Computer Science and Application 计算机科学与应用, 2023, 13(3), 626-634
Published Online March 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.133062

Hans汉斯

基于深度学习方法的在线动作检测技术综述

张　婉，张睿萱，谢　昭*，刘家仁，金宇奇，沈玉龙

合肥工业大学计算机与信息学院，安徽　宣城

收稿日期：2023年2月23日；录用日期：2023年3月24日；发布日期：2023年3月31日

摘　要

动作检测技术，是在算法观测整个视频后自动识别出其中出现的动作类别和始末时间，在机器人、智能家居、城市安防等领域均有应用。然而实际生活中，很多场景需要在某些事件刚发生时给予反馈，这需要检测算法以一种在线形式接收视频信息，传统的动作检测算法因为观测信息不完全，效果很差。本文基于当前在线动作检测算法的研究现状，概述了目前用于在线检测的主流方法，总结了目前研究将遇到的挑战。

关键词

在线动作检测，机器视觉，深度学习

A Review of Online Action Detection Techniques Based on Deep Learning Methods

Wan Zhang, Ruixuan Zhang, Zhao Xie*, Jiaren Liu, Yuqi Jin, Yulong Shen

School of Computer and Informatics, Hefei University of Technology, Xuancheng Anhui

Received: Feb. 23rd, 2023; accepted: Mar. 24th, 2023; published: Mar. 31st, 2023

Abstract

Action detection technology, in which an algorithm observes the entire video and then automatically identifies the type of action that occurs in it and the start and end times, is used in robotics, smart homes, urban security and other areas. However, in real life, many scenarios require feedback when certain events first occur, which requires detection algorithms to receive video information in an online format. Traditional action detection algorithms are ineffective because of in-

*通讯作者。

文章引用: 张婉, 张睿萱, 谢昭, 刘家仁, 金宇奇, 沈玉龙. 基于深度学习方法的在线动作检测技术综述[J]. 计算机科学与应用, 2023, 13(3): 626-634. DOI: 10.12677/csa.2023.133062

complete observation information. Based on the current state of research in online action detection algorithms, this paper provides an overview of the mainstream methods currently used for online detection and summarises the challenges that current research will encounter.

Keywords

Online Action Detection, Computer Vision, Deep Learning

1. 引言

随着近几年计算机视觉的飞速发展，动作检测作为视频领域一个重要分支，逐步成为学术热点，并取得了丰硕的成果[1] [2]。动作检测任务目标为在长视频中找出其包含的动作类别，并指明每个动作开始和结束时间。目前多为离线动作检测，常用于视频的检索和分类。然而在实际的应用场景中：

1) 对危险行为的检测和预防，如监控视频中的异常行为检测、驾驶员酒后或疲劳驾驶检测、行人闯红灯行为检测等，使用传统的动作检测算法只能接受行为带来的风险，而无法做出事先的预防。

2) 实时人机交互系统，如医院的服务类机器人、生活中的安防机器人等，必须实时检测动作并做出实际反应。

3) 在虚拟现实中，系统根据用户动作参数，及时预测出用户当前进行动作，从而虚拟世界中人物做出相应对策，提高用户体验。

包括以上需求在内的多种应用场景刺激了动作检测向实时检测方向拓展。在线动作检测[3]，接收的信息以流形式，在动作刚开始识别出其动机，从而及时做出决策，解决上述实时性问题。具体任务图如图 1 所示。

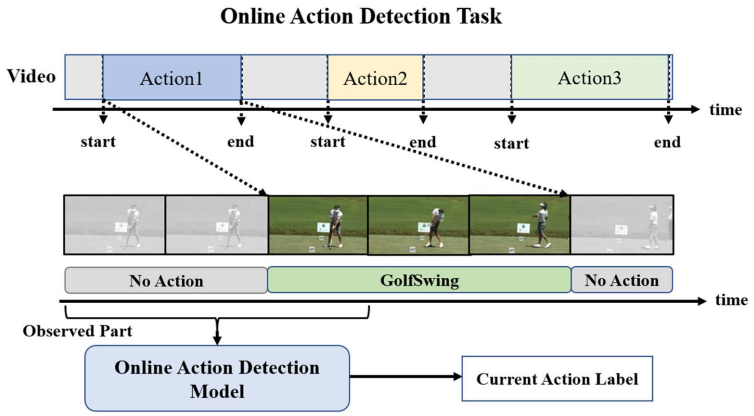

Figure 1. Online action detection task
图 1. 在线动作检测任务图

在线动作检测方法从模型架构来看主要分为两类：基于循环神经网络方法和基于 Transformer [4]方法。在线动作检测任务最早于 2016 年由 Geest 等人[3]提出，由于处理数据以流形式呈现，故最初模型以

序列模型为基础，在此上进行改进。随后 4 年内，出现了 TRN [5] (Temporal Recurrent Networks)，IDN [6] (Information Discrimination Network)等模型，该类模型均以 RNN 为基础，在下文均有介绍。随着 2017 年 Transformer [4]首次提出解决机器翻译任务，在计算机视觉领域，尤其是视频领域，由于视频属于序列与 Transformer [4]框架契合，很多视频领域模型均以 Transformer [4]作为基础，在 2020 年后出现了 OadTR [7] (online action detection with transformers)，LSTR [8]，Colar [9]等模型。相比较以 RNN 为基础模型，该类模型能够更好获取全局信息。

除了围绕以上两种主要模型的进展外，还有学者研究对数据集的预处理。Hyunjun Eun 等人采用 CNN 进行特征提取，并提出一种架构简单的时间过滤网络 TFN [10]。TFN [10] 提供一个过滤模块训练相关系数的得分，以此来反映当前信息与当前动作的相关性，从而过滤掉背景和不相关的动作。TFN [10]在 THUMOS-14 [11]和 TVSeries [3]数据集上都表现出良好的表现。

目前在线动作检测领域，英文综述极度匮乏。在中文综述中，未获取相关文献。对于该项新任务，具备较大应用前景，需要更多关注。本文聚焦于介绍从 2016 年至今主要解决在线动作检测的模型和数据集，对任务难点和未来发展方向做了总结。

2. 在线动作检测算法

目前已经发表的有关在线动作检测的算法并不是很多，大致上从其发展流程来看可分为以循环神经网络作基础和以 Transformer [4]作基础两类，下面着重介绍每一类代表模型，分析其优缺点。

2.1. 基于循环神经网络方法

循环神经网络(Recurrent Neural Network, RNN)最早于 20 世纪 80 年代提出，用于处理序列数据。随着深度学习发展，出现了 LSTM [12]，GRU [13]等变体。研究表明在用于较多数据情况下，以循环神经网络为基础的网络效果优于传统的如隐马尔可夫模型等。

早在 2012 年 Hoai 等人[14]便已经提出在线动作检测的类似概念，但由于数据所限，当时的方法采用滑动窗口机制，即将视频分为相互重叠的小片段，对每个片段进行检测和分类。采用这种方式效果慢，且整体看并不是以在线流的形式。Geest 等人[3]在 2016 年首次提出在线动作检测，介绍了数据集 TVSeries，给出解决任务的基本框架。随后 Li 等人[15]首次提出了联合分类回归循环神经网络(Joint Classification-Regression Recurrent Neural Network)，以解决动作序列长度不确定的流序列动作检测问题，文中采用人体骨架信息作为输入，网络中主体采用 LSTM [12]，将流序列动作标签分类和开始结束时间预测同步进行，文中同时公开 OAD 数据集，用于后续的在线动作检测工作。2018 年，Geest 等人[16]提出一种双流反馈网络，对特征和帧间时间依赖性分开处理。2019 年 Xu 等人[5]提出 TRN [5]，在此之间在线动作检测仅仅基于过去观察的历史信息，受人们辨别当前动作时首先预判未来这一现象影响，TRN [5]首先预测近邻的未来信息，然后将其与观测信息结合，从而实现辨别当前动作。2020 年 EUN 等人提出 IDN [6]，IDN 是在 GRU [13]基础上的一种拓展，主要针对正在输入的信息，判别其与当前动作相关性来确定是否使用输入信息，这样降低了背景或者其他动作的信息的干扰。2021 年 Wang 等人[17]对四种基本的时间建模方法——时间池、时间卷积、递归神经网络和时间注意力模型进行了研究实证，解释了它们对在线动作检测的效果。并探索了出了 6 种混合时态模型，揭示了时态模型之间的互补性，最终通过实验得出了扩张因果卷积 DCC 和非局部 M-NL 或 LSTM 的简单混合，显著改善了个体模型。Kim 等人[18]基于预测未来以更好辨别当前动作这一思想基础上，设计了时间平滑网络来更好地适应在线动作检测任务。2022 年，Sunah Min 等人[19]。拓展 LSTM 引入信息高程单元 IEU，提出了信息高程网络 IEN，通过 IEU 的附加信息提升门将过去的、被遗忘的、但与当前信息相关的信息提升到单元状态，优化对未

来的预测。

总的来看，以循环神经网络作为基础的模型，优点在于保持了视频的顺序性，能够更好处理局部帧之间的关系。然而实际上过去某些帧对未来可能存在大的影响。故该类方法缺点在于无法获得较长跨度帧与帧之间的关系，且在采用自回归预测情况下容易造成错误累积，即由中间环节预测错误导致最终结果错误。

2.2. 基于 Transformer 方法

Transformer [4]于 2017 年提出，用于解决机器翻译等任务。近些年越来越多计算机视觉模型采用其作为基础。Transformer [4]的核心在于采用了自注意力机制，降低序列中不重要部分的比重。

最早将 Transformer [4]用于在线动作检测的模型为 OadTR [7]，于 2021 年 Wang 等人提出。模型思想和前面 TRN [5]类似，均是预测未来以更好检测当前动作。用于架构中数据并行输入，在推理速度上明显优于以 RNN 为基础的模型。Xu 等人[8]提出 LSTR [8]，通过长时和短时记忆来建模视频数据，模型动态利用长时间窗口来对较大尺度历史信息建模从而得到粗略特征，以短时间窗口得到最近的精细特征。2022 年郭洪基等人[20]提出一种基于不确定性的在线动作检测时空注意力，通过输入计算动作预测的不确定性并赋予注意力权重，输入模型的特征更具有相关性和可辨别性。作者将该方法分别应用于 TRN [5]，OadTR [7]，和 LSTR [8]，结果表明在 THUMOS-14 [11]，TVSeries [3]和 HDD [21]数据上均提高了模型的性能。该研究方法在冗杂的背景下和小规模的数据集下表现出更好的泛化能力。Yang 等人[9]提出 Colar，首次在在线动作检测中使用类别级建模和范例参考机制，将历史帧作为范例指导当前帧从而获得长期依赖关系，同时类别级建模为时态依赖性提供补充。Chen 等人[22]提出了 GateHub (gated history unit with background suppression)，结合了 Transformer 和 RNN，使模型同时具备了对长时间信息的建模能力和选择信息进行编码的能力。具体来说，GateHub 还提出了未来增强 FaH，通过观察到的可用随后帧增强历史特征的信息性；提出了位置引导门控交叉注意力机制(position-guided gated cross-attention)，根据预测当前帧的信息量来增强或抑制部分历史信息；引入了背景抑制目标函数，降低背景帧误报为动作帧的可能性。

总的来看，以 Transformer [4]作为基础的模型，结果上大大优于以循环神经网络为基础的模型。Transformer [4]能够获得长时时序依赖关系，且推理速度快。但其缺点在于无法获取局部序列间关系，且模型计算量大，需要的数据量大。实际上，如何获取和使用长时时序依赖和短时时序依赖是该领域一个重要的研究方向。

3. 数据集与评价指标

3.1. 数据集

目前常用与在线动作检测任务数据集如表 1 所示。

Table 1. Dataset introduction
表 1. 数据集

数据集	数据来源	类别个数	数据规模
THUMOS'14	生活实际	20	训练集为 UCF101 子集，验证集和测试集共 2584 个实例
TVSeries	影视作品	30	6 种节目，共计 16 小时
ActivityNet	Youtube	200	15,000 个视频，共计 200 小时
OAD	生活实际	10	59 个长视频，共计 216 分钟

THUMOS’14 [11]主要由真实生活场景人类动作组成，共包含 20 类别，常用于动作识别和动作检测任务。训练集以 UCF101 [23]为基础，对每个视频做了剪辑，包含 13,320 个实例，验证集和测试集未被剪辑，包含背景信息，两者分别包含 1010 和 1574 个实例。目前方法常用 VGG [24]作特征预提取后的特征为输入。

TVSeries [3]数据主要源于热门影视作品，数据集共包含 6 个节目，每个节目时间大致为 150 分钟，共计 16 小时。数据集中标签共分为 30 种，手工标注出了动作开始和结束时间，由于数据集动作主体繁多、行事风格多样且不同动作类别可能在时间上交叉，故该数据集难度相对较大。

ActivityNet [25]数据源于 Youtube，是目前最大的用于动作检测数据集。数据集共包含 15,000 个视频，共 200 小时，其中 2/3 用于训练，整个数据集被分为 200 个动作标签。

OAD [16]数据集首次伴随在线动作检测任务而提出，该数据集主要记录居家日常生活。数据集包含 10 个动作类别，共计 216 分钟。视频中每个动作持续时间以及动作间间隔时间都是不固定的，故数据贴切生活实际。

3.2. 评价指标

和动作检测指标相同，现有的研究中主要采用的指标有 *AP*，*mAP*，*cAP*，*mcAP*。在线动作检测中算法会逐个得到每一帧对应的标签，定义某类别的 *AP* 计算方式如式(1)，(2)：

$$Prec = \frac{TP}{TP + FP} \tag{1}$$

$$AP = \frac{\sum_{k} Prec(k) \times I(k)}{\sum TP} \tag{2}$$

其中当第 *k* 帧为真正例时，*I*(*k*)为 1。*TP* 表示真正例数量，*FP* 表示假正例数量。

mAP 是对所有类别的 *AP* 做一个平均，计算公式如式(3)。

$$mAP = \frac{\sum_{m} AP(m)}{M} \tag{3}$$

其中 *M* 为类别数量。

由于正例和负例不均衡现象，严重影响 *AP* 计算结果。目前有模型采用 *cAP*，缓解上述问题，计算公式如式(4)，(5)：

$$cPrec = \frac{TP}{TP + \frac{FP}{w}} \tag{4}$$

$$cAP = \frac{\sum_{k} cPrec(k) \times I(k)}{\sum TP} \tag{5}$$

其中 *w* 表示正例和反例个数之间的比值。

mcAP 为所有类别 *cAP* 的平均值，计算公式如式(6)。

$$mcAP = \frac{\sum_{m} cAP(m)}{M} \tag{6}$$

3.3. 结果比较

本节主要对现有模型在 THUMOS’14 和 TVSeries 两个数据集上主要指标做汇总。

数据集 THUMOS’14 上，大多数模型选择预训练 TSN (Temporal Segment Networks)模型做特征预提取，且根据预训练时选择特征不同分为 ActivityNet 和 Kinetics。在线动作检测模型效果如表 2 所示。

Table 2. Dataset THUMOS’14 online motion detection effect (mAP)
表 2. 数据集 THUMOS’14 在线动作检测效果(mAP)

模型名称	预训练特征	mAP	年份
CDC [26]	ActivityNet	44.4	2017
RED [27]		45.3	2017
FATS [18]		51.6	2021
IDN [6]		50.0	2020
LAP [28]		53.3	2020
TFN [10]		55.7	2021
OadTR [7]		58.3	2021
Colar [9]		59.4	2022
LFB [29]		61.6	2019
LSTR [8]		**65.3**	**2021**
FATS [18]	Kinetics	59.0	2021
IDN [6]		60.3	2020
TRN [5]		62.1	2019
PKD [30]		64.5	2020
WOAD [31]		67.1	2021
LFB [29]		64.8	2019
OadTR [7]		65.2	2021
Colar [9]		66.9	2022
LSTR [8]		69.5	2021
GateHub [22]		**70.7**	**2022**

数据集 TVSeries 上，和前述数据集做法类似根据预训练时选择特征不同分为 ActivityNet 和 Kinetics，且模型同时比较不同观测比例下在线动作检测。在线动作检测模型效果如表 3 所示。

Table 3. Dataset TVSeries online motion detection effect (mcAP)
表 3. 数据集 TVSeries 在线动作检测效果(mcAP)

模型名称	预训练特征	观测比例									
		0%~10%	10%~20%	20%~30%	30%~40%	40%~50%	50%~60%	60%~70%	70%~80%	80%~90%	90%~100%
TRN [5]	ActivityNet	78.8	79.6	80.4	81.0	81.6	81.9	82.3	82.7	82.9	83.3
IDN [6]		80.6	81.1	81.9	82.3	82.6	82.8	82.6	82.9	83.0	83.9
TFN [10]		83.1	84.4	85.4	85.8	87.1	88.4	87.6	87.0	86.7	85.6
OadTR [7]		79.5	83.9	86.4	85.4	86.4	87.9	87.3	87.3	85.9	84.6
Colar [9]		80.2	84.4	**87.1**	85.8	86.9	**88.5**	88.1	87.7	86.6	85.1
LSTR [8]		**83.6**	**85.0**	86.3	**87.0**	**87.8**	**88.5**	**88.6**	**88.9**	**89.0**	**88.9**

Continued

IDN [6]	Kinetics	81.7	81.9	83.1	82.9	83.2	83.2	83.2	83.0	83.3	86.6
PKD [30]		82.1	83.5	86.1	87.2	88.3	88.4	89.0	88.7	88.9	87.7
OadTR [7]		81.2	84.9	87.4	87.7	88.2	89.9	88.9	88.8	87.6	86.7
Colar [9]		82.3	85.7	88.6	88.7	88.8	**91.2**	89.6	89.9	88.6	87.3
LSTR [8]		84.4	85.6	87.2	87.8	88.8	89.4	89.6	89.9	90.0	90.1
GateHub [22]		**84.5**	**87.6**	**89.5**	**90.0**	**90.2**	91.0	**91.3**	**91.3**	**91.3**	**90.7**

4. 结束语

随着深度学习不断发展，学术界提出越来越多贴切生活实际的方向，在线动作检测可用于许多潜在的视频监控应用、交互式机器人服务、ADAS (Advanced Driver Assistance System)、体育赛事分析和冲突行为警告系统等。与此同时，得益于算力水平不断提高，目前检测算法能够在较短时间内处理相当的数据。

与一次性读取所有帧的离线动作检测不同，在线动作检测只观察部分动作，许多关键信息是未知的。因此，在线动作检测的挑战在于根据已知的信息尽快做出判断，这就要求系统有很高的判断力和灵敏度[32]。

对于视频中时间序列的处理，在早期算力不足和数据匮乏情况下，采用循环神经网络，但从循环神经网络架构来看，其优点在于能自发地获得帧与帧之间的先后关系，保持了视频的顺序性；但其也有大量缺点，首先训练时存在梯度消失和梯度爆炸问题，其次循环神经网络每个单元的记忆相对短，无法获取全局语义信息，这使得早出现的帧对晚出现帧无影响，此外，采用自回归形式仍会出现错误累积现象。随着 Transformer [4]架构首次用于解决自然语言处理任务，在计算机视觉领域，很多方向的模型均采用 Transformer [4]作为基础。采用该架构优点在于能捕获全局信息，对于动作检测，可以发掘早期行为对当前行为的影响，同时推理速度较循环神经网络快很多。但该架构缺点在于 Transformer [4]效果好依托于训练数据量大，同时模型无法获取局部帧之间关系。

目前最新方法尽管从演示效果看已取得较好精度，但仍具备很大提升空间，以下 4 点是对该领域未来展望：

1) 开发范围更广，更贴切实际生活，贴切任务目标的数据集，从而开发更大模型，在更多动作上做检测。

2) 在训练过程中如何获取长时趋势和短时线索，以及如何将它们相结合，这是一个值得考虑的问题，也是目前主流方法研究方向。

3) 在线动作检测突出解决检测中实时性问题，如何在序列不断变长时进行高效的计算，这是该项技术能否落地的一个关键。

4) 动作实例和背景之间的边界模糊是固有的。如何定义统一的起止时间点，并使其在不同人之间无争议，仍然值得研究[32]。

参考文献

[1] Vaudaux-Ruth, G., Chan-Hon-Tong, A. and Achard, C. (2021) SALAD: Self-Assessment Learning for Action Detection. *WACV*, Waikoloa, 3-8 January 2021, 1268-1277. https://doi.org/10.1109/WACV48630.2021.00131

[2] Shi, D.F., Zhong, Y.J., Cao, Q., *et al*. (2022) ReAct: Temporal Action Detection with Relation Queries. *ECCV*, Tel Aviv, 23-27 October 2022, 105-121. https://doi.org/10.1007/978-3-031-20080-9_7

[3] De Geest, R., Gavves, E., Ghodrati, A., *et al*. (2016) Online Action Detection. *ECCV*, Amsterdam, 11-14 October 2016,

269-284. https://doi.org/10.1007/978-3-319-46454-1_17

[4] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *NIPS*, Long Beach, 12 June 2017, 5998-6008.

[5] Xu, M.Z., Gao, M.F., Chen, Y.-T., *et al*. (2019) Temporal Recurrent Networks for Online Action Detection. *ICCV*, Seoul, 27 October-2 November 2019, 5531-5540.

[6] Eun, H., Moon, J., Park, J., *et al*. (2020) Learning to Discriminate Information for Online Action Detection. *CVPR*, Seattle, 13-19 June 2020, 806-815.

[7] Wang, X., Zhang, S.W., Qing, Z.W., *et al*. (2021) Oadtr: Online Action Detection with Transformers. *ICCV*, Virtual, 21 June 2021, 7545-7555. https://doi.org/10.1109/ICCV48922.2021.00747

[8] Xu, M.Z., Xiong, Y.J., Chen, H., *et al*. (2021) Long Short-Term Transformer for Online Action Detection. *NeurIPS*, Virtual, 7 July 2021, 1086-1099.

[9] Yang, L., Han, J.W. and Zhang, D.W. (2022) Colar: Effective and Efficient Online Action Detection by Consulting Exemplars. *CVPR*, New Orleans, 2 March 2022, 3150-3159. https://doi.org/10.1109/CVPR52688.2022.00316

[10] Eun, H., Moon, J., Park, J., *et al*. (2021) Temporal Filtering Networks for Online Action Detection. *Pattern Recognition*, **111**, Article ID: 107695. https://doi.org/10.1016/j.patcog.2020.107695

[11] Idrees, H., Zamir, A.R., Jiang, Y.-G., *et al*. (2017) The THUMOS Challenge on Action Recognition for Videos "in the Wild". *Computer Vision and Image Understanding*, **155**, 1-23. https://doi.org/10.1016/j.cviu.2016.10.018

[12] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[13] Cho, K., *et al*. (2014) Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *The Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, 25-29 October 2014, 1724-1734.

[14] Hoai, M. and De la Torre, F. (2014) Max-Margin Early Event Detectors. *International Journal of Computer Vision*, **107**, 191-202. https://doi.org/10.1007/s11263-013-0683-3

[15] Li, Y., Lan, C., Xing, J., *et al*. (2016) Online Human Action Detection Using Joint Classification-Regression Recurrent Neural Networks. *ECCV*, Amsterdam, 11-14 October 2016, 203-220. https://doi.org/10.1007/978-3-319-46478-7_13

[16] De Geest, R. and Tuytelaars, T. (2018) Modeling Temporal Structure with LSTM for Online Action Detection. *WACV*, Lake Tahoe, 12-15 March 2018, 1549-1557. https://doi.org/10.1109/WACV.2018.00173

[17] Wang, W., Peng, X., Qiao, Y. and Cheng, J. (2022) An Empirical Study on Temporal Modeling for Online Action Detection. *Complex & Intelligent Systems*, **8**, 1803-1817. https://doi.org/10.1007/s40747-021-00534-3

[18] Kim, Y.H., Nam, S. and Kim, S.J. (2021) Temporally Smooth Online Action Detection Using Cycle-Consistent Future Anticipation. *Pattern Recognition*, **2021**, Article ID: 107954. https://doi.org/10.1016/j.patcog.2021.107954

[19] Min, S. and Moon, J. (2022) Information Elevation Network for Online Action Detection and Anticipation. 2022 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), New Orleans, 19-20 June 2022, 2549-2557. https://doi.org/10.1109/CVPRW56347.2022.00287

[20] Guo, H., Ren, Z., Wu, Y., Hua, G. and Ji, Q. (2022) Uncertainty-Based Spatial-Temporal Attention for Online Action Detection. *European Conference on Computer Vision* (*ECCV*), Tel Aviv, 23-27 October 2022, 69-86. https://doi.org/10.1007/978-3-031-19772-7_5

[21] Ramanishka, V., Chen, Y.-T., Misu, T. and Saenko, K. (2018) Toward Driving Scene Understanding: A Dataset for Learning Driver Behavior and Casual Reasoning. *CVPR*, Salt Lake City, 18-22 June 2018, 7699-7707.

[22] Chen, J.W., Mittal, G., Yu, Y., Kong, Y. and Chen, M. (2022) Gatehub: Gated History Unit with Background Suppression for Online Action Detection. *CVPR*, New Orleans, 18-24 June 2022, 19925-19934. https://doi.org/10.1109/CVPR52688.2022.01930

[23] Soomro, K., Zamir, A.R. and Shah, M. (2012) Ucf101: A Dataset of 101 Human Actions Classes from Videos in the Wild.

[24] Simonyan, K. and Zisserman, A. (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. *International Conference on Learning Representations*, San Diego, 7-9 May 2015.

[25] Heilbron, F.C., Escorcia, V., Ghanem, B. and Niebles, J.C. (2015) Activitynet: A Large-Scale Video Benchmark for Human Activity Understanding. *CVPR*, Boston, 7-12 June 2015, 961-970. https://doi.org/10.1109/CVPR.2015.7298698

[26] 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. *CVPR*, Honolulu, 21-26 July 2017, 1417-1426. https://doi.org/10.1109/CVPR.2017.155

[27] Gao, J.Y., Yang, Z.H. and Nevatia, R. (2017) RED: Reinforced Encoder-Decoder Networks for Action Anticipation.

BMVC, London, 4-7 September 2017.

[28] Qu, S.Q., Chen, G., Xu, D., Dong, J.H., Lu, F. and Knoll, A. (2020) LAP-Net: Adaptive Features Sampling via Learning Action Progression for Online Action Detection.

[29] Wu, C.-Y., Feichtenhofer, C., Fan, H.Q., He, K.M., Krahenbuhl, P. and Girshick, R. (2019) Long-Term Feature Banks for Detailed Video Understanding. *CVPR*, Long Beach, 16-20 June 2019, 284-293.

[30] Zhao, P.S., Wang, J.J., Xie, L.X., Zhang, Y., Wang, Y.F. and Tian, Q. (2020) Privileged Knowledge Distillation for Online Action Detection.

[31] Gao, M.F., Zhou, Y.B., Xu, R., Socher, R. and Xiong, C.M. (2021) WOAD: Weakly Supervised Online Action Detection in Untrimmed Videos. *CVPR*, Virtual, 19-25 June 2021, 1915-1923.

[32] Hu, X.J., Dai, J.Z., Li, M., Peng, C.L., Li, Y. and Du, S.D. (2022) Online Human Action Detection and Anticipation in Videos: A Survey. *Neurocomputing*, **491**, 395-413. https://doi.org/10.1016/j.neucom.2022.03.069