

基于区域时空二合一网络的动作检测方法

汤 强, 朱 煜, 郑兵兵, 郑 婕

(华东理工大学信息科学与工程学院, 上海 200237)

摘要: 视频动作检测研究是在动作识别的基础上, 进一步获取动作发生的位置和时间信息。结合 RGB 空间流和光流时间流, 提出了一种基于 SSD 的区域时空二合一动作检测网络。改进了非局部时空模块, 在光流中设计了像素点筛选器来提取运动关键区域信息, 只对空间流中筛选出的动作关键区域进行相关性计算, 有效获得动作长距离依赖并改善非局部模块计算成本较大的缺陷, 同时降低了视频背景噪声的干扰。在基准数据集 UCF101-24 上进行了实验, 结果表明所提出的区域时空二合一网络具有更好的检测性能, 视频级别的平均精度 (video_mAP) 达到了 43.17%@0.5。

关键词: 视频动作检测; SSD; 双流网络; 非局部模块; UCF101-24

中图分类号: TP391.4

文献标志码: A

近年来, 随着互联网技术的发展和数码设备的不断普及, 视频数据呈爆发式增长, 视频智能分析成为当前计算机视觉领域的研究热点, 在人机交互系统、自动驾驶、视频监控和城市安全系统等领域具有广泛的应用。其中, 动作检测是视频分析的主要任务和难点之一。动作检测是在动作识别^[1-2]的基础上, 对视频中发生的动作分类后, 进一步定位动作发生的空间位置及动作开始和结束的时间。

早期较为成功的动作检测算法^[3-5]主要是通过滑动窗搜索法、特征表示以及 SVM 等方法来实现。后来, 受到动作识别中密集轨迹 (Dense Trajectories, DT) 算法的启发, 研究者们开始使用 DT 算法在时间维度上进行动作定位^[6-7]。Van 等^[8]进一步将 DT 算法从时序动作检测扩展至时空动作检测, 取得了一些进展。但这些早期的方法主要通过人工设计特征, 泛化能力和整体性能都较为有限。

随着深度学习在目标检测领域的广泛应用, 一些研究者们认为动作检测和目標检测具有一定的相似性, 所以尝试将 Faster-RCNN 和 SSD 等目标检测网络迁移到动作检测任务中。同时, 由于双流卷积神经网络能有效结合视频的空间和时序信息, 在动

作识别任务中具有优异的性能, 目前许多动作检测网络也沿用双流网络的框架形式。其中, Peng 等^[9]将 Faster-RCNN 和双流网络结合, 提出了基于 Faster-RCNN 的端到端双流动作检测网络。使用 Faster-RCNN 对视频中的每一帧 RGB 图像和光流图进行目标检测, 然后通过兴趣区域和置信度融合的方式将空间流和时间流的帧级别检测结果进行合并, 最后通过动作管道生成算法将相邻帧之间的检测结果匹配连接, 生成最终的时空动作管道。Singh 等^[10]使用双流 SSD 检测器, 并结合实时光流算法实现更快速的动作检测。然而, 相较于图片而言, 视频具有很强的时间相关性, 内容随时间不断变化。上述双流法虽然融入了时序信息, 但无法获取长距离的时序相关性, 对于一些时间跨度比较长的动作实例表现欠佳。同时, 视频往往包含大量背景信息, 使得对视频进行特征表示时, 存在大量的冗余信息, 对分类和检测结果产生干扰。因此, 有效利用视频中运动特征明显的关键区域信息, 构建长距离时序相关性, 对动作检测十分重要。

针对以上问题, 本文提出了一种基于 SSD 的区域时空二合一动作检测网络, 引入非局部机制 (Non-

收稿日期: 2020-11-26

作者简介: 汤 强(1995—), 男, 江苏宿迁人, 硕士生, 主要研究方向为视频理解、图像处理。E-mail: TQ1508420095@163.com

通信联系人: 朱 煜, E-mail: zhuyu@ecust.edu.cn

local)^[11]并加以改进,在获取长距离时序相关性的同时,通过时间流特征图中的关键动作区域信息对空间流特征图进行筛选,增强网络对于动作主体特征的学习,减少背景区域对分类和检测结果的影响。实验结果表明本文方法能有效提高动作检测的精度。

1 区域时空二合一网络模型

基于 SSD 的区域时空二合一动作检测网络结构如图 1 所示。

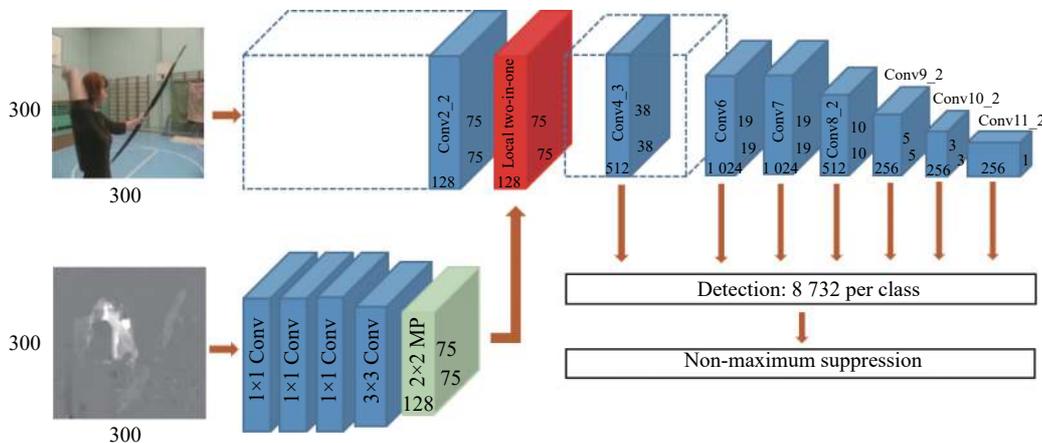


图 1 区域时空二合一网络结构图

Fig. 1 Region spatiotemporal two-in-one network structure

图 1 中蓝色方块表示卷积层,绿色方块表示最大池化层,红色方块表示区域时空二合一模块。上支路的输入为 300×300 的单帧 RGB 图像,下支路输入为 300×300 的单张光流图。首先,RGB 图像经过 SSD 的 Conv1 和 Conv2 卷积层,得到大小为 75×75 、通道数为 128 的空间流特征图。光流图则分别经过 3 层 1×1 卷积层、1 层 3×3 卷积层以及 1 层 2×2 最大池化层得到大小为 75×75 、通道数为 128 的时间流特征图。然后,将这两组尺寸相同的特征图同时输入到区域时空二合一模块,实现动作目标特征增强和时空特征融合。通过区域时空二合一模块后,时空混合特征图经过 SSD 剩余的部分进行边界框回归和分类,并得到帧级别的动作检测结果。最后,帧级别的动作检测结果经过动作管道生成器进行匹配和连接,最终得到视频级别的动作检测结果。

1.1 双流融合

与以往通过改变网络结构去学习动作的时空特征不同,Simonyan 等^[12]从输入数据角度入手,结合视频相邻帧之间运动信息可以通过光流表达的特点,设计了一种双流卷积神经网络用于视频动作识别,但所提出的双流网络的时间流和空间流相对独立,导致彼此无法共享学习的动作信息。针对这一问题,Feichtenhofer 等^[13]提出了一种新的双流融合技术,将原始双流中网络末端的结果融合改进为网络中段的特征图融合。这种特征图融合方式可以在像

素级别上连接包含动作空间信息的特征图和包含动作时间信息的特征图,使得后续的卷积网络可以同时学习动作的空间特征和时序特征。与文献 [14] 相似,本文也借鉴这一网络中段双流融合思路,但与文献 [14] 直接将两流特征图相加不同,本文设计了区域时空二合一模块来获取更好的融合特征。

1.2 区域时空二合一模块

长距离时序相关性的建立对于动作检测任务十分重要。文献 [11] 提出了一类新的神经网络,通过非局部运算来捕捉长距离相关性,有利于时序问题的处理,本文将其引入到动作检测任务中。非局部操作可以增强特征图中每个像素点之间的相互关系,使得动作目标能够获取动作整体相关性,即更好的时序信息。但是非局部操作需要计算全局的像素点,计算成本较大。同时视频背景像素加入全局计算,将增加噪声,对结果产生干扰。因此,本文对非局部加以改进,设计了像素标记筛选器 M_S ,只对动作特征明显的目标区域进行相关性计算,从而在获取长距离时序相关性的同时,减少网络计算成本并且排除背景的干扰。

一般来说,视频中的背景区域在相邻帧之间几乎不会发生变化,而产生动作的主体区域在相邻帧之间时常会发生明显的位移或角度变化。在光流图中,像素值较大的部分通常为动作目标区域,因此,本文结合光流图的数据特征对非局部机制进行改

进, 提出了一种区域时空二合一模块, 其结构如图 2 所示。主要分为 3 个部分: 像素点筛选、区域相关性计算、像素点填充。

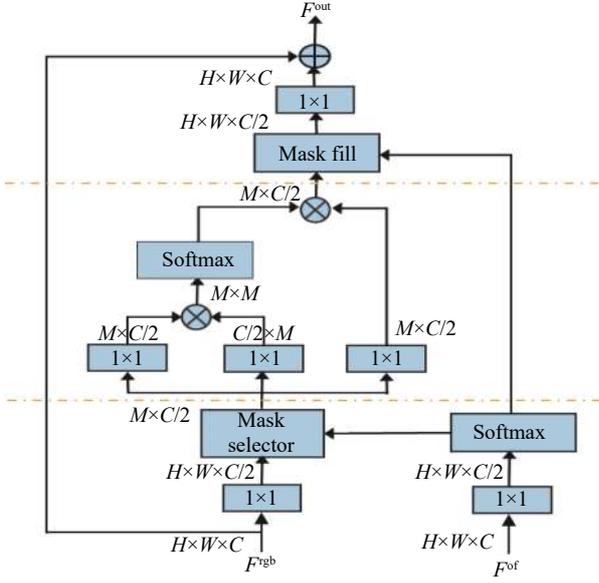


图 2 区域时空二合一模块结构

Fig. 2 Region spatiotemporal two-in-one module structure

1.2.1 像素点筛选 像素点筛选可以表示为

$$s = M_S \{ \delta(F^{rgb}), \text{softmax}(\varphi(F^{of})), \varepsilon \} \quad (1)$$

其中: F^{rgb} 表示空间流特征图; F^{of} 表示时间流特征图。像素点筛选的流程如下: 首先, F^{rgb} 和 F^{of} 分别经过 1×1 的卷积操作 δ 和 φ 缩减通道维度到 128, 减少计算量; 然后, 时间流特征图 $\varphi(F^{of})$ 经过 softmax 将像素值归一化到 0~1 之间, 并与空间流特征图 $\delta(F^{rgb})$ 一同输入标记筛选器 M_S 。

M_S 的原理如图 3 所示。图 3 中上半部分为空

间流特征图 $\delta(F^{rgb})$, 下半部分为归一化后的时间流特征图 $\text{softmax}(\varphi(F^{of}))$ 。对于时间流特征图中的像素点, 若像素值大于阈值 ε , 则在该点的相应位置打上标记。其中 ε 的取值通过特征图的大小设定, 例如对于 28×28 的特征图, ε 的取值为 $1/(28 \times 28) = 0.00128$ 。将 1 设为被除数是因为时间流特征图经过 softmax 后像素值的总和为 1, 因此通过阈值设定可以将平均值以上的像素点标记出来。然后将这些标记信息映射到空间流特征图上, 筛选出空间流特征图中相对应位置的像素点。因此, 经过标记筛选器后得到的输出 s 的大小为 $M \times C/2$, 其中 M 为空间流特征图中筛选出的像素点, $C/2$ 为输出特征图的通道数。

1.2.2 区域相关性计算 区域相关性计算方式与非局部机制的非局部计算相同, 只是输入数据的大小有所区别, 不再是特征图的所有像素点间计算, 而是对由 M_S 筛选出的关键动作区域的像素点间计算。区域相关性的计算公式如 (2) 所示:

$$y = \text{softmax}(s^T W_\theta W_\phi s) g(s) \quad (2)$$

其中: $W_\theta W_\phi$ 为网络要学习的参数; $g(s)$ 为 s 在 g 空间的映射, 本文取 $g(s)$ 为特征图本身。

1.2.3 像素点填充 计算相关性之后, 再通过标记的位置将像素值逐一放回筛选之前的特征图中, 如式 (3) 所示:

$$z = M_F \{ y, \text{softmax}(\varphi(F^{of})), \varepsilon \} \quad (3)$$

最后式 (4) 通过 1×1 的卷积 ω 还原原始模块输入的通道数, 并通过残差连接与原始空间流输入相加, 得到最终的输出结果 F^{out} :

$$F^{out} = \omega(z) + F^{rgb} \quad (4)$$

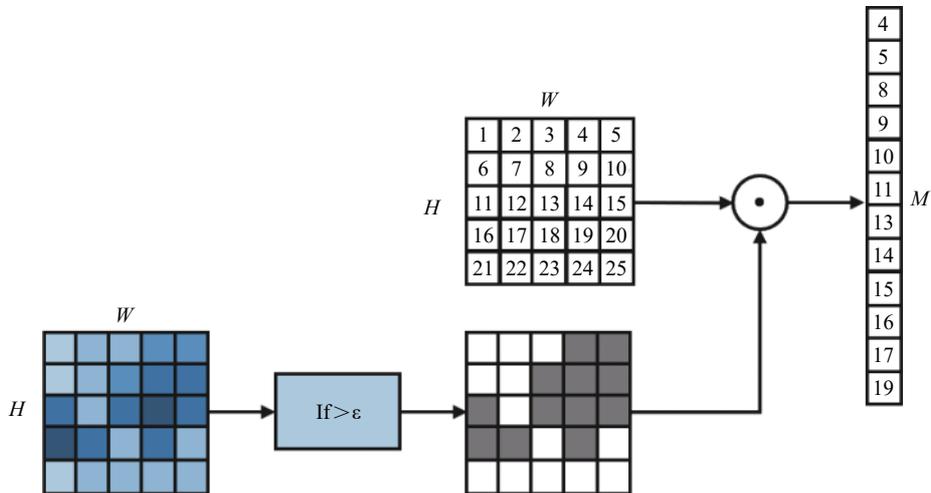


图 3 标记筛选器原理图

Fig. 3 Schematic diagram of mark selector

区域时空二合一模块的核心点在于该模块运用时间流提取的特征图中包含动作目标区域信息的特点。设计 M_S 将空间流特征图中该动作关键区域的像素点筛选出来单独进行相关性计算,从而在获取长距离时序相关性的同时,改善非局部模块计算成本较大的缺陷,并大大降低了视频背景噪声的干扰。

1.3 动作管道生成

动作检测任务的最终目的是得到一组连续检测集合,即动作管道 $T_c = \{b_{1s}, \dots, b_{1e}\}$ 。其中, c 表示动作管道的类别, b_{1s} 表示动作管道开始的帧级别检测框, b_{1e} 表示动作管道结束的帧级别检测框。本文借鉴文献 [10] 中的动作管道生成算法,使用一种简单而有效的管道生成方式为每个动作类别并行地构建多个动作管道。

该算法的输入是具有特定类别置信度的帧级别检测框。在时刻 t , 对于类别 c 的前 n 个检测框 $\{b_c^j, j=1, \dots, n\}$, 算法通过选取一个最佳匹配检测框来逐渐增加动作管道的长度。生成的动作管道需要满足以下条件: (1) 同一动作管道连续帧的检测结果在空间上的重叠率必须大于阈值 λ ; (2) 每个帧级别检测框只能属于一个动作管道。

2 实验结果与分析

2.1 实验数据集

本文采用动作检测基准数据集 UCF101-24 进行实验,该数据集是动作识别基准数据集 UCF-101 的子集。为了适合动作检测任务,研究者们从 UCF-101 中剔除面部表情等部分动作类别,并进行标签补充,最后保留 24 个动作类别。UCF101-24 中的视频数据不仅有类别标签,还有动作目标的空间坐标,以及动作开始和结束的时间。UCF101-24 数据集中每个视频仅包含一个动作类别,但可能包含多个动作实例。一般每个视频有 1.5 个动作实例,每个动作实例覆盖视频时长的 70%,而对于某些类别,动作实例的平均覆盖时长可能低至 30%。

2.2 评价指标

借鉴目标检测的评价指标,动作检测也使用平均精度(mean Average Precision, mAP)作为算法的主要评价标准。根据帧级别的检测结果和视频级别的检测结果,评价指标分为帧级别的平均精度(frame_mAP)和视频级别的平均精度(video_mAP)。

frame_mAP 为数据集中所有类别 frame_AP 的平均值,计算公式如下:

$$\text{frame_mAP} = \frac{\sum_{n=1}^N \text{frame_AP}(n)}{N} \quad (5)$$

其中: N 为动作类别总数; \sum 表示求和。frame_mAP 与目标检测中 AP 的计算方式相同,为计算精确率-召回率曲线下的面积大小。

video_mAP 为检测算法最后生成的动作管道准确率的均值,只有当组成管道的所有帧的检测区域与真实区域的平均重叠率大于阈值 σ 且动作管道的类别分类准确时,才认为这段动作管道为真实正样本。

2.3 实验环境与参数设置

实验环境为 NVIDIA GTX 1080Ti GPU、Ubuntu 16 操作系统、Pytorch0.4.0 和 Python3.6.4。

网络的输入数据为 $3 \times 300 \times 300$ 的单张 RGB 视频帧和 $2 \times 300 \times 300$ 的单张光流图。每一次送入网络的批次大小为 8,初始学习率为 0.0001。每次训练迭代 120000 次,并在迭代 70000 次和 90000 次时进行学习率衰减,学习率的衰减率为 0.5。

2.4 实验结果

2.4.1 UCF101-24 上各类别的 Frame-AP 表现 表 1 示出了 SSD 动作检测网络^[10] 和区域时空二合一网络在数据集具体 24 个动作类别的 frame_AP 对比结果。统计了 IoU 阈值为 0.5 时各类别的 frame_AP,其中最后一行是所有类别的 frame_AP 取平均得到的 frame_mAP, Δ 表示区域时空二合一网络相对于基础网络的提升。

与 SSD 动作检测网络相比,区域时空二合一网络在“打篮球”、“滑雪”等 18 个动作类别的帧级别检测精度均有所提升,“滑雪”动作的检测精度增幅高达 9.64%,每个类别平均增幅达到 2.61%。而“扣篮”等 6 个动作的检测精度虽然有所下降,但幅度均小于 0.3%。性能提升的 18 个动作类别都属于包含大量背景信息的动作实例,并且动作跨时也比较长,表明本文方法有效建立了长距离时序相关性,同时降低了背景的干扰。其中,“滑雪”动作数据集示例如图 4 所示,可以看到数据集中的滑雪动作并不都是左上角的只有白雪覆盖的背景,相反,数据集中大部分都伴随着山峰、树木或赛道背景等噪声干扰。而对于“扣篮”动作的检测性能稍微下降,主要是因为“扣篮”动作持续时间较短,往往只有一、两秒,而且动作主体集中在体积较小的篮球上,而“打篮球”则包括前期运球、准备投篮、投篮、篮球进框等多个步骤,持续时间较长,并且动作主体集中于人。因此,整体而言,区域时空二合一网络可以有效提高动作检测

表 1 UCF101-24 数据集中各类别在 IoU 阈值为 0.5 时的 frame_AP(%) 对比结果

Table 1 Comparison of frame_AP of UCF101-24 at IOU threshold of 0.5

Class	frame_AP/%		Δ (diff)
	SSD	This paper	
Basketball	28.91	32.37	3.46
Basketball_dunk	49.90	49.61	-0.29
Biking	78.36	78.27	-0.09
Cliff_diving	50.19	57.95	7.76
Crick_bowling	27.68	31.44	3.76
Diving	78.97	80.70	1.73
Fencing	87.95	88.16	0.21
Floor_gymnastics	83.38	85.44	2.06
Golf_swing	43.44	44.83	1.39
Horse_riding	88.57	88.41	-0.16
Ice_dancing	71.61	72.40	0.79
Long_jump	56.77	59.44	2.67
Pole_vault	55.04	56.72	1.68
Rope-climbing	81.36	82.12	0.76
Salsa_spin	69.26	69.01	-0.25
Skate_boarding	68.63	71.71	3.08
Skiing	68.09	77.73	9.64
Skijet	84.44	87.45	3.01
Soccer_juggling	79.97	80.14	0.17
Surfing	82.88	86.50	3.62
Tennis_swing	37.26	37.18	-0.08
Trampoline_jumping	60.63	60.54	-0.09
Volleyball_spiking	35.51	36.50	0.99
Walking_with_dog	74.26	74.44	0.18
frame_mAP	64.29	66.21	1.92



图 4 “滑雪”动作数据集示例

Fig. 4 Example of action Skiing in dataset

精度。

2.4.2 与其他动作检测算法的 video_mAP 对比 当 IoU 阈值分别为 0.2、0.5、0.75 和 0.50:0.95 时, video_mAP 值如表 2 所示。其中, 0.5:0.95 为阈值 0.5 至 0.95 每隔 0.05 测得的平均值。

表 2 动作检测算法在 UCF101-24 数据集上的 video_mAP 结果对比

Algorithm	video_mAP/%			
	0.20	0.50	0.75	0.50:0.95
Literature [9]	71.80	35.90	1.60	8.80
Literature[10]	69.8	40.9	15.5	18.7
Literature [15]	66.70	35.90	7.90	14.40
Literature [16]	73.5	37.8	-	-
Literature [17]	71.53	40.07	13.91	17.90
Literature [18]	56.7	36.6	-	-
Literature [19]	72.9	41.4	-	-
This paper	74.22	43.17	14.82	19.05

可以看出, 与近年来提出的动作检测网络相比, 本文提出的区域时空二合一网络在不同阈值下的 video_mAP 均有所提高。

2.4.2 UCF101-24 可视化与检测结果示例 为了验证本文提出的区域时空二合一网络的性能, 对区域时空二合一模块的输入特征图和输出特征图进行可视化, 并给出了 UCF-24 数据集中部分示例检测结果。可视化结果如图 5 所示。图中的第 1 行是输入模块中的时间流特征图, 第 2 行是输入模块中的空间流特征图, 第 3 行是模块输出的时空二合一特征图。

从第 1 行的时间流数据可以看出, 原始输入光流图在经过若干层卷积之后可以得到动作目标区域明显的时间流特征图。对比使用时间流特征图筛选前后的第 2 行和第 3 行特征图, 尤其是最后一列, 不难发现, 经过区域时空二合一模块后的空间流特征图中背景噪声受到了抑制, 最后输出的时空二合一特征图中动作目标区域得到增强和突出, 更为明显, 如图中做运动的男子及其手部和腿部变化。

图 6 示出了 UCF-24 数据集中部分示例的检测结果, 绿色字体代表预测的动作类别, 红色框为预测的动作区域, 黄色框为真实的动作区域。示例中黄色框和红色框的重合比例很高, 表明本文提出的网络具有较好的检测性能。

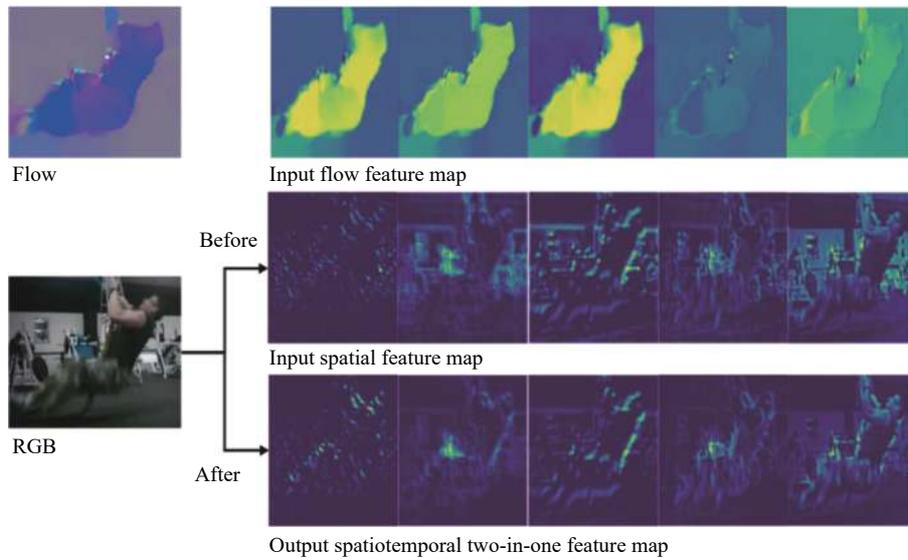


图 5 区域时空二合一模块前后特征图可视化结果示例

Fig. 5 Example of visualization of the region spatiotemporal two-in-one module



图 6 区域时空二合一网络对于 UCF-24 数据集中部分示例检测结果

Fig. 6 Detection results of proposed network for ucf-24 dataset

3 结束语

本文通过对非局部模块进行分析和改进,在光流中设计了像素点筛选器来提取运动关键区域信息,再融入空间流中。提出了一种区域时空二合一模块,并借鉴双流融合的原理将区域时空二合一模

块应用于 SSD 检测框架中。提出了一种新型的动作检测网络——区域时空二合一网络。最后通过基准数据集 UCF101-24 进行网络训练和性能验证。实验证明区域时空二合一网络在获取长距离时序相关性的同时,改善非局部模块计算成本较大的缺陷,并大大降低了视频背景噪声的干扰,拥有更好的检测性能。

参考文献:

- [1] 黄晴晴,周凤余,刘美珍. 基于视频的人体动作识别算法综述[J]. 计算机应用研究, 2020, 37(11): 3213-3219.
- [2] 朱煜,赵江坤,王逸宁,等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报, 2016, 42(6): 848-857.
- [3] TIAN Y, SUKTHANKAR R, SHAH M. Spatiotemporal deformable part models for action detection[C]//the IEEE International Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2013: 2642-2649.
- [4] TRAN D, YUAN J. Max-margin structured output regression for spatio-temporal action localization[C]//Advances in Neural Information Processing Systems. USA: NIPS, 2012: 350-358.
- [5] YUAN J, LIU Z, WU Y. Discriminative video pattern search for efficient action detection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, 33(9): 1728-1743.
- [6] GAIDON A, HARCHAOU Z, SCHMID C. Temporal localization of actions with actoms[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013, 35(11): 2782-2795.
- [7] ONEATA D, VERBEEK J J, SCHMID C. Efficient action localization with approximately normalized fisher vectors[C]//the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2014: 2545-2552.
- [8] VAN G J C, JAIN M, GATI E, *et al.* APT: Action localization proposals from dense trajectories[C]//the British Machine Vision Conference. UK: BMVA Press, 2015: 1-12.
- [9] PENG X, SCHMID C. Multi-region two-stream R-CNN for action detection[C]//European Conference on Computer Visio. Amsterdam, Netherlands: Springer, 2016: 744-759.
- [10] SINGH G, SAHA S, SAPIENZA M, *et al.* Online real-time multiple spatiotemporal action localisation and prediction[C]//the IEEE International Conference on Computer Vision. Italy: IEEE, 2017: 3637-3646.
- [11] WANG X, GIRSHICK R, GUPTA A, *et al.* Non-local neural networks[C]//the IEEE Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2018: 7794-7803.
- [12] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[C]//Advances in Neural Information Processing Systems. Canada: NIPS, 2014: 568-576.
- [13] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]//the IEEE International Conference on Computer Vision and Pattern Recognition. USA: IEEE, 2016: 1933-1941.
- [14] 杨天明,陈志,岳文静. 基于视频深度学习的时空双流人物动作识别模型[J]. *计算机应用*, 2018, 38(3): 895-899, 915.
- [15] SAHA S, SINGH G, SAPIENZA M, *et al.* Deep learning for detecting multiple space-time action tubes in videos[C]//International Computer Vision Summer School. Italy: ICVSS, 2016: 1-13.
- [16] YANG Z H, GAO J Y, NEVATIA R. Spatio-temporal action detection with cascade proposal and location anticipation[C]//British Machine Vision Conference. UK: BMVC, 2017: 1-12.
- [17] BEHL H S, SAPIENZA M, SINGH G, *et al.* Incremental tube construction for human action detection[C]//British Machine Vision Conference. Newcastle, UK: BMVC, 2018: 1-12.
- [18] SONG Y, KIM I. Spatio-temporal action detection in untrimmed videos by using multimodal features and region proposals[J]. *Sensors*, 2019, 19(5): 1085-1103.
- [19] ALWANDO E H P, CHEN Y T, FANG W H. CNN-Based multiple path search for action tube detection in videos[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020, 30(1): 104-116.

Action Detection Based on Region Spatiotemporal Two-in-One Network

TANG Qiang, ZHU Yu, ZHENG Bingbing, ZHENG Jie

(School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China)

Abstract: With the explosive growth of video data, video intelligent analysis has become the current academic and industrial research hotspot. Video Action Detection is to obtain the location and time information of actions based on Action Recognition. In this paper, we propose a region spatiotemporal two-in-one action detection network based on single shot multi-box detector(SSD) with the combination of RGB space flow and optical flow. To design the improved

nonlocal spatiotemporal module in the network, a pixel filter is proposed in optical flow to extract the information of key motion regions, and then the correlation calculation is only performed on the selected key motion regions in the spatial flow. The proposed module can get long-range dependence of actions effectively and improve the computational cost of the nonlocal module, and reduce the interference of video background noise. The proposed network is tested on the benchmark dataset UCF101-24, and the video_map reaches 43.17% @ 0.5. The results show that the proposed network has better performance.

Key words: video action detection; SSD; two-stream network; non local model; UCF101-24