计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(10):109−117 [doi: 10.15888/j.cnki.csa.008122]

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

基于 3D-SVD 的时空行为定位算法①

王紫烟[1], 张立华[1,2,3,4], 翟　鹏[1,5], 杜洋涛[1]

[1](复旦大学 智能机器人研究院, 上海 200433)
[2](季华实验室, 佛山 528200)
[3](智能机器人教育部工程研究中心, 上海 200433)
[4](吉林省人工智能与无人系统工程研究中心, 长春 130012)
[5](上海智能机器人工程技术研究中心, 上海 200433)
通讯作者: 张立华, E-mail: lihuazhang@fudan.edu.cn

摘　要: 随着摄像头的普及, 基于人工智能的行为分析技术在智能视频领域扮演着越来越重要的角色. 现有的算法大多采用光流网络或者 3D 网络来获取行为的时间信息, 但是光流网络和一般的 3D 网络计算量大, 在同时进行分类和定位这两项任务时计算效率低. 针对这一问题, 本文构建了一个能够进行空间定位和分类的双流框架, 在 3D 网络分支中采用 SVD 的思想分解 3D 卷积核以减少 3D 网络的参数, 并利用动态规划算法高效的搜索最佳行为管道, 在训练的过程中采用 mixup 算法对数据集进行扩充, 增强训练的效果. 最后, 在 UCF101-24 和 J-HMDB-21 这两个被广泛使用的行为定位数据集上进行了实验验证, 相比于基线算法, 两个数据集的 Frame-mAP 分别提高了 7.1% 和 4.8%, 其中, J-HMDB-21 在不同 *IOU* 下的 Video-mAP 分别提高了 5.2% 和 4.8%. 实验结果表明, 本文提出的算法能有效提高行为定位能力, 与其它算法相比获得了较好的结果.

关键词: 行为定位; SVD; 数据增强; 行为管道

引用格式: 王紫烟,张立华,翟鹏,杜洋涛.基于 3D-SVD 的时空行为定位算法.计算机系统应用,2021,30(10):109–117. http://www.c-s-a.org.cn/1003-3254/8122.html

Spatio-Temporal Action Localization Algorithm Based on 3D-SVD

WANG Zi-Yan[1], ZHANG Li-Hua[1,2,3,4], ZHAI Peng[1,5], DU Yang-Tao[1]

[1](Institute of AI and Robotics, Fudan University, Shanghai 200433, China)
[2](Ji Hua Laboratory, Foshan 528200, China)
[3](Engineering Research Center of AI and Robotics, Ministry of Education, Shanghai 200433, China)
[4](Engineering Research Center of AI and Unmanned Vehicle Systems of Jilin Province, Changchun 130012, China)
[5](Shanghai Engineering Research Center of AI and Robotics, Shanghai 200433, China)

Abstract: With the popularity of video surveillance, action analysis technology based on artificial intelligence is playing an increasingly important role in the field of intelligent surveillance. Most of the existing algorithms depend on an optical flow network or a 3D network to obtain the time information of actions. However, the optical flow network and the general 3D network require a large amount of computation, and the computational efficiency is low when classification and localization are carried out simultaneously. To solve this problem, this study builds a dualflow framework capable of spatial localization and classification and follows the idea of SVD to decompose the 3D convolution kernel in the branch of the 3D network, thus reducing the 3D network parameters. In addition, the dynamic programming algorithm is employed to efficiently search the optimal action tubes, and the mixup algorithm is used to expand the data sets during training, thereby enhancing the training results. Finally, experimental verification is carried out on UCF101-24 and

① 基金项目: 上海市科委项目 (19511132000)
Foundation item: Project of Science and Technology Commission of Shanghai Municipality (19511132000)
收稿时间: 2021-01-06; 修改时间: 2021-02-07; 采用时间: 2021-02-23

J-HMDB-21, two widely used data sets for action localization. Compared with the baseline algorithm, the Frame-mAP of the two data sets is improved by 7.1% and 4.8%, and the Video-mAP of J-HMDB-21 under different *IoU*s is enhanced by 5.2% and 4.8%. Experimental results show that the proposed algorithm can substantially improve the ability of action localization, with better results compared with other algorithms.

Key words: action localization; SVD; data augmentation; action tubes

时空行为定位技术是一种针对目标行为的智能视频分析技术, 即在视频帧进行行为分类并产生与行人空间位置相关的定位框序列. 旨在不需要人为干预的情况下, 利用计算机视觉和视频分析方法对摄像头下的人类行为或视频网站上的视频进行自动分析, 在智能体育和智能监控领域有着广泛的应用.

时空行为定位技术在时间和空间上同时定位人体的行为, 这在计算机视觉领域是一项非常重要的任务. 为了解决这个任务, 早期 Weinzaepfel 等[1] 采用 CNN 网络和时空运动直方图描述符对轨迹进行评分来确定行为的空间位置, 并采用多尺度滑动窗口进行时间定位. 在双流网络和 3D-CNN 网络应用于行为识别任务之后, 由于其优秀的性能, 此后的时空行为定位工作大多基于这两个框架进行. 这些算法通常又包括两个部分: 生成帧级动作和生成帧间关联. 为了生成更准确的帧级动作, Peng 等[2] 将 RPN 扩展到光流数据上训练运动的 RPN, 以此来提高帧级行为检测的性能. Yang 等[3] 提出了级联方案, 采用级联区域提议网络生成帧级动作.

帧间关联一般又称为行为管道的构建, 多采用贪婪算法和动态规划算法. 为了更高效地找到行为管道的多重路径, Alwanda 等[4] 开发了一种低成本的动态规划算法并利用相邻帧的时间一致性修正了不准确的行为边界框. 此外, 因为行为管道可以利用视频帧的时间连续性特征定位行为的时空位置, 所以为了更好的利用这一特性, Hou 等[5] 在 3D-CNN 的基础上提出了一个 tol 池化层, 缓解了行为管道上时空的再变化问题. Kalogeiton 等[6] 则对每个行为框进行精修来得到更准确的行为管道. Li 等[7] 利用 LSTM 结构捕捉时间信息, 并使用维特比算法连接每一帧的行为框形成行为管道. 在以上工作的基础上, He 等[8] 构建了一个新的行为定位框架, 并利用 TPN 生成了通用的行为管道.

以上的工作在行为定位任务上均有着不错的表现, 但是其中很多工作基于双流网络或者 3D-CNN 进行, 双流网络中的光流网络分支计算效率低, 而 3D-CNN 参数量大, 在一定程度上也存在着精度与计算效率之间的矛盾. 为了缓解该矛盾, Qiu 等[9] 提出了 P3D 网络, 在这个网络中用二维空间卷积和一维时间卷积来模拟 3D 卷积以降低参数量. 在 P3D 的基础上, Tran 等[10] 做了大量的实验探索类似的架构, 并将其重新演绎为 (2+1)D.

本文为了解决 3D-CNN 计算量大的问题, 从 3D 卷积核自身出发, 将二维层面的 SVD 思想扩展到 3D-CNN 中得到 3D-SVD, 有效的降低了 3D 卷积网络的参数量, 并基于 3D-SVD 提出了一个时空行为定位网络框架. 首先, 在数据集的处理上, 我们加入了 mixup 算法进行数据增强, 丰富了数据集的内容. 其次, 我们构建双流网络架构对行为进行识别并定位, 采用空间定位网络和时空特征提取网络融合的方式, 并使用 3D-SVD 对三维卷积网络进行优化. 最后, 采用序列重排序算法和动态规划算法对行为管道进行构建, 可以有效降低行为的空间漂移对定位结果的影响. 根据实验结果表明, 我们的网络在两个公开的数据集上指标都有所提升.

1 相关工作

本文的主要研究内容包括时空行为定位, 行为管道构建和数据增强 3 个部分. 时空行为定位和行为管道构建可以对视频中的行为进行定位和分类. 而时空行为定位网络需要大量的视频数据来进行训练, 因此数据增强也是时空行为定位任务中常见的子任务.

1.1 数据增强

数据增强是一种数据扩增技术, 可以在有限的数据集上进行扩充得到更多的数据来帮助训练. 常用的数据增强技术有图像翻转、裁剪、缩放等几何层面上的增强方式, 也有增加噪声、进行填充、颜色变换等颜色层面上的增强方式, 这两种图像增强方式都是在单个图像上进行操作的图像增强技术. 除此之外, 还有在多个图像上进行操作从而产生新图像的图像增强技

术. 其中, SMOTE 算法[11] 利用插值来改变数据集的类不平衡现象, SamplePairing 算法[12] 将不同的图像分别进行处理后再叠加来得到新的样本. 近几年生成对抗网络[13] 逐渐兴起, 这种网络可以通过一个生成网络随机的生成图像, 再通过一个判别网络判断生成的图像是否“真实”. 这样通过网络的学习, 来随机生成与数据集分布一致的图像集合, 将有限的数据内容变得更加的丰富.

1.2 时空行为定位

时空行为定位任务可以同时完成行为检测和行为分类两个任务. 行为检测实质上是一个目标检测任务, 可以检测出目标行为在时空的具体位置. 一般目标检测任务是帧级层面上的检测任务, RCNN[14] 作为基于区域的检测算法, 使用选择搜索算法在图像上提取出可能包含物体的区域, 然后使用分类网络得到每个区域内物体的类别. 在此基础上, Faster RCNN[15] 提出了 RPN 代替了 RCNN 中的选择搜索算法, Fast-RCNN[16] 共享了卷积计算提高了特征的利用效率. 为了进一步提高目标检测算法的实时性, YOLO[17] 和 SSD[18] 将检测任务统一为一个端到端的回归问题, 目前的 YOLO 版本能够现阶段最优的检测结果. 在帧级目标检测的基础上, 时空行为定位任务还需要对行为进行时序上的检测, 大多采用构建行为管道的方法, 辅以双流网络[19] 和 3D-CNN[20] 网络来进行时空行为定位. 最近有工作将 2D 特征和 3D 特征构建双流网络[21], 进行了通道融合得到了很好的结果. 但上述时空行为定位方法采用的 3D-CNN 网络具有很大的参数量, 导致整体网络计算负担过大.

1.3 行为管道构建方法

构建行为管道即从视频片段每一帧检测到的一系列行为框中找到最优的行为框路径, 将其链接为行为管道, 行为管道的构建方法决定了时空行为定位的准确性. 行为管道构建实际上是一个最优路径搜索问题, 有学者采用贪婪算法[12] 增量的生成多个行为管道, 再利用动态规划的方法找出最优的行为管道. 还有学者采用维特比算法[7] 链接不断递增的行为框以此形成多通道的行为序列. 为了更高效地搜索到行为管道, 优化动态规划算法来增强整体算法效率也成了研究方向之一, 基于此有学者开发了一种低成本的能在单次运行中找到多重路径的算法[4]. 除此之外, HISAN[22] 在动态规划的基础上采用了 SR 算法减少了边界框在链接过程中遮挡和背景的影响, 并采用多路径搜索算法进行优化, 一次迭代就能找到所有可能的路径.

2 基于 3D-SVD 的行为定位算法

本文构建了一个端到端的框架, 可以定位视频中的多个行为, 在这个框架下可以同时提取到关键帧的二维特征和输入片段的三维特征. 基于 3D-SVD 的行为定位算法整体框架如图 1 所示, 分为 3 个主要部分: 空间定位网络、时空特征提取网络和行为管道构建. 接下来, 介绍本文框架的具体结构.

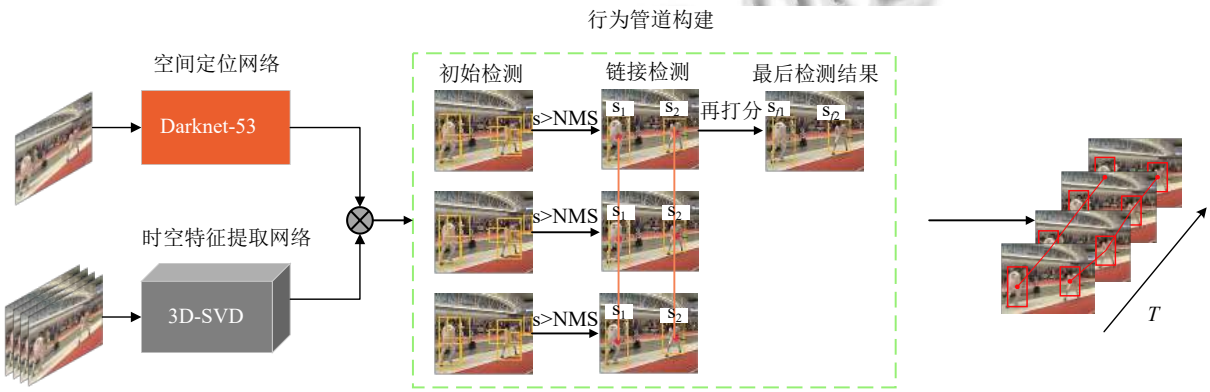

图 1 整体框架图

2.1 双流网络结构

空间定位网络和时空特征提取网络组成了一个双流网络. 空间定位网络分支采用 Darknet-53[23] 作为主干网络, 利用视频关键帧的二维特征来实现行为的空间定位. 时空特征提取网络分支在传统三维卷积网络的基础上采用 SVD 的思想, SVD 矩阵分解如图 2(a) 所示. 本文将 SVD 的矩阵分解思想扩展到三维层面, 将 3D 卷积核进行分解, 这样分解矩阵能够共享视频不

同维度的权值, 减少传统三维卷积网络的参数量, 我们将其称为 3D-SVD.

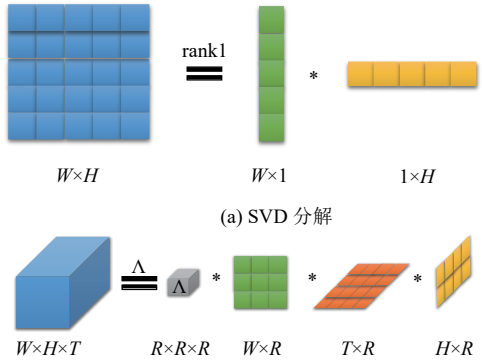

图 2 SVD 和 3D-SVD 分解对比

3D-SVD 将 3D 卷积核分解的过程可以用 Tucker 分解来表示, 如图 2(b) 所示. 假设 X 是大小为 $t\times h\times w$ 的三阶张量, 进行 Tucker 分解后为:

$$X \approx g\times T\times H\times W \tag{1}$$

其中, g 的大小为 $r_1\times r_2\times r_3$, T 的大小为 $t\times r_1$, H 的大小为 $h\times r_2$, W 的大小为 $w\times r_3$, 即:

$$x_{ijk}\approx\sum_{m=1}^{r_1}\sum_{n=1}^{r_2}\sum_{l=1}^{r_3}\left(g_{mnl}\cdot t_{im}\cdot h_{jn}\cdot w_{kl}\right) \tag{2}$$

当 g 为 $h\times t\times w$ 时, 可以得到视频的 3 个视角, 正常视角 W-H、沿着时间维度的高度信息视角 H-T 和沿着时间维度的宽度信息视角 W-T, 与 CoST 网络[24] 相似.

设输入的特征图的大小为 $T\times H\times W\times C_1$, C_1 是输入通道. 那么 3 个视角的输出特征图为:

$$\begin{cases} x_h = x\otimes w_{1\times3\times3} \\ x_w = x\otimes w_{3\times1\times3} \\ x_t = x\otimes w_{3\times3\times1} \end{cases} \tag{3}$$

然后将 3 组特征图进行加权求和:

$$y=[\alpha_{hw}\ \alpha_{tw}\ \alpha_{th}]\begin{bmatrix} x_h \\ x_w \\ x_t \end{bmatrix} \tag{4}$$

如式 (3) 所示, 3 个视角的卷积核共享权重, 3D-SVD 能够对视频 3 个视角的特征进行融合, 从而实现视频的行为分类.

3D-ResNeXt-101 网络[25] 在 Kinetics 数据集上获得了很好的表现, 因此将 3D-ResNeXt-101 网络作为时空特征提取网络的主干网络. 3D-ResNeXt-101 的网络结构如表 1 所示, 在此基础上将其中的中间层卷积替换为如图 3(b) 所示的 3D-SVD 结构即可有效的减少参数量, 并能得到所需要的时空信息.

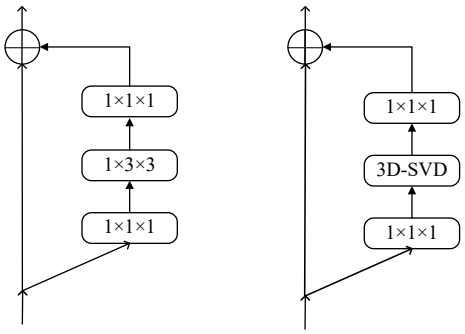

图 3 残差单元对比

表 1 3D-ResNeXt-101 结构

层的名称	输出大小	卷积结构
conv1	8×112×112	1×7×7, 64, 步数2
conv2	8×56×56	3×3×3最大池化, 步数2 $\begin{bmatrix} 1\times1\times1,128 \\ 1\times3\times3,128 \\ 1\times1\times1,256 \end{bmatrix}\times3$
conv3	8×28×28	$\begin{bmatrix} 1\times1\times1,256 \\ 1\times3\times3,256 \\ 1\times1\times1,512 \end{bmatrix}\times4$
conv4	4×14×14	$\begin{bmatrix} 1\times1\times1,512 \\ 1\times3\times3,512 \\ 1\times1\times1,1024 \end{bmatrix}\times23$
conv5	4×7×7	$\begin{bmatrix} 1\times1\times1,1024 \\ 1\times3\times3,1024 \\ 1\times1\times1,2048 \end{bmatrix}\times3$
平均池化		

在时空特征提取网络这个分支中, 输入是由一系列连续帧所组成的视频片段, 经过 3D-ResNeXt-101 网络后输出为 [帧数×高度×宽度×3] 的特征图, 为了和空间定位分支网络输出的特征图相匹配, 将时空特征提取网络输出的特征图的深度维数减少到 1.

2.2 基于序列重排序的行为管道构建

在进行定位和分类的过程之中, 行为的空间漂移会导致其定位精度的降低. 因此, 采用序列重排序算法, 可以减少运动漂移对检测的影响从而链接到行为管道的正确路径.

设分类分数为 $y_{i,j}=\left\{y_{i,j}^0,\cdots,y_{i,j}^{cls}\right\}$, cls 为需要识别的类别数量, $y_{i,j}^0$ 指第 i 帧没有动作存在的背景的分数. 我们用分类分数来调整行为类别的检测框分数:

$$s_c\left(x_{i,j}\right)=s_c'(x_{i,j})\times y_{i,j}^c \tag{5}$$

接着采用非最大值抑制算法将边界框的数量减少到$N_{nms}<N$, 将每一帧中这样得到的边界框与相邻帧中重叠最大的边界框链接到一起. 如果这两个框重叠的部分低于设置的阈值, 则终止链接. 我们采用下面的方式对最后一帧进行重新打分:

$$s_c(x_{T_s,j}) = \max\left(\frac{\sum_{i=T_s}^{T_e} s_c(x_{i+1,l_i})}{T_e - T_s + 1}, s_c(x_{T_s,J})\right) \tag{6}$$

其中,

$$l_i = \arg\max_l IoU(x_{i+1,l}, x_{i,l_{i-1}}), i \geq T_s \tag{7}$$

$$x_{T_s,l_{T_s}} = x_{T_s,j} \tag{8}$$

然后将每一帧产生的检测框链接在一起形成行为管道. 由上可知, 行为框的集合为$\{x_{i,j}\}$, 行为框检测分数的集合为$\{s_c(x_{i,j})\}$. 设时间 T 有 K 帧图像, 找到从 i=1 到 i=K 中的包含单个行为的一组行为框, 这组行为框组成一个行为管道. 行为管道的累积分数最大值为:

$$\sum_{T_1,\cdots,T_{2N_{nms}}} \sum_{i=1}^{K-1} A_c(x_{i,j}, x_{i+1,g}) \tag{9}$$

其中,

$$A_c(x_{i,j}, x_{i+1,g}) = s_c(x_{i+1,g}) + \alpha \times IoU(x_{i,j}, x_{i+1,g}) \tag{10}$$

式 (10) 中, α为权重参数, j 和 g 是路径$\{T_n\}_{n=1}^{2N_{nms}}$中不同的路径.

2.3 数据增强—mixup 算法

在实际的行为定位任务中, Okan 等[21] 采用了图像抖动、改变图像饱和度、色调和曝光度等技术对行为定位数据集的训练部分进行了图像增强. 这些图像增强操作可以有效的生成不同光照条件, 不同视角以及不同环境下的图像, 提高了训练的效果. 但是这些变换都是基于单个图像进行的操作. 行为定位任务实际环境复杂, 所以我们需要一种增加数据多样性的数据增强方法来增加算法的鲁棒性. 考虑到系统的效率, 我们增加了同样是利用了插值特性的 mixup 算法对已有的数据集进行进一步的图像增强操作. 这种算法是一种利用了线性插值增强新样本数据的数据增强方法, 基于领域风险最小化原则的数据增强方法.

设 x 为数据, y 为数据标签, $P(x,y)$为两者的联合分布, $l(\cdot)$为损失函数, 经验风险为:

$$R(f) = \int l(f(x), y) dP(x,y) \tag{11}$$

训练集$\{x,y\}$用狄拉克函数近似表示为:

$$P_\delta(x,y) = \frac{1}{n}\sum_{i-1}^{n} \delta(x = x_i, y = y_i) \tag{12}$$

$$R_\delta(f) = \int l(f(x), y) dP_\delta(x,y) = \frac{1}{n}\sum_{i=1}^{n} l(f(x_i), y_i) \tag{13}$$

数据量完成由 n 到 m 的增广之后, 式子变换为:

$$P_v(\tilde{x},\tilde{y}) = \frac{1}{n}\sum_{i=1}^{n} v(\tilde{x},\tilde{y}|x_i,y_i) \tag{14}$$

$$R_v(f) = \frac{1}{m}\sum_{i=1}^{n} l(f(\tilde{x}_i), \tilde{y}_i) \tag{15}$$

其中, v 代表邻域分布, mixup 采用了线性插值的方法, 用线性表达代表邻域分布.

$$\mu(\tilde{x},\tilde{y}|x_i,y_i) = \frac{1}{n}\sum_{j}^{n} E_\lambda[\delta(\tilde{x},\tilde{y})] \tag{16}$$

设(x_i,y_i)和(x_j,y_j)为在训练集中随机选取的两个样本点, 则插值产生的新的数据点为:

$$\tilde{x} = \lambda \cdot x_i + (1-\lambda)x_j \tag{17}$$

$$\tilde{y} = \lambda \cdot y_i + (1-\lambda)y_j \tag{18}$$

$$\lambda = Beta(\lambda, \beta),\ 0<\lambda<1 \tag{19}$$

3 实验结果及分析

为了评估算法的性能, 本节在 UCF101-24 和 J-HMDB-21 两个流行且具有挑战性的数据集上进行了训练和测试实验, 并将实验结果与其它算法的结果进行比较和分析.

本文的实验均在配有 4 张 32 GB 显存的 DGX STATION 上进行, 在 Linux 操作系统下使用 PyTorch 框架作为运行环境.

3.1 实验数据集和评价指标

UCF101-24 是 UCF-101 的子类数据集, 包含 24 种行为类别和 3207 个带有行为边界框标注的视频, 提供了行为的类别和空间标注信息.

J-HMDB-21 是 HMDB-51 的子类数据集, 包含 21 种行为类别和 928 个短视频. 在每个视频的所有帧中都有一个行为实例.

两个公开数据集中的数据来源于视频网站上真实场景下所拍摄的视频, 主要包含体育运动行为和日常生活行为, 如图 4 和图 5 所示.

评价指标: 在时空行为定位任务中最常用的评价指标为 Frame-mAP 和 Video-mAP, 前者是对于帧的度量, 代表了每帧检测的召回曲线下的区域, 后者是对于行为管道的度量, 视频平均每帧与真实标签的 *IOU* 超过了实验设定的阈值, 并且准确的预测了行为的类别, 则行为管道是正确的实例. 最后计算每个行为类别的平均精度.

图 4 UCF101-24 数据集

图 5 J-HMDB-21 数据集

3.2 实现细节

本文采用了双流网络的框架, 需要对提取空间信息的 2D 网络参数和提取时间信息的 3D 网络参数进行初始化. 采用了在 PASCAL VOC 上进行了预训练的 2D 模型和在 Kinetics 上进行了预训练的 3D 模型. 在本文的模型之中, 两个网络的参数可以进行联合更新. 设置初始的学习速率为 0.0001.

对于时空特征提取网络输入的视频剪辑长度, 因为长序列往往包含更多的时间信息, 因此采用 16 帧的剪辑长度, 并将下采样率设置为 1. 在模型进行训练之前, 除了采用 mixup 算法进行数据增强之外, 同时采用了图像水平翻转、随机剪裁、改变图像色调和饱和度这样基础的数据增强操作, 将图像统一随机缩放为 224×224 大小的图像输入网络.

3.3 双流网络消融实验

本文采用了双流网络框架. 在传统的双流网络中, 单独的二维卷积网络和单独的光流网络都无法对行为的时空信息进行准确的判断, 所以一般采用两个网络融合的结果. 为了对每条网络分支的作用做出更准确的判断, 本文设置了消融实验来判断双流结构是否能得到更好的结果.

实验采用 Frame-mAP、定位召回率和行为分类准确率 3 个指标. 其中, 定位召回率指的是正确定位的行为数与真实标签行为总数之比.

在两个数据集上分别进行的消融实验表明, 空间定位网络和时空特征提取网络进行融合后能得到更好的平均精准度, 比单独的时空特征提取网络提高了 9.5% 和 15.9%, 如表 2 和表 3 所示. 此外, 空间定位网络在定位上能得到更好的结果, 定位精度比时空特征提取网络高 3% 和 14.4%, 时空特征提取网络在行为分类上能得到更好的结果, 分类精度比空间定位网络高 8.3% 和 16.1%. 因此空间定位网络更关注空间特征, 时空特征提取网络更关注时间特征. 采用这两个网络融合的方法能更好的融合时空信息.

表 2 在 UCF101-24 上的实验结果 (%)

模型	Frame-mAP	定位召回率	分类准确率
空间定位网络	59.2	92.1	84.3
时空特征提取网络	72.6	89.2	92.6
双流网络	**82.1**	**92.8**	**94.1**

表 3 在 J-HMDB-21 上的实验结果 (%)

模型	Frame-mAP	定位召回率	分类准确率
空间定位网络	47.3	93.9	52.1
时空特征提取网络	54.6	79.5	68.2
双流网络	**70.5**	**95.4**	**71.0**

3.4 数据增强算法消融实验

为了更直观的判断 mixup 算法对于本文行为定位方法的影响, 进行了关于数据增强算法的消融实验, 结果如表 4 所示. 根据表 4 可知, 增加 mixup 算法能有效的扩充数据集, 使训练过程更加的有效, 得到更好的结果.

表 4 不同数据增强下的 Frame-mAP (%)

数据增强算法	UCF101-24	J-HMDB-21
基础数据增强	80.9	69.2
本文	**81.7**	**70.5**

3.5 比较实验

本小节比较了本文提出的算法和其它相关算法在 UCF101-24 和 J-HMDB-21 两个公开数据集上的 Frame-mAP 和 Video-mAP, 并在不同的 IOU 上进行了对比实验. 本小节对比的算法皆为近几年论文产出结果, 其中一部分方法在某些指标上拥有先进的结果[5,6,26], 一部分采用了与本文相似的双流网络结构和动态规划算法, 具有比较意义[2,4,27,28].

实验结果如表 5 所示, 对比两个数据集上的 Frame-mAP 指标, 本文提出的方法相对于之前的方法分别提升了 7.1% 和 5.8%, 具有良好的性能.

表 5 不同模型 Frame-mAP 对比 (%)

方法	UCF101-24	J-HMDB-21
TCNN[5]	41.4	61.3
ACT[6]	69.5	65.7
STEP[26]	75.0	—
Peng等[2]	64.8	56.9
本文	**82.1**	**70.5**

此外, 我们对比了两个数据集上的 Video-mAP 性能指标, 如表 6 和表 7 所示. 在 *IoU* 阈值分别为 0.2 和 0.5 的情况下, 本文提出的方法在 J-HMDB-21 数据集上总是优于当前的方法, 分别提高了 5.2% 和 5.3%, 另外在 UCF101-24 数据集上的改进稍逊色于在 J-HMDB-21 上的结果, 这是由于 J-HMDB-21 拥有更多相似子行为序列的行为类别. 根据目前的实验, 随着 *IoU* 数值的改变, Video-mAP 也会随之变化, 实验结果表明, 在 *IoU* 为 0.2 时, 能得到最好的结果.

表 6 不同模型在 UCF101-24 上 Video-mAP 对比 (%)

方法	*IoU*=0.2	*IoU*=0.5
TCNN[5]	47.1	—
ACT[6]	**77.2**	**51.4**
MPS[4]	72.9	41.1
Peng等[2]	73.5	32.1
Singh等[27]	73.5	46.3
Saha等[28]	66.6	36.4
本文	74.2	48.6

表 7 不同模型在 J-HMDB-21 上 Video-mAP 对比 (%)

方法	*IoU*=0.2	*IoU*=0.5
TCNN[5]	78.4	76.9
ACT[6]	74.2	73.7
Tpnet[29]	74.8	74.1
Peng等[2]	74.1	73.1
Singh等[27]	73.8	72.0
Saha等[28]	72.6	71.5
本文	**83.6**	**81.7**

3.6 结果可视化

最后, 对图 6 的时空定位网络可视化输出结果进行分析. 由图 6(a)–图 6(c) 可得, 本文所用的方法在背景简单的情况下可以准确的进行视频行为定位任务, 在视频序列中定位行为发生的空间位置并识别行为的类别. 图 6(d) 和图 6(e) 则表明, 面对同一类行为的时空定位, 在背景有与行为类别无关的行为发生时, 可能会产生误判的行为. 同时, 图 6(d) 和图 6(e) 与 Saha 等[28] 的可视化结果进行对比表明, 本文的方法在行为产生重叠的情况下也能得到准确的结果.

(a) 简单背景 1 (b) 简单背景 2 (c) 简单背景 3 (d) 复杂背景 1 (e) 复杂背景 2

图 6 定位和行为识别可视化

4 结论与展望

本文提出了一种基于 3D-SVD 的时空行为定位算法, 用于解决行为定位任务中 3D 网络参数量过大的问题. 本文算法基于双流网络的框架实现, 在双流网络的框架下同时训练了空间定位网络和时空特征提取网络, 将 SVD 算法引入 3D 卷积中, 构建了能将 3D 卷积核

进行分解的 3D-SVD, 降低了网络的参数量, 实现了行为的定位和分类; 利用 mixup 算法进行了数据增强, 辅以基础数据增强操作对数据集进行增广; 并采用序列重排序算法和动态规划算法构建了更为合适的行为管道. 在两个常用的公开数据集上进行实验的结果表明, 本文的模型在各指标上能获得较优的结果.

参考文献

1 Weinzaepfel P, Harchaoui Z, Schmid C. Learning to track for spatio-temporal action localization. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 3164–3172.

2 Peng XJ, Schmid C. Multi-region two-stream R-CNN for action detection. Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016. 744–759.

3 Yang ZH, Gao JY, Nevatia R. Spatio-temporal action detection with cascade proposal and location anticipation. arXiv: 1708.00042, 2017.

4 Alwando EHP, Chen YT, Fang WH. CNN-based multiple path search for action tube detection in videos. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(1): 104–116. [doi: 10.1109/TCSVT.2018.2887283]

5 Hou R, Chen C, Shah M. Tube Convolutional Neural Network (T-CNN) for action detection in videos. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 5823–5832.

6 Kalogeiton V, Weinzaepfel P, Ferrari V, *et al*. Action tubelet detector for spatio-temporal action localization. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 4415–4423.

7 Li D, Qiu ZF, Dai Q, *et al*. Recurrent tubelet proposal and recognition networks for action detection. Proceedings of the 15th European Conference on Computer Vision. Cham: Springer, 2018. 306–322.

8 He JW, Deng ZW, Ibrahim MS, *et al*. Generic tubelet proposals for action localization. Proceedings of 2018 IEEE Winter Conference on Applications of Computer Vision. Lake Tahoe: IEEE, 2018. 343–351.

9 Qiu ZF, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 5534–5542.

10 Tran D, Wang H, Torresani L, *et al*. A closer look at spatiotemporal convolutions for action recognition. Proceedings of 2018 IEEE/CVF Conference on COMPUTER Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6450–6459.

11 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321–357. [doi: 10.1613/jair.953]

12 Inoue H. Data augmentation by pairing samples for images classification. arXiv: 1801.02929, 2018.

13 Goodfellow IJ , Pouget-Abadie J, Mirza M , *et al*. Generative adversarial networks. Advances in Neural Information Processing Systems, 2014, 3: 2672–2680.

14 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.

15 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149. [doi: 10.1109/TPAMI.2016.2577031]

16 Girshick R. Fast R-CNN. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 1440–1448.

17 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.

18 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016. 21–37.

19 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: NIPS, 2014. 568–576.

20 Hara K, Kataoka H, Satoh Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6546–6555.

21 Köpüklü O, Wei XY, Rigoll G. You only watch once: A unified CNN architecture for real-time spatiotemporal action localization. arXiv: 1911.06644, 2019.

22 Pramono RRA, Chen YT, Fang WH. Hierarchical self-attention network for action localization in videos.

Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 61–70.

23 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv: 1804.02767, 2018.

24 Li C, Zhong QY, Xie D, *et al*. Collaborative spatiotemporal feature learning for video action recognition. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7864–7873.

25 Touvron H, Vedaldi A, Douze M, *et al*. Fixing the train-test resolution discrepancy. Proceedings of the 33rd Conference on Neural Information Processing Systems. Vancouver: NeurIPS, 2019. 8250–8260.

26 Yang XT, Yang XD, Liu MY, *et al*. STEP: Spatio-TEmporal Progressive learning for video action detection. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 264–272.

27 Singh G, Saha S, Sapienza M, *et al*. Online real-time multiple spatiotemporal action localisation and prediction. Proceedings of 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 3657–3666.

28 Saha S, Singh G, Sapienza M, *et al*. Deep learning for detecting multiple space-time action tubes in videos. arXiv: 1608.01529, 2016.

29 Singh G, Saha S, Cuzzolin F. Predicting action tubes. Proceedings of European Conference on Computer Vision. Cham: Springer, 2018. 106–123.