计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(5):262−272 [doi: 10.15888/j.cnki.csa.009109]

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

基于深度学习的翻越行为检测①

王　林, 赵　甜

(西安理工大学 自动化与信息工程学院, 西安 710048)
通信作者: 赵　甜, E-mail: 1164182600@qq.com

摘　要: 翻越行为检测对疫情管控、社会治安等有着重要意义, 一定程度上可以减少因为违规翻越行为而造成的意外事故. 针对目前翻越行为检测任务实时性差、需要先验知识的问题, 本文将 Faster RCNN+SlowFast 时空行为检测算法应用在翻越行为检测任务中, 将翻越行为进行拆分检测. 为提高时空行为检测算法中目标的检测精度和速率将目标检测模块 Faster RCNN 改为实时性高且轻量化的 YOLOv5; 其次针对同一行为不同视角下广泛的类内多样性的问题, 将 Fast 支路和 Slow 支路的 residual block 分别改为 AC residual block 和 SE residual block 来加强网络对关键特征与细粒度特征的学习能力, 最后设计翻越行为检测算法进行攀爬与下降两种状态的连续性检测, 实验结果显示该网络平均准确率达 93.5%, 在翻越行为检测中表现出良好的性能.

关键词: 翻越行为; 行为识别; 时空行为检测; SlowFast

引用格式: 王林,赵甜.基于深度学习的翻越行为检测.计算机系统应用,2023,32(5):262–272. http://www.c-s-a.org.cn/1003-3254/9109.html

Crossing Behavior Detection Based on Deep Learning

WANG Lin, ZHAO Tian

(School of Automation and Information, Xi'an University of Technology, Xi'an 710048, China)

Abstract: Crossing behavior detection is of great significance for epidemic control and social security and can reduce accidents caused by illegal crossing behavior to a certain extent. In view of the problems of poor real-time performance and the need for prior knowledge in the current crossing behavior detection task, this study applies the Faster RCNN+SlowFast spatiotemporal behavior detection algorithm to the crossing behavior detection task to split and detect the crossing behavior. In order to improve the detection accuracy and speed of the target in the spatiotemporal behavior detection algorithm, the target detection module, namely Faster RCNN is changed to lightweight YOLOv5 with high real-time performance. Then, according to the extensive in-class diversity under different perspectives of the same behavior, the residual block of the Fast branch and Slow branch is changed to AC residual block and SE residual block, respectively, so as to strengthen the network's learning ability to key features and fine-grained features. Finally, the crossing behavior detection algorithm is designed to detect the continuity of climbing and descending states. Experimental results show that the average accuracy of the network reaches 93.5%, which shows excellent performance in crossing behavior detection.

Key words: crossing behavior; action recognition; spatiotemporal behavior detection; SlowFast

日常生活中经常会出现因为发生违规翻越行为而造成的意外事故, 尤其是在疫情管控期间比如 2022 年 3 月江西省某路口有 7 人违规翻越及破坏疫情防控围挡设施, 给疫情防控带来了非常大的风险; 2022 年 4 月,

① 基金项目: 陕西省科技计划重点项目 (2017ZDCXL-GY-05-03)
收稿时间: 2022-11-07; 修改时间: 2022-12-10; 采用时间: 2023-01-06; csa 在线出版时间: 2023-03-30
CNKI 网络首发时间: 2023-03-30

违法行为人王某某明知某高校疫情期间采取不允许校外人员进入的管控措施仍通过翻越铁栅栏方式进入校园, 扰乱该高校的防疫秩序等. 这些新闻报道暴露出了违规翻越行为对社会治安、疫情防控造成的不良影响以及监控发现不及时的问题. 目前人们日常的监控系统往往只是记录事物的发生, 对于意外事故的检测均采取后台人工监测, 大部分只有在出现意外事故后才进行查找, 这样不仅耗费人力资源也没有将视频资源利用起来. 本文研究把视频理解领域中的行为识别技术应用于安防监控中自动化地识别出视频中的翻越行为并及时的警告与制止具有重要意义及实践价值.

国内外学者尝试了许多将现有图像分类任务中的卷积神经网络算法迁移到视频行为识别领域中, 但与图像分类任务不同的是网络还需提取不同帧的时序特征, 根据提取时序特征的不同将现有方法分为基于 2D 卷积的行为识别和基于 3D 卷积的行为识别. 基于 2D 卷积的行为识别比较经典的是时空双流网络结构 (two-stream CNN)[1], 该方法利用光流来捕捉视频帧之间的时序信息, 对此还有在该基础上改进的双流网络的变体及 TSN 网络[2]、I3D 网络[3] 等, 虽然光流法对行为识别的准确率有一定提高但是提取光流特征的内存消耗和计算代价非常大无法实现端到端的识别. 基于 3D 卷积的行为识别相比于 2D 卷积多了一个时间维代替光流来学习复杂的时序特征, 可以实现端到端的特征提取和分类. 比较经典的是 C3D 网络[4], 网络设计是将 VGGNet 的卷积核由 3×3 的 2D 卷积扩展为 3×3×3 的 3D 卷积, 虽然模型简单但是三维卷积核使得网络的参数量和计算成本呈指数增长, 该算法设计属于 3D 卷积进行行为识别的一个里程碑性的工作, 在此之后就有了其他的 3D 网络比如 R3D[5]、SlowFast[6] 算法等.

国内外学者就异常行为检测的方法有 Khunchai 等人[7] 提出利用实时人体姿态估计进行关键点提取后再将关键点转换为与 x 轴中的参考点进行角度位置比较进而检测人体行为的方法. Sun 等人[8] 提出采用背景阈值分割法检测前景目标后得到关键区域进行逻辑判断识别危险行为的方法. Bian 等人[9] 提出将边缘检测算法提取图像边缘作为三维卷积神经网络的输入来识别打架、攀爬和摔倒等异常行为. Priyadarshini 等人[10] 提出将穿戴传感器得到的三维数据利用梯度提升树与决策树方法集成进行行为分类识别. 李自强等人[11] 提出一种相同自编码结构的孪生网络, 并在编码器与解码器间加入了记忆增强模块的方法对视频异常行为进行检测. Yi 等人[12] 提出将残差网络结合通道注意力与空间注意力进行行为识别, 使网络更关注有用信息, 提高动作识别的准确率. Yu 等人[13] 提出将翻越行为分为 4 个连续的行为, 使用扩展的星形骨架特征进行人体建模, 根据轮廓点与质心的关系特征构建离散隐马尔可夫模型 HMM的方法来对翻越行为进行检测. 胡瑗等人[14] 提出一种利用轨迹特征进行行为检测的方法, 其中使用角点跟踪 KLT 算法进行目标跟踪, 连接随时间顺序的采样点生成原始轨迹, 对轨迹进行聚类, 并且提出将复杂行为拆分进行识别的方法. Kolekar 等人[15] 提出使用 SVM 分类器对步行、爬上、爬下等的行为特征进行训练的方法对翻越行为进行拆分检测. 张泰等人[16] 提出先使用目标检测再进行目标跟踪然后学习轨迹特征进行翻越行为识别的方法.

目前大多数翻越行为检测算法都为手工特征提取方法或将动作转化为跟踪轨迹与翻越交接线进行曲线判断的方法, 大部分不能直接对视频画面中的动作直接进行检测, 存在检测速率慢、鲁棒性不高、无法应用在不同场景中的问题. 针对目前存在的问题, 本文提出基于时空行为检测算法的翻越行为检测模型, 其中包括时空行为检测模块 (目标检测模块、多标签行为分类模块)、行为判别模块两部分, 其中时空行为检测模块以短时间为一个周期, 对短时间内视频关键帧中每个人的位置信息及其行为信息进行检测, 在实际的应用中可以达到近似实时的效果. 其次针对时空行为检测算法在自制数据集遮挡情况下目标框定位不准确、行为误检的问题, 对时空行为检测算法进行改进, 最后设计行为判别模块对拆分行为进行判断实现翻越行为检测.

1 相关工作

1.1 Faster RCNN+SlowFast 时空行为检测算法

时空行为检测算法要实现两个任务, 一是目标的定位任务, 二是对目标进行多标签行为分类任务, 如图 1 所示. 时空行为检测算法处理视频输入帧以 3 s 为一个周期, 步幅为 1 s, 对 1.5 s 时的关键帧进行标注, 对关键帧提供短时间的上下文信息. 数据在一个相对简短的时间背景下不仅可以提高时空行为检测

算法的检测速度还可以消除只通过静态帧进行行为判断产生理解歧义的问题. 对目标的定位采用现有的目标检测模型 Faster RCNN 直接得到目标框尺寸, 完成定位任务. 然后借鉴 Faster RCNN 算法相关思想将目标框投影到 3D 特征提取网络得到的特征块上获得相应的特征矩阵, 具体实现是将每个 2D RoI 延时间复制成 3D RoI, 采用 RoIAlign 方法对感兴趣的特征区域处理为统一尺寸, 最后将得到的特征在时间维度进行平均池化, 再连接一个全连接层由 Sigmoid 分类器进行多标签行为预测. 目标检测模块没有参与和行为检测模块 SlowFast 联合训练, 目标检测模块采用现有已训练好的检测器.

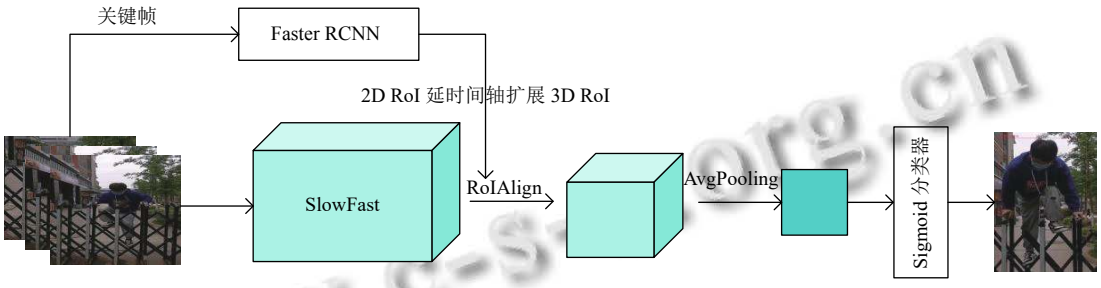

图 1 时空行为检测算法网络结构图

1.2 Faster RCNN 网络

Faster RCNN 为经典的两阶段目标检测算法, 由 RPN 区域生成网络和 Fast RCNN 网络组成. 如图 2 所示将图像输入主干网络得到相应的特征图, 使用 RPN 区域生成网络生成候选框后获得相应的特征矩阵, 之后再通过 RoI pooling 层将特征矩阵缩放在 7×7 大小, 再展平通过一系列全连接层进行分类和目标边界框修正.

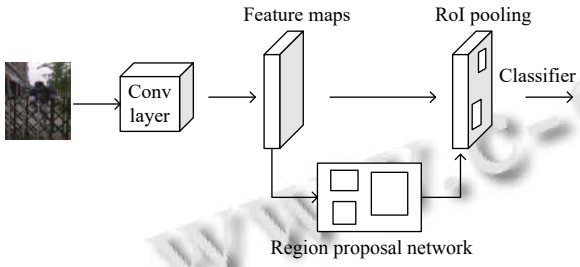

图 2 Faster RCNN 网络结构图

1.3 SlowFast 网络

SlowFast 网络架构[6] 是受生物学视网膜神经细胞种类分布启发, 网络为不同帧率运行的双流单一架构, 通过设置不同的采样间隔和卷积通道数实现两条路径不同的时间分辨率和空间分辨率, 网络结构图如图 3 所示. 慢速路径以低时间分辨率和高空间分辨率来捕获视频空间物体的语义特征, 设该分支采用的时间分辨率 τ, 通道数为 D, 采样的帧数为 T, 则原始的视频长度为 $\tau \times T$ 帧. 该分支只有在高层才使用 3D 卷积, 在底层过早地使用 3D 卷积会降低准确率, 只有在空间的感受野足够大之后, 帧之间的时序关系才明显. 另一条快速路径以高时间分辨率和低空间分辨率来捕获视频中物体快速变化的运动特征, 该支路运动信息变化比较快, 输入 αT (α>1) 帧, 高帧率数据以及较少通道 βD (β=1/8) 在更精细的时间维度捕获快速变化的运动信息. 计算量和通道数的平方成正比, 由于 Fast 支路通道数较少, 因此 Fast 支路十分的轻量级, 只占用整体 20% 的计算量. Fast 支路为了保证时序上的保真度以获得精细的运动信息, 在时间维度上不进行任何汇合操作, 始终保持时间维度为 αT, 直到最后的全局平均汇合.

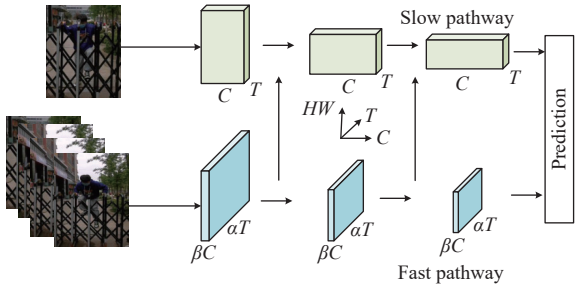

图 3 SlowFast 网络结构图

Fast 和 Slow 分支有着很强的互补性, 信息融合时为 Fast 分支向 Slow 分支进行单向信息融合, 由于 Slow 分支的维度 $D \times T \times S \times S$ 和 Fast 分支的维度 $\beta D \times \alpha T \times S \times S$ 不一致, 因此需要执行转换的横向连接来匹配两条支路

输出的特征向量. 有 3 种单向连接方式分别为时间转换通道、时间步长采样, 时间步长卷积, 时间步长卷积的效果最好, 本文也采用时间步长卷积进行信息融合, 在网络的 $pool_1$、res_2、res_3 与 res_4 层之后进行时间步长卷积后进行单向特征融合, 最后将 Slow 和 Fast 分支的特征向量进行拼接, 经过全连接层进行类别预测. SlowFast 为一种思想, 网络具体实现其 backbone 使用的是基于 3D 的 ResNet50, SlowFast 网络实例如表 1 所示.

表 1 SlowFast 网络结构 (backbone 为 ResNet50)

Stage	Slow pathway	Fast pathway	S/F output $T\times S^2$
rawclip	—	—	64×224^2
data layer	Stride16, 1^2	Stride2, 1^2	$4/32\times224^2$
$conv_1$	1×7^2, 64	5×7^2, 64	$4/32\times112^2$
$Pool_1$	1×3^2, max	1×3^2, max	$4/32\times56^2$
res_2	{1×1^2, 64 1×3^2, 64 1×1^2, 256}×3	{3×1^2, 8 1×3^2, 8 1×1^2, 32}×3	$4/32\times56^2$
res_3	{1×1^2, 128 1×3^2, 128 1×1^2, 512}×4	{3×1^2, 64 1×3^2, 64 1×1^2, 256}×4	$4/32\times28^2$
res_4	{3×1^2, 256 1×3^2, 256 1×1^2, 1 024}×6	{3×1^2, 32 1×3^2, 32 1×1^2, 128}×6	$4/32\times14^2$
res_5	{3×1^2, 512 1×3^2, 512 1×1^2, 2048}×3	{3×1^2, 64 1×3^2, 64 1×1^2, 256}×3	$4/32\times7^2$
Global average pool, concat, fc			#classes

2 改进的时空行为检测模型

2.1 YOLOv5 网络

目标检测模块在整个网络模型中不仅需要完成定位任务同时在 SlowFast 网络中也需要利用目标框获得特征矩阵进行行为分类任务, 因此目标检测模块定位准确性将影响网络的整体性能. 本文研究的人体翻越行为存在目标遮挡、多尺度等问题, 原 Faster RCNN 表现较差, 故本文选用在实时性和模型参数量都更优于 Faster RCNN 算法的 YOLOv5 进行目标定位, 该模型包括 backbone 特征提取模块、neck 特征融合模块、head 预测模块 3 部分. 特征提取模块即网络的主干采用 New CSP-Darknet53, 该模块在 YOLOv4 主干 CSP-Darknet 的基础上做了一点小的改进添加了 Focus 模块, 该模块将相邻像素划分为一个 patch, 将每个 patch 相同位置的像素点进行了拼接, 长宽尺度减小, 通道数变为之前的 4 倍, 该模块也可以用一个 6×6 卷积层代替效率更高. 在特征融合模块采用 SPPF, SPPF 模块在 SPP 的模块上进行了改进, 都能在一定程度上解决目标多尺度的问题, 但是 SPPF 模块相比于 SPP 模块的效率更高些; New CSP-PAN 模块有自顶向底和自底向顶两条路径的信息融合并且在 PAN 网络的基础上引入了 CSP 结构; 预测模块分了 3 组特征图分支分别通过 1×1 卷积核进行不同尺寸目标的预测, YOLOv5 网络模型如图 4 所示.

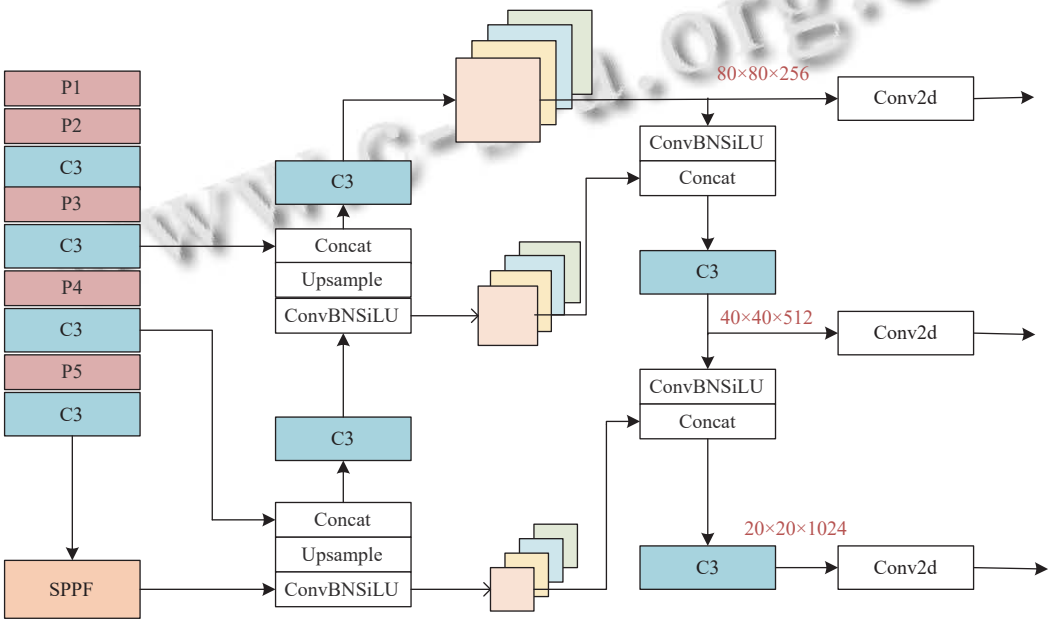

图 4 YOLOv5 网络结构图

2.2 改进 SlowFast 网络

视频行为理解任务中人体行为存在丰富的多样性, 如同一行为不同执行目标外观差异较大, 其上下文信息也不同, 同时拍摄角度、光线等的影响, 在视频动作检测任务中极容易出现误检现象. 3DCNN 在时空建模方面很有效, 但是无法捕获视频中包含的足够信息, 要提高分类准确率还需要学习更加细粒度的特征. 为提高网络的检测精度, 本文对 SlowFast 网络做了两部分改进, 一是在主要提取行为运动特征的 Fast 分支引入 ACTION 注意力模块, 来提高网络对细粒度特征的学习能力; 二是在主要提取视频空间特征 Slow 支路引入 SE 注意力模块, 加强网络对行为空间特征关键特征的提取能力, 总体上提升网络的性能.

2.2.1 ACTION 模块

ACTION 模块[17] 是由 STE、CE、ME 这 3 个互补的注意力模块组成, 可以提取视频关键的时空特征、动作时序特征在不同通道间的权重、动作相邻帧之间的变化轨迹特征, 网络结构图如图 5 所示.

STE 时空注意力模块通过生成时空掩码 M 来生成时空 attention map, 为了减少直接对 3D 卷积进行操作而增加模型的计算量, 首先需要对输入 $X \in R^{[N,T,C,H,W]}$ 做一个通道全局化得到相对应的通道轴的全局时空张量 $F_1 \in R^{[N,T,1,H,W]}$, 然后 reshape $F_1^* \in R^{[N,1,T,H,W]}$. 这样就可以将 F 输入到一个 $3\times3\times3$ 的卷积层 K 中, 表示为:

$$F_{o1}^* = K_1 \times F_1^* \tag{1}$$

再对 F_{o1}^* 重构为 $F_{o1} \in R^{[N,T,1,H,W]}$, 通过激活函数 Sigmoid 得到:

$$M_1 = \delta(F_{o1}) \tag{2}$$

最终输出为:

$$Y_1 = X + X \odot M_2 \tag{3}$$

网络结构图如图 6(a) 所示.

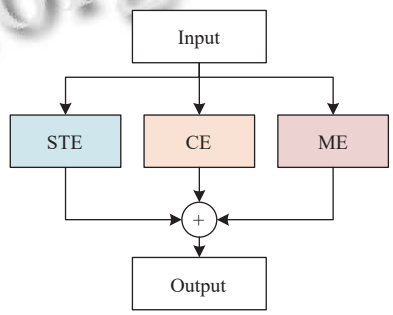

图 5 ACTION 模块

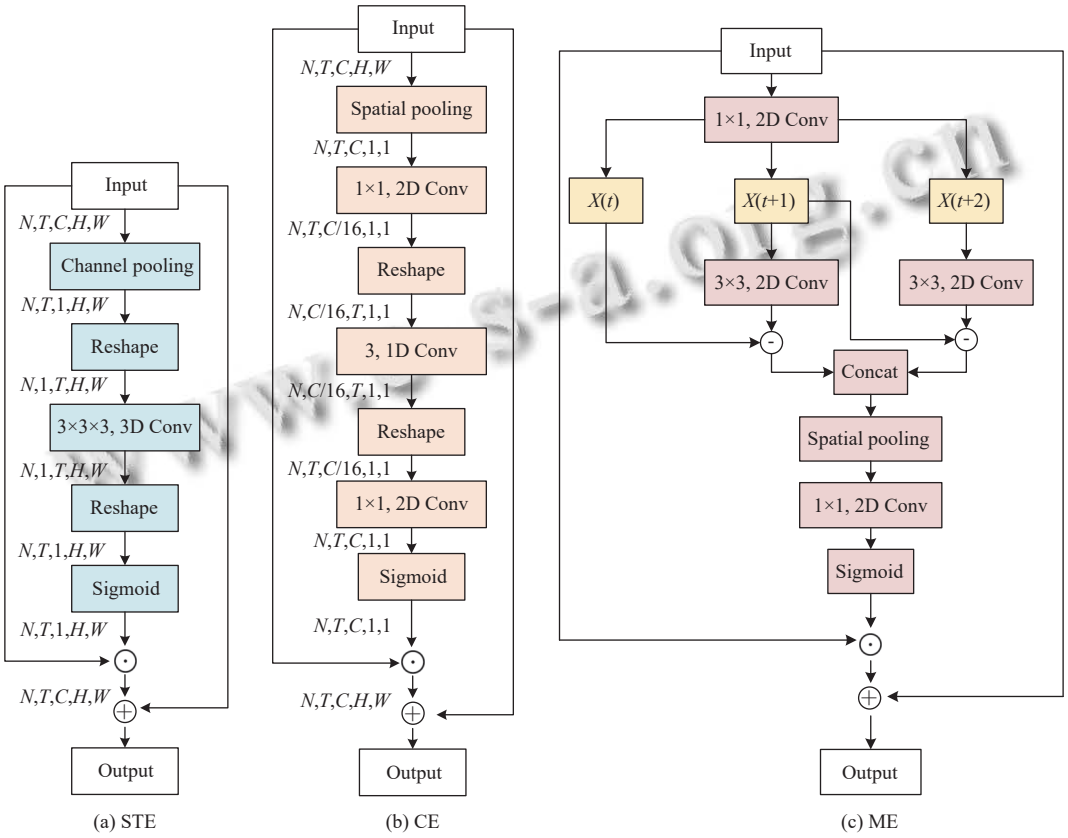

图 6 STE、CE、ME 模块

CE 的结构和 SE 的结构相似, CE 结构不同的是在两个 FC 层之间插入了一个一维卷积层来表征信道特征的时间信息, 可以自适应校准通道特征响应, 网络结构图如图 6 (b) 所示. 对于输入$X \in \mathrm{R}^{[N,T,C,H,W]}$首先对空间进行平均池化得到$F_2 \in R^{[N,T,C,1,1]}$, 求解 F_2 的表达式为:

$$F_2 = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X[:,:,:,i,j] \tag{4}$$

用 1×1 的卷积核对 F_2 的通道数进行压缩得到:

$$F_r = K_1 \times F_2 \tag{5}$$

将$F_r \in R^{[N,T,C/r,1,1]}$重构为$F_r^* \in R^{[N,C/r,T,1,1]}$

再使用内核大小为 3 的 1 维卷积核K_2来处理:

$$F_{\text{temp}}^* = K_2 \times F_r^* \tag{6}$$

Reshape 后, 再通过一个 1×1 的 2D 卷积核K_3, 并使用激活函数, 表示为:

$$F_{o2} = K_3 \times F_{\text{temp}} \tag{7}$$

$$M_2 = \delta(F_{o2}) \tag{8}$$

$$Y_2 = X + X \odot M_2 \tag{9}$$

ME 模块如图 6(c) 所示提取相邻帧之间的运动信息, 运动特征建模表示为:

$$F_m = K \times F_r[:,t+1,:,:,:] - F_r[:,t,:,:,:] \tag{10}$$

其中, K为3×3的卷积; 整体网络运动特征建模首先通过 1×1 的卷积对输入的通道进行压缩, 然后分别通过式 (10) 计算相邻帧的运动特征并进行串联得:

$$F_M = [F_m(1), \cdots, F_m(t-1), 0] \tag{11}$$

接着依次经过空间平均池化、通道调整、Sigmoid 函数得到权重矩阵M_3, 最后得到输出为:

$$Y_3 = X + X \odot M_3 \tag{12}$$

最终 ACTION 模块的输出为:

$$Y = Y_1 + Y_2 + Y_3 \tag{13}$$

ACTION 注意力模块在特征级别对网络内部特征进行建模, 可以激活视频中的多类型信息更好的表征动作. 原 SlowFast 网络由堆叠的残差块中 3×1×1、1×3×3 卷积核来提取时空特征来表征动作, 但是广泛的类内多样性需要学习更加细粒度的时空特征, 在每个残差块前加 ACTION 模块在获取多类型的激励信号后再进行卷积获取更加细粒度的特征, 可以提高多标签行为分类的精度. 如图 7 所示.

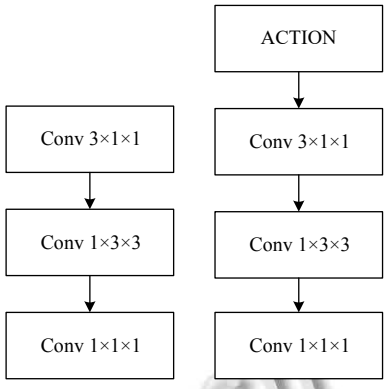

图 7 AC residual block 网络结构图

2.2.2 引入 SENet 空间注意力

在 SlowFast 网络中, Slow 支路网络设置了为快速通道的 β 倍的通道数来学习空间信息, 加倍的通道数意味着有更多的特征, 学习过程中同等程度对待这些特征会使网络学习到一些对行为分类任务没有帮助的特征, 在分类任务中产生误检现象. 为了更好地提取视频中行为的空间信息, 本文对 Slow 支路的 residual block 添加 SENet 模块, 称为 SE residual block, 如图 8 所示.

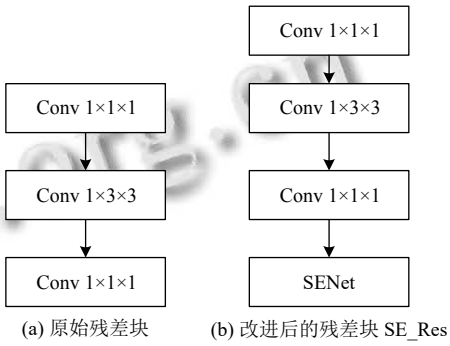

图 8 SE residual block 网络结构图

SENet 首先在时域上进行扩展, 在每个时域上进行一个空间的全局平均池化, 每个通道得到一个标量, 再送入两层全连接层得到 C 个通道各自的权值, 根据信息的重要程度对通道信息赋予不同的权值, 最后对网络重要的 feature map 赋予大权重一些无效的 feature map 赋予小权重, 突出通道的关键信息和抑制视频背景干扰像素的影响, 同时残差块中 SENet 模块的引入也增加了原网络的非线性使其可以更好地拟合不同通道之间的关系, 网络结构图如图 9 所示.

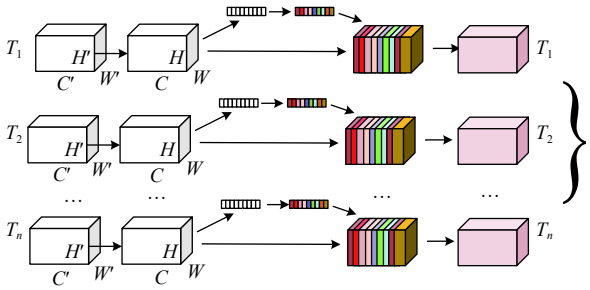

图 9 SENet 网络结构图

2.3 损失函数

本文采用 BCELoss 二分类交叉熵损失函数, 与多类别分类任务不同的是不同行为类别之间并不属于互斥关系可以同时出现, 例如某人可以一边走路一边说话, 不能简单地使用 Softmax 将输出归一化为相加为 1 的 [0, 1] 之间的概率值. 在本文行为多标签分类任务中需先将网络输出数据输入 Sigmoid 函数将取值缩放到 [0, 1] 之间, 再计算 Loss. Sigmoid 函数、BCELoss 损失函数公式为:

$$y = \frac{1}{1+\mathrm{e}^{-z}} \tag{14}$$

$$L(y,t) = -\frac{1}{n}\sum w \times (t\ln(y)) + (1-t)\times \ln(1-y) \tag{15}$$

其中, y为网络预测值, t为标签值, w为权重系数.

3 翻越行为检测模块

本文将翻越行为定义为攀爬与下降两种状态同时发生的连续性动作, 为检测两种状态是否为连续性发生, 设计翻越行为检测算法. 算法流程如算法 1.

算法 1. 翻越行为检测算法流程

(1) 初始化, 设定行为计数 *n* 变量用于读取检测结果, 设置攀爬行为检测标识变量 *Flag*, 用于记录上一个动作是否为攀爬行为.
(2) 读取检测到的行为.
(3) If *Flag*==0, 从翻越行为开始检测; 跳到步骤 (4).
If *Flag*!=0, 即上一个动作是攀爬行为; 跳到步骤 (5).
(4) If *label*==“climb”, 将 Flag 置 1.
If *label*!=“climb”, *Flag* 不变.
等待读取下一个行为; 等待读取下一个行为; 跳到步骤 (2).
(5) If *label*==“fall”, 输出翻越行为记录, 将 *Flag* 置 0; 跳到步骤 (2).
If *label*==“fall”, 可能存在连续攀爬行为, 跳到步骤 (4).

4 翻越行为检测系统

本文提出将非原子性的翻越行为拆分为攀爬与下降的原子性行为, 利用时空行为检测算法先对原子性行为进行识别再设计算法对翻越这种组合行为进行检测的方法. 该系统主要由监控视频输入模块、视频分割模块、目标检测模块、行为识别模型、翻越行为判别模块组成, 如图 10 所示.

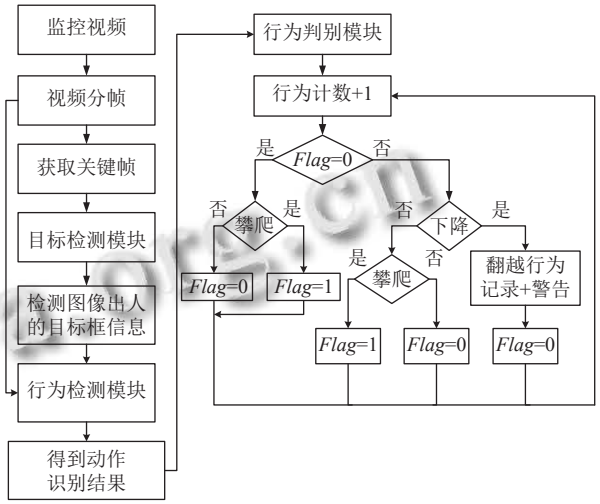

图 10 翻越行为检测系统

5 实验分析

5.1 实验环境

本文实验环境为 CentOS Linux 7 操作系统; GPU 为 GeForce GTX 1080Ti (×4), 11 GB; CPU 为 Intel(R) Xeon(R)CPU E5-2640v4@2.20 GHz; 深度学习框架为 PyTorch 框架.

5.2 实验数据集

本文主要研究人员的异常翻越行为, 没有通用的数据集需要自建视频行为数据集. 采集不同场景下的数据集比较困难, 本文采用视频拼接一部分 HMDB51 公共数据集的视频, 一部分进行实地拍摄. 实地拍摄场景有 3 个, 包括 A 场景楼梯栏杆、B 场景防盗栅栏、C 场景交通场景下护栏等 (拍摄视频场景均在安全且被允许的情况下进行采集的), 数据集部分帧如图 11 所示.

图 11 数据集部分帧展示

按照比例将自制数据集随机划分为训练集和测试集, 按照 AVA 数据集[18] 的标注格式进行数据标注, 每个片段 30 s, 共有 200 个片段. 为了平衡异常行为和其他行为的视频数量, 本文将 AVA 数据集进行了选择和裁剪, 对于自定义的数据集标注采用半自动化方式进行标注每秒标注一帧, 数据集制作流程如图 12 所示. 先用 Faster RCNN 进行目标框检测, 将生成的 CSV 文件导入 VIA 进行行为标注, 标注类别为 stand、walk、climb、fall 这 4 大类, 最终生成检测框标注文件、时空行为标注文件、不参与帧标注文件、标签标注文件 4 类文件.

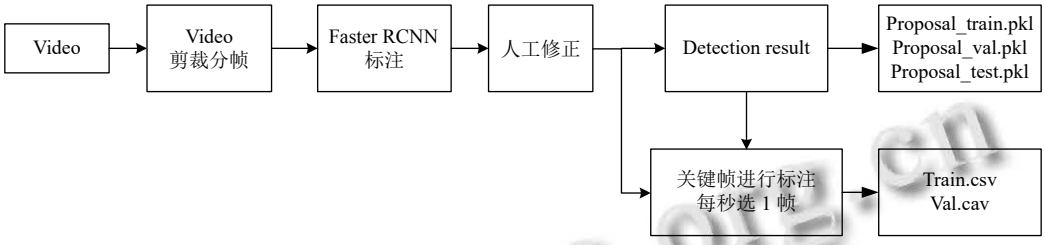

图 12 数据集制作流程

5.3 实验结果与分析

(1) 目标检测模块评估

原网络采用 Faster RCNN 算法检测精度一般, 运行时间较慢. 为提高栅栏遮挡情况下人物的识别精度及速度将目标检测模块进行替换, 为评估不同网络性能分别将自制 200 个视频数据集在 YOLOv5 及 Faster RCNN 上进行训练, 以模型大小、平均帧率 FPS 及 *mAP* 来评估模型的检测性能, 如表 2 所示.

表 2 目标检测模型性能评估

算法	模型大小 (M)	FPS	*mAP*@0.5 (%)
Faster RCNN	41.9	21.2	80.1
YOLOv5	7.4	49.7	93.6

从表 2 可以看出, 本文使用自制行人检测数据集, YOLOv5 相比原 Faster RCNN 的参数量减少了 34.5M, *mAP* 提高了 13.5 个百分点, 检测速率提高了 28.5 FPS. Faster RCNN 算法在遮挡情况下对行人检测时, 存在检测框尺寸与目标框大小不匹配, 无法完全覆盖被测目标, 而改进后的 YOLOv5 对其不同尺度的被测物体都能进行很好的匹配识别, 模型检测对比图如图 13 所示.

使用 YOLOv5 训练使得在攀爬的整个过程中使得攀爬目标的置信度始终高于 0.7, 才会被视为正样本, 所得的 Bounding box 才能被后面的 SlowFast 行为识别网络所采用, 检测同一视频帧的效果图如图 14 所示.

(2) 行为识别模型评估与分析

为验证本文改进模块在行为识别网络中的有效性和可靠性, 采用平均准确率 (mean average precision, *mAP*) 作为评价指标, IoU 取 0.5, 以 SlowFast 原网络为基准, 并将改进分为 4 组进行消融实验. 其中, “√”表示在网络中使用了该改进方法, “×”表示在网络中没有使用该方法. 评价指标公式为:

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \tag{16}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \tag{17}$$

$$mAP = \int_0^1 p(R)dR \tag{18}$$

其中, *Recall* 指所有正样本都被检测到的概率, *Precision* 指所有正样本都被预测正确的概率.

(a) YOLOv5 模型 (b) Faster RCNN 模型

图 13 Faster RCNN 与 YOLOv5 模型的检测对比图

网络训练的学习率设为 0.1, 使用 SGD 优化算法进行训练, 权重衰减为 0.9, 使用 CosineAnnealingLR 策略调整学习率, Dropout 为 0.5, batch size 为 8, epoch

为 200; 对图像进行正则化处理; 通过时间裁剪、空间裁剪、短边缩放进行数据增强; 数据预处理后 RGB 图尺寸归一化为 224×224×3; 每次训练连续采用 32 帧图片, 两条支路分别使用不同时间分辨率分别采样 16 帧和 4 帧, 图 15 为原 SlowFast 与改进后 SlowFast 的训练损失曲线图.

图 14 翻越遮挡情况下 YOLOv5 检测的不同帧图片

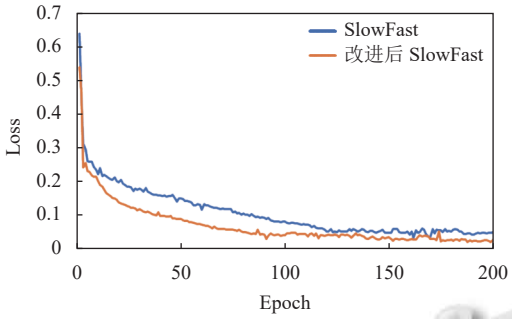

图 15 原 SlowFast 与改进后 SlowFast 的训练损失曲线图

为了更直观地表现出不同改进对不同行为检测的提升情况, 在表 3 中分别对比了 4 种类别行为的检测结果.

从表 3 可以看出, 原模型 1 对攀爬与下降两种主要行为的识别性能一般, 平均 *mAP* 为 81.18%. 模型 2 将 Faster 支路的主干更改添加 ACTION 模块之后, 整个识别精度整体提高, 尤其对时序特征要求更高的动作, 攀爬行为 *mAP* 提高 12.16%, 下降行为 *mAP* 提高了 18.79%, 主要是因为 Fast 支路添加 ACTION 模块后可以在特征级别上很好的融合时间和空间信息, 使得对时序信息特征提取更加完整. 模型 3 在 Slow 支路添加 SEnet 对模型的精度只有小部分提升, 攀爬与下降行为 *mAP* 分别提高 6.92% 和 6.13%, 但是对“stand”这类时序要求不高的行为却带来了比添加 ACTION 模块更高的提升, 说明添加 SEnet 可以有效提取特征图的关键信息进而提升模型整体的准确率, 但是时序特征提取还是不够充分. 模型 4 由于 Faster 支路有向 Slow 支路的单向融合, 将两个支路都进行改进之后的, 总体模型上带来了更高的提升, 平均 *mAP* 提升了 12.1%, 满足了基本需求. 不同场景下翻越行为检测视频帧的结果如图 16 所示.

表 3 消融实验的对比结果 (%)

模型	ACTION	SE	类别的*mAP*@0.5				*mAP*@0.5
			stand	walk	climb	fall	
1	×	×	92.23	90.72	77.67	64.11	81.18
2	√	×	93.14	95.79	89.83	82.90	90.41
3	×	√	96.32	93.18	84.59	70.24	86.08
4	√	√	98.92	96.13	90.81	84.29	93.28

(3) 翻越行为检测模块识别性能评估

本文将翻越行为分为攀爬和下降, 之后由翻越行为检测模块进行翻越行为检测, 对输出结果进行统计, 结果如表 4 所示.

从表 4 可以看出, 该检测算法对于 A 场景楼梯栏杆、B 场景防盗栅栏检测效果好一些, 但是对 C 场景交通场景下护栏下的检测效率较差, 其中一个原因可能是数据集相对较少, 再一个原因可能是因为对于交通场景下跨域护栏的攀爬和下降行为边界不明显且持续时间较短.

图 16 不同场景下翻越行为检测视频帧的结果图

表 4 不同视频场景的对比结果

视频场景	视频数 (识别数)	Accuracy (%)
A	70 (66)	94.28
B	80 (79)	98.75
C	50 (42)	86.00
总计	200	93.50

(4) 本文算法与其他翻越行为检测算法对比

由于目前没有针对翻越行为检测任务的公共数据集, 故本文算法与其他算法的翻越行为检测都是在自制数据集上进行, 可能存在相同行为检测判断的差异, 但由表 5 可以看出, 本文所提的算法准确率高于现有的混合高斯+KLT、3D-ResNet-101+Darknet19 方法、YOLOv3CMP+Multitracker, 与 YOLO+GOTURN 算法准确率相差不大, 但相比于 YOLO+GOTURN 算法, 本文所提出的算法有更加广泛的适用场景, 是直接对行为进行检测的算法.

表 5 本文方法与其他算法对比 (%)

模型	正确率
混合高斯+KLT[16]	92.36
YOLO+GOTURN[19]	93.71
3D-ResNet-101+Darknet19[20]	88.3
YOLOv3+CMP+Multitracker[21]	90
Ours	93.5

6 结论与展望

本文将时空行为检测算法应用在翻越行为检测任务中, 识别精度优于其他算法或与其他算法相当, 可以实现对翻越异常行为的准确识别, 一定程度上促进了时空行为识别的进一步研究, 但在此过程中仍存在一些误识别的情况, 因此在未来的研究中会继续改进模型使模型更加准确和轻量化, 未来如果可以将网络足够轻量化嵌入在监控设备中, 在记录事物发生的同时记录行为将会极大地提高监控设备的作用.

参考文献

1 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS). Montreal: MIT Press, 2014. 568–576.

2 Wang LM, Xiong YJ, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 20–36.

3 Carreira J, Zisserman A. Quo vadis, Action recognition? A new model and the Kinetics dataset. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 4724–4733. [doi: 10.1109/CVPR.2017.502]

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

5 Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3D residual networks for action recognition. Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW). Venice: IEEE, 2017. 3154–3160. [doi: 10.1109/ICCVW.2017.373]

6 Feichtenhofer C, Fan HQ, Malik J, *et al*. SlowFast networks for video recognition. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul: IEEE, 2019. 6201–6210.

7 Khunchai S, Kruekaew A, Getvongsa N. A fuzzy logic-based system of abnormal behavior detection using PoseNet for smart security system. Proceedings of the 37th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC). Phuket: IEEE, 2022. 912–915. [doi: 10.1109/ITC-CSCC55581.2022.9894998]

8 Sun RC, Song XM, Tang Y, *et al*. A balcony dangerous behavior detection algorithm based on video image. Proceedings of the 2019 WRC Symposium on Advanced Robotics and Automation (WRC SARA). Beijing: IEEE, 2019. 296–301. [doi: 10.1109/WRC-SARA.2019.8931955]

9 Bian CL, Xu YM, Wang L, *et al*. Abnormal behavior recognition based on edge feature and 3D convolutional neural network. Proceedings of the 35th Youth Academic Annual Conference of Chinese Association of Automation (YAC). Zhanjiang: IEEE, 2020. 1–6. [doi: 10.1109/YAC51587.2020.9337685]

10 Priyadarshini RK, Banu AB, Nagamani T. Gradient boosted decision tree based classification for recognizing human behavior. Proceedings of the 2019 International Conference on Advances in Computing and Communication Engineering (ICACCE). Sathyamangalam: IEEE, 2019. 1–4. [doi: 10.1109/ICACCE46606.2019.9080014]

11 李自强, 王正勇, 陈洪刚, 等. 基于外观和动作特征双预测模型的视频异常行为检测. 计算机应用, 2021, 41(10): 2997–3003. [doi: 10.11772/j.issn.1001-9081.2020121906]

12 Yi ZW, Sun ZH, Feng JC, *et al*. 3D residual networks with channel-spatial attention module for action recognition. Proceedings of the 2020 Chinese Automation Congress (CAC). Shanghai: IEEE, 2020. 5171–5174. [doi: 10.1109/CAC51589.2020.9326923]

13 Yu E, Aggarwal JK. Detection of fence climbing from monocular video. Proceedings of the 18th International Conference on Pattern recognition (ICPR). Hong Kong: IEEE, 2006. 375–378.

14 胡瑗, 夏利民, 王嘉. 基于轨迹分析的行人异常行为识别. 计算机工程与科学, 2017, 39(11): 2054–2059. [doi: 10.3969/j.issn.1007-130X.2017.11.013]

15 Kolekar MH, Bharti N, Patil PN. Detection of fence climbing using activity recognition by support vector machine classifier. Proceedings of the 2016 IEEE Region 10 Conference (TENCON). Singapore: IEEE, 2016. 398–402. [doi: 10.1109/TENCON.2016.7848029]

16 张泰, 张为, 刘艳艳. 周界视频监控中人员翻越行为检测算法. 西安交通大学学报, 2016, 50(6): 47–53. [doi: 10.7652/xjtuxb201606008]

17 Wang ZW, She Q, Smolic A. ACTION-Net: Multipath excitation for action recognition. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 13209–13218. [doi: 10.1109/CVPR46437.2021.01301]

18 Gu CH, Sun C, Ross DA, *et al*. AVA: A video dataset of spatio-temporally localized atomic visual actions. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 6047–6056. [doi: 10.1109/CVPR.2018.00633]

19 周巧瑜, 曹扬, 詹瑾瑜, 等. 基于 YOLO 和 GOTURN 的景区游客翻越行为识别. 计算机技术与发展, 2022, 32(1): 134–140. [doi: 10.3969/j.issn.1673-629X.2022.01.023]

20 Suo F, Li GH, Zhu CF, *et al*. Analysis of illegal behavior in power station based on video surveillance. Proceedings of the 10th IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Chongqing: IEEE, 2022. 381–385. [doi: 10.1109/ITAIC54216.2022.9836940]

21 孙宝聪. 基于图像检测的机场人员异常行为分析技术研究. 数字通信世界, 2020, (1): 26, 38. [doi: 10.3969/J.ISSN.1672-7274.2020.01.012]

(校对责编: 牛欣悦)