候选提议φ的提议特征$f_\varphi=(f_{p_s},f_{p_c},f_{p_e})$. 该提议特征具有很好的鲁棒性, 在引入开始区间和结束区间后, 使得行为提议特征具备了上下文信息.

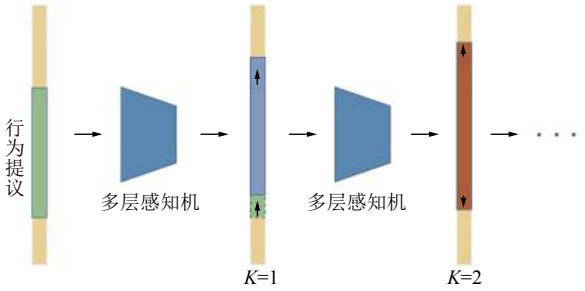

图 2 边界回归网络处理行为提议边界

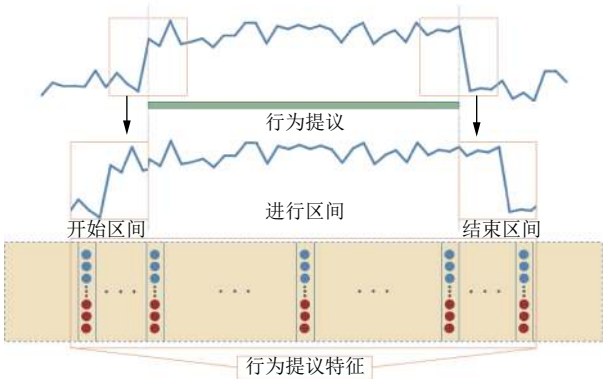

图 3 行为提议特征构建

2.6 行为分类

深度学习网络常用的分类器, 本文选择使用多层感知机网络作为特征构建后的多分类器. 对于时序行为检测任务, 多层感知机网络输出 n+1 个概率值, 其中 n 表示数据集中行为的数量, 1 表示背景类. 在 ActivityNet v1.3 中, n=200, 在 THUMOS 2014 中, n=20. 每个概率值表示属于某一类行为的概率, 将最大概率值对应的行为作为行为分类的结果.

为了获取较好的实验结果, 本文使用一个多任务损失函数来联合训练边界回归和行为分类网络. 损失函数如式 (3) 所示. 时序行为检测任务需要对行为定位和识别, 这两个任务息息相关, 如果单独训练网络会降低识别的泛化能力, 可能会出现对某一任务的过拟合现象. 而联合训练可以较好的解决这个问题, 联合训练可以在有限的数据集内完成训练, 由于引入了额外的相关训练数据, 有助于网络学习到更适合任务需求的参数, 可以提高模型的泛化能力. 行为的类别和发生时间是个体属性的不同方面, 具有较强的相关性, 使用联合训练可以使得定位与识别任务真正地结合起来, 学习到的内容彼此受益, 提高时序行为检测的准确率.

$$L=L_{cls}+\lambda L_{reg} \tag{3}$$

其中, L_{cls}是分类损失函数, 对于本文中多分类任务而言, 使用多分类交叉熵函数作为损失函数. L_{reg} 是边界回归损失函数, λ是超参数. 回归损失函数为:

$$L_{reg}=\frac{1}{N}\sum_{i=1}^{N}\sum_{z=1}^{n}l_i^z\left[R\left(o'^z_{s,i}-o^z_{s,i}\right)+R\left(o'^2_{e,i}-o^z_{e,i}\right)\right] \tag{4}$$

其中, R 是曼哈顿距离, N 是 batch size, n 是行为类别的总数量, l_i^z是标签, 当第 i 个样本属于 z 类时, $l_i^z=1$, 否则, $l_i^z=0$. o'是回归偏移量, o 是真实数据. 学习率设置为 0.005, batch size 设置为 128.

3 实验

为了验证本文算法的有效性, 本文在 ActivityNet v1.3[1]和 THUMOS 2014[2]数据集上进行实验. ActivityNet v1.3 数据集是常用的时序行为检测数据集, 包括 200 类不同的动作, 同时提供了边界和种类信息标注. THUMOS 2014 中没有训练集, 有 20 类行为带有标注. 本文分别在两个数据集上进行实验, 在各自提供的数据子集上训练网络, 并使用预训练的网络进行测试, 将实验结果与现有方法进行对比分析.

3.1 数据集

ActivityNet v1.3[1]是一个用于时序行为检测的大型数据集, 其中包含 19994 个带有 200 类动作标注的长视频, 在 2017 年和 2018 年的 ActivityNet 挑战中使用了该数据集. ActivityNet 按照 2: 1: 1 的比例分为训练集、验证集和测试集.

THUMOS 2014[2]有 1010 个视频用于验证, 1574 个视频用于测试. 这些视频中包含 20 类带有行为标注的目标动作. 该数据集没有训练集, 使用 UCF101 数据集作为训练集. 由于训练集没有提供时间注释, 本文在验证集上训练模型并在测试集上进行实验测试. 因此将带有 20 类行为标注的 220 个视频用于训练. 在本文的实验中, 将本文提出的方法与 THUMOS 2014 和 ActivityNet v1.3 上的现有技术进行比较, 并进行结果分析.

3.2 实验网络参数设置

本文实验环境选择深度学习框架 Caffe 平台实现. 使用 SGD 方法学习模型中的参数, batch size 为 128, momentum 为 0.9. 双流卷积神经网络采用 ResNet 网络用作空间网络, BN-Inception 网络用作时间网络. 空间网络和时间网络的初始学习率分别设置为 0.001 和 0.005. 在 ActivityNet v1.3 中, 空间网络和时间网络迭代训练次数分别为 9500 次和 20 000 次, 学习率分别在迭代每 4000 次和 1000 次后缩小 0.1. 在 THUMOS 2014 中, 空间网络和时间网络分别进行 1000 次和 6000 次的迭代训练, 学习率在每 400 和 2500 次时缩小 0.1. 在特征提取过程中, 单元间隔均被设置为 16. 在 TAG 方法中使用的二元行为分类器使用每个数据集的训练集进行训练. 在边界回归过程中, K=3.

3.3 实验结果分析

评价标准: AvtivityNet v1.3[1]和 THUMOS 2014[2]都有统一的评价标准, 因此按照它们的评价标准测试不同 IoU 阈值的平均预测精度 mAP. 在 ActivityNet v1.3 数据集中, 所需测试的 IoU 阈值为{0.5, 0.75, 0.95}, IoU 阈值范围[0.5: 0.05: 0.95]的 mAP 的平均值用于比较不同方法之间的性能. 在 THUMOS 2014 数据集中, 所需测试的 IoU 阈值为{0.1, 0.2, 0.3, 0.4, 0.5}. 阈值为 0.5 时得出的平均预测精度用于比较不同方法的实验结果.

将本文算法与其它时序行为检测方法在 THUMOS 2014 数据集和 ActivityNet v1.3 数据集上进行比较, 如表 1、表 2 所示. 从表 1、表 2 中可以发现, 在这两个数据集上, 本文提出的算法识别准确率优于其它算法, 识别效果较好. 本文使用双流卷积神经网络所获取的特征结合了运动表层特征和时序信息两部分, 更好的发掘了视频所包含的信息. 行为提议在经过多层感知机迭代处理后边界信息更为准确, 之后的三段式特征设计融合了上下文信息, 一方面建立了较为全面的行为描述, 另一方面提高了行为识别准确率.

4 结论与展望

为了充分获取视频中的时空信息, 使用双流卷积神经网络构建特征描述符, 之后通过 TAG 方法产生候选行为提议, 经过多次迭代处理后获取更为准确的边界信息, 将行为提议扩展为三段式特征设计, 并对目标行为进行识别. 该方法在结合时序信息的基础上, 生成了质量较高的动作提名, 时序边界更为准确, 识别率也有所提升. 实验结果表明该方法能在 THUMOS 2014 数据集合 ActivityNet v1.3 数据集上得到较好的效果. 但是行为提议生成和回归的方法着眼于局部信息, 缺少与行为提议全局特征的结合分析, 时序定位的准确度仍有不足. 下一步的研究将会引入行为提议的特征共同分析定位准确度, 获得更为准确的时序边界.

表 1 不同时序行为检测算法在 THUMOS 2014 数据集上的准确率 (%)

THUMOS 2014, mAP					
算法名称	0.1	0.2	0.3	0.4	0.5
Wang *et al.*[14]	18.2	17.0	14.0	11.7	8.3
Oneata *et al.*[15]	36.6	33.6	27.0	20.8	14.4
Richard *et al.*[16]	39.7	35.7	30.0	23.2	15.2
S-CNN[9]	47.7	43.5	36.3	28.7	19.0
Yeung *et al.*[8]	48.9	44.0	36.0	26.4	17.1
TCN[17]	-	-	-	33.3	25.6
STPN[18]	52.0	44.7	35.5	25.8	16.9
本文算法	**68.9**	**59.0**	**53.4**	**40.6**	**30.1**

表 2 不同时序行为检测算法在 ActivityNet v1.3 数据集上的准确率 (%)

ActivityNet v1.3(testing), mAP				
算法名称	0.5	0.75	0.95	Average
Singh *et al.*[19]	28.67	17.78	2.88	17.68
SCC[20]	40.00	17.90	4.70	19.30
TCN[17]	37.49	23.47	4.47	23.58
R-C3D + Boundary[21]	27.82	15.00	2.82	15.68
本文算法	**45.73**	**30.56**	**5.42**	**29.39**

参考文献

1 Heilbron FC, Escorcia V, Ghanem B, *et al*. ActivityNet: A large-scale video benchmark for human activity understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA. 2015. 961–970.

2 Idrees H, Zamir AR, Jiang YG, *et al*. The THUMOS challenge on action recognition for videos “in the wild”. Computer Vision and Image Understanding, 2017, 155: 1–23. [doi: 10.1016/j.cviu.2016.10.018]

3 Buch S, Escorcia V, Shen CQ, *et al*. SST: Single-stream temporal action proposals. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 2911–2920.

4 Heilbron FC, Niebles JC, Ghanem B. Fast temporal activity

proposals for efficient detection of human actions in untrimmed videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 1914–1923.

5 Escorcia V, Heilbron FC, Niebles JC, *et al*. Daps: Deep action proposals for action understanding. European Conference on Computer Vision. The Netherlands. 2016. 768–784.

6 Gao JY, Yang ZH, Chen K, *et al*. Turn tap: Temporal unit regression network for temporal action proposals. Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy. 2017. 3628–3636.

7 Shou Z, Wang DG, Chang SF. Temporal action localization in untrimmed videos via multi-stage CNNs. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 1049–1058.

8 Yeung S, Russakovsky O, Mori G, *et al*. End-to-end learning of action detection from frame glimpses in videos. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 2678–2687.

9 De Geest R, Gavves E, Ghodrati A, *et al*. Online action detection. European Conference on Computer Vision. The Netherlands. 2016. 269–284.

10 Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 4489–4497.

11 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada. 2014. 568–576.

12 Zhao Y, Xiong YJ, Wang LM, *et al*. Temporal action detection with structured segment networks. Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy. 2017. 2914–2923.

13 Pont-Tuset J, Arbeláez P, Barron JT, *et al*. Multiscale combinatorial grouping for image segmentation and object proposal generation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(1): 128–140. [doi: 10.1109/TPAMI.2016.2537320]

14 Wang L, Qiao Y, Tang X. Action recognition and detection by combining motion and appearance features. THUMOS14 Action Recognition Challenge, 2014, 1(2): 2.

15 Oneata D, Verbeek J, Schmid C. The Lear submission at THUMOS 2014. In THUMOS Action Recognition Challenge, 2014.

16 Richard A, Gall J. Temporal action detection using a statistical language model. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 3131–3140.

17 Dai XY, Singh B, Zhang GY, *et al*. Temporal context network for activity localization in videos. Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy. 2017. 5793–5802.

18 Nguyen P, Han B, Liu T, *et al*. Weakly supervised action localization by sparse temporal pooling network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA. 2018. 6752–6761.

19 Singh B, Marks TK, Jones M, *et al*. A multi-stream Bi-directional recurrent neural network for fine-grained action detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 1961–1970.

20 Heilbron FC, Barrios W, Escorcia V, *et al*. SCC: Semantic context cascade for efficient action detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 1454–1463.

21 Kong WJ, Li NN, Liu S, *et al*. BLP-boundary likelihood pinpointing networks for accurate temporal action localization. arXiv preprint arXiv: 1811. 02189, 2018.