计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(2):147−153 [doi: 10.15888/j.cnki.csa.007770]

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

基于深度学习的围栏跨越行为检测方法①

房　凯

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通讯作者: 房　凯, E-mail: 2322358928@qq.com

摘　要: 在作业现场的安全管理中, 对于非施工人员围栏跨越的监管一直是必不可少的. 但目前施工场地普遍存在作业面广、施工人员管理困难等问题, 导致人工监察的方式效率低下. 而基于视频的人体行为检测技术作为计算机视觉领域重要的研究热点, 在公共安全监控方面有着广泛应用. 因此针对传统人工监察的不足, 结合当前计算机视觉技术, 提出一种智能化的围栏跨越违规检测与识别方法. 该方法通过监控不断获取视频帧, 以视频帧组成的剪辑作为输入, 使用三维卷积和二维卷积分别提取时序和空间特征, 将两部分特征融合后进行分类和边界框回归. 最后通过设置对比试验以验证此方法效果, 实验结果表明, 该方法具有一定的泛化性.

关键词: 计算机视觉; 围栏跨越; 行为检测

引用格式: 房凯.基于深度学习的围栏跨越行为检测方法.计算机系统应用,2021,30(2):147–153. http://www.c-s-a.org.cn/1003-3254/7770.html

Deep Learning-Based Detection Method of Fence Crossing Action

FANG Kai

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: In the safety management of the operation sites, the supervision of fence crossing by non-construction personnel has always been essential. However, at present, there are many problems in the construction sites, such as a wide range of operation and a difficulty in the management of construction personnel, leading to the inefficiency of manual supervision. As an important research hot spot in the field of computer vision, video-based human action detection is widely used in public security monitoring. Therefore, in view of the shortcomings of the traditional manual supervision, in combination with the current computer vision technology, an intelligent detection and recognition method for fence crossing violations is proposed in this paper. In this method, video frames are acquired continuously through monitoring, and clips composed of video frames are taken as input. In addition, temporal and spatial features are extracted by 3D and 2D convolutions respectively. After fusion of the two parts of features, classification and boundary box regression are carried out. Furthermore, a comparative experiment is conducted to verify the effect of this method. The experimental results show that the proposed method can detect the fence crossing behavior accurately in a short time, featuring strong working ability.

Key words: computer vision; fence crossing; action detection

1　引言

针对视频行为分析技术[1]的应用可以有效的提升公共场所的管制水平, 对维护社会稳定和人身安全有着重要意义. 而将针对视频的行为分析技术运用在安防领域[2]中, 不仅可以降低人工监控的程度以减少人力物力, 还可以避免因人力因素导致的重要监控信息的遗漏, 有效提高工作效率, 从而达到对重大事故的预警及监控作用, 避免事故的发生, 因此具有重要的研究意义.

① 收稿时间: 2020-06-09; 修改时间: 2020-07-07; 采用时间: 2020-07-17; csa 在线出版时间: 2021-01-27

围栏作为施工现场实行封闭式管理的重要工具, 在建筑施工作业中, 是明令要求必须提前设置的. 在作业现场对一些存在安全隐患的地方安装围栏隔离起来, 最大程度的为施工安全提供保障, 减少不必要的损失和伤害[3]. 但目前施工现场中对围栏跨越的监管大多依赖人工监察, 而且施工场地普遍存在作业面广、施工人员管理困难, 安监人员难以及时准确了解现场人员的分布和作业情况, 加之工地中各单位安全责任划分不明确, 通常导致安全监督检查力度不够, 所以这种人工监察的方式效率非常低下. 而且尽管围栏按照要求设置, 但存在多数人员安全意识不强, 对围栏跨越的危险性意识不到位.

在这种背景下, 如果能设计一种智能化的围栏跨越违规检测算法, 可以大大提升对于非施工人员跨越围栏情况的监管效率, 实现智能化的安全管理, 及时发现跨越围栏人员并发出警报, 为人员的安全做出了一定的保障.

2 相关工作

近年来, 深度学习[4] 在计算机视觉中得到了广泛的应用, 基于深度学习的动作识别[5] 是一种端到端的方法, 使用深度网络从原始视频中自动学习特征[6] 输出分类结果. 根据深度学习网络的结构的不同, 基于深度学习的动作识别方法[7–9] 主要分为基于双流卷积网络的动作识别和基于三维卷积网络的动作识别.

2.1 基于双流卷积网络的动作识别

视频的处理相对于单帧图像来说更为复杂, 主要原因在于单帧图像仅仅包含空间位置信息, 而视频不仅具有单帧图像的空间特征, 还包含帧与帧之间的时序特征[10]. 因此, 在视频处理方面, 需要同时考虑空间和时间两大部分, 这就要求深度网络具备同时处理不同维度特征的能力[11].

Simonyan 与 Zisserman[12] 在 2014 年发布了双流网络 Two-stream, 包括两部分卷积神经网络来处理视频数据, 表示为空间流和时间流网络, 分别提取相关特征实现人体动作识别. 众所周知, 视频不仅包含空间信息, 还具备时间维度上的信息, 空间信息是相对于组成视频的所有独立的单个帧图像, 每张图像上的像素包含了图像中实际目标的相关信息, 例如一个正在打球的人; 另外时间维度上的信息是指视频中的一帧图像与下一帧图像之间的过度关系, 它是随时间产生的一种光流信息, 即速度矢量场[11]. 因此针对视频独有的特性, 采用双流卷积神经网络进行处理, 其中空间流卷积网络处理单帧图像, 提取空间特征, 时间流卷积网络处理多帧光流, 提取时序特征, 分别经过 Softmax 层处理, 最后使用分类器融合两部分特征, 实现人体动作较为准确地识别.

但是, 上述的空间流和时间流卷积神经网络均为 2D 卷积, Two-stream 双流卷积神经网络的基本网络架构如图 1 所示.

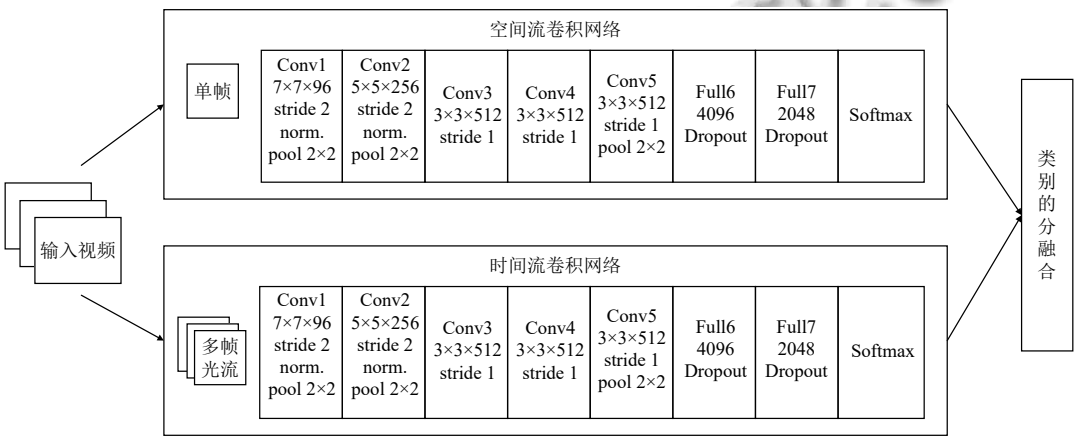

图 1 Two-stream 双流网络结构图

在双流卷积神经网络提出之前, 动作识别的相关研究主要是从处理单帧图像的角度出发, 通过结合分析关键帧中人体运动姿态及其背景实现动作识别. 这种方法的主要问题是没有利用视频本事特有的时间特征, 仅分析每一帧图像中的空间特征, 因此识别效果有限. 双流卷积神经网络正是为了解决此问题而提出, 相比于仅处理单帧图像, 双流卷积神经网络可一次性输入两帧图片, 这样在处理空间信息的基础上还考虑到

了一个动作本身持续性的时间变化特征, 通过综合利用两部分特征[13–16] 极大地提升了动作识别的准确性.

2.2 基于三维卷积网络的动作识别

二维卷积仅可以用来处理单帧图像, 对于视频本身的时间维度上的信息难以处理. 因此三维卷积的作用就显现出来, 它可以看作是对二维卷积的直接扩展, 在原本处理单帧图像空间特征的基础上, 多了一个维度来捕获时序信息. 3D CNN 架构由 Ji 等[17] 提出, 3D 卷积通过堆叠多个连续的帧组成一个立方体, 然后使用 3D 卷积核进行处理. 2D 卷积与 3D 卷积的本质区别在于, 处理视频数据时 2D 卷积操作后生成的特征图还是二维的, 相应的多通道信息被完全压缩, 而 3D 卷积操作后生成的特征图仍然是三维的, 因此保留了视频时间维度上的信息. Tran 等[18] 在前者的基础上提出了一种 C3D (Convolutional 3D) 的现代深层架构, 如图 2 所示, C3D 网络包含 8 次卷积操作, 其中卷积核大小均为 3×3×3, 步长为 1×1×1, 5 次最大池化操作, 除第一层池化的池化核大小和步长为 1×2×2, 其余均为 2×2×2, 最后网络经过两次全连接层和 Softmax 层输出最终结果. 实验结果表明, 此 C3D 方法在视频动作识别精度上要优于之前的方法, 并且其不需要额外的计算光流, 直接可以完成空间信息和时序信息特征的提取操作.

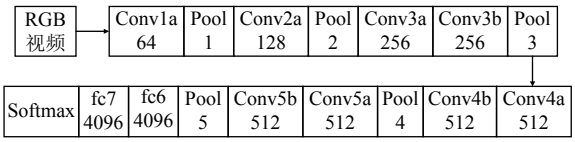

图 2 C3D 网络结构图

为了优化 3D 卷积本身神经网络层数的限制, 进一步提高使用三维卷积进行视频动作识别的研究水平, Carreira 等[19] 沿时间维度重复使用在 ImageNet 上预先训练的二维滤波器, 将用于图像分类非常深的网络拓展为空间-时间特征提取器. Qiu 等[20] 提出了另一种构建深度三维卷积网络的方法: 伪三维残差网 (Pseudo-3Dresidualnet, P3D ResNet).

三维网络相比于双流网络更加简单直接, 可以更直观的捕捉短时间内的时间动态, 但三维网络通常考虑比较短的时间间隔, 因此无法捕获长期的时间信息.

3 基于深度学习的围栏跨越行为检测

针对作业现场围栏跨越违规行为检测问题, 本文从计算机视觉角度提出一种智能化的检测与识别方法. 考虑到二维卷积可以用来解决空间定位问题, 而三维卷积在处理视频时相对传统的双流网络更加简单直接. 因此提出此方法, 通过结合二维卷积及三维卷积, 其中三维卷积用于提取输入剪辑中的时序特征, 输出特征维度为$C' \times H' \times W'$; 二维卷积则提取当前帧空间特征, 解决定位问题, 输出特征维度为$C'' \times H'' \times W''$.

本文拟采用的三维卷积架构为 3D-SE-ResNext-101, 在 3D-ResNext-101 的基础上引入 SE 模块, 相同深度的情况下提升了精度; 采用 Darknet-19 作为二维卷积架构, 提取视频中的空间位置特征; 最后将得到的特征进行通道融合, 然后分类回归, 实现围栏跨越行为检测与识别. 具体流程如图 3 所示.

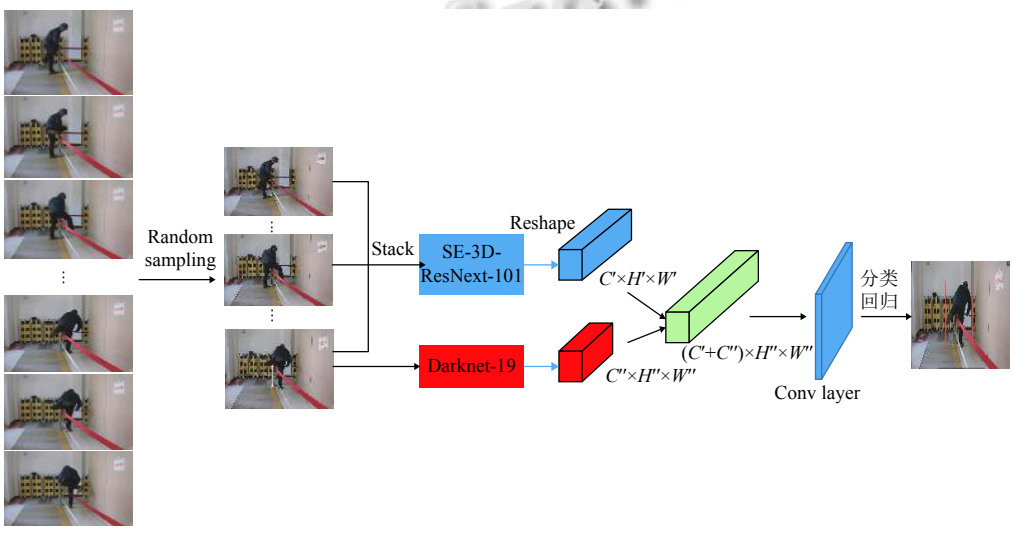

图 3 围栏跨越行为检测流程图

3.1 3D-SE-ResNext-101

三维卷积不仅可以在空间维度上, 而且可以在时间维度上应用卷积运算来捕获运动信息. 众所周知, 残差网络可以有效解决神经网络随深度增加而出现训练效果变差的问题, 其内部多个残差块使用跳跃连接, 可以有效解决梯度消失现象. 3D-ResNext 基本 block 单元如图 4 所示.

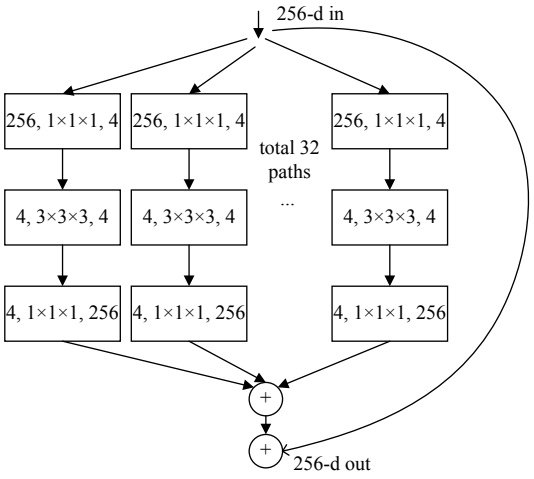

图 4 3D-ResNext 基本单元

SE 模块主要包括 Squeeze 和 Excitation 两个操作, 可以适用于任何映射:

$$F_{\text{tr}}: X \to U, X \in R^{H' \times W' \times C'}, U \in R^{H \times W \times C} \tag{1}$$

以卷积为例, 卷积核为$V = [v_1, v_2, \cdots, v_C]$, 其中$v_c$表示第 c 个卷积核. 那么输出$U = [u_1, u_2, \cdots, u_c]$:

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s \tag{2}$$

其中, *代表卷积操作, 而v_c^s代表一个 3D 卷积核, 其输入 channel 上的空间特征, 它学习特征空间关系, 但是由于对各个 channel 的卷积结果做了 sum, 所以 channel 特征关系与卷积核学习到的空间关系混合在一起. 而 SE 模块就是为了抽离这种混杂, 使得模型直接学习到 channel 特征关系. 对于 3D-ResNext, SE 模块嵌入到残差结构中的残差学习分支中, 如图 5 所示.

3.2 Darknet-19

为了解决空间定位问题, 并行提取当前帧的二维特征. 我们采用 Darknet-19 作为基本架构, 因为它在准确性和效率之间取得了很好的平衡. 如表 1 所示, 包含 19 个卷积层和 5 个最大池化层, 同时使用 batch normalization 来加速收敛.

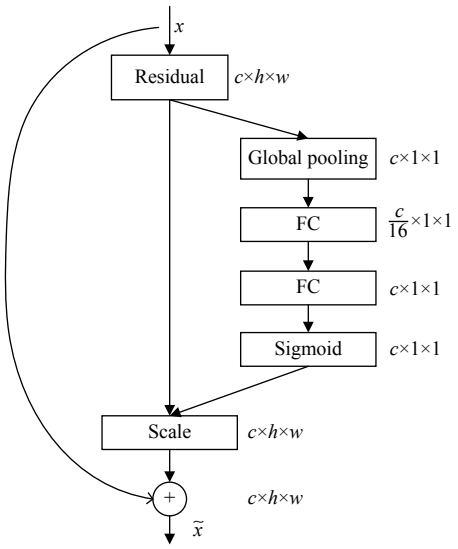

图 5 3D-SE-ResNext module

表 1 Darknet-19

Type	Filters	Size/Stride	输出
Convolutional	32	3×3	224×224
Maxpool		2×2/2	112×112
Convolutional	64	3×3	112×112
Maxpool		2×2/2	56×56
Convolutional	128	3×3	56×56
Convolutional	64	1×1	56×56
Convolutional	128	3×3	56×56
Maxpool		2×2/2	28×28
Convolutional	256	3×3	28×28
Convolutional	128	1×1	28×28
Convolutional	256	3×3	28×28
Maxpool		2×2/2	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Convolutional	256	1×1	14×14
Convolutional	512	3×3	14×14
Maxpool		2×2/2	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	512	1×1	7×7
Convolutional	1024	3×3	7×7
Convolutional	1000	1×1	7×7
Avgpool		Global	1000
Softmax			

3.3 损失函数设计

对于最终输出特征图尺寸$H' \times W'$中的每个网格单元 (gridcell), 用 K-means 方法事先选择 5 个先验框, 因此最终输出大小为$[(5 \times (NumCls + 5)) \times H' \times W']$, 其中 $NumCls$ 表示行为分类得分个数, 还有 4 个坐标和 1 个置信度得分. 对于训练集中的 ground truth, 中心落在哪

个 cell, 那么该 cell 的 5 个 Anchor box 对应的边界框就用来预测它, 最终选择 *IOU* 值最大的边界框负责预测; 与 ground truth 匹配的先验框负责计算坐标误差, 置信度误差以及分类误差, 而其它 4 个边界框只计算置信度误差. 损失函数计算公式如下:

$$loss=\sum_{i=0}^{W}\sum_{j=0}^{H}\sum_{k=0}^{A}(L_1+L_2+L_3) \tag{3}$$

式中, W, H 分别指的是特征图的宽与高; A 指的是先验框数目. L_1, L_2, L_3 如式 (4)~式 (6) 所示.

$$L_1=1_{MaxIOU<Thresh}\lambda_{\text{noobj}}*\left(-b_{ijk}^{o}\right)^2 \tag{4}$$

计算各个预测框和所有 ground truth 之间的 *IOU* 值, 若最大值也小于阈值, 则标记此为 background.

$$L_2=1_{t<12\,800}\lambda_{prior}*\sum_{r\in(x,y,w,h)}\left(prior_k^r-b_{ijk}^r\right)^2 \tag{5}$$

计算先验框与预测框的坐标误差.

$$\begin{aligned}L_3=&1_k^{truth}*\lambda_{\text{coord}}*\sum_{r\in(x,y,w,h)}\left(truth^r-b_{ijk}^r\right)^2\\&+1_k^{truth}*\lambda_{\text{obj}}*\left(IOU_{truth}^k-b_{ijk}^o\right)^2\\&+1_k^{truth}*\lambda_{\text{class}}*\left(\sum_{c=1}^{c}\left(truth^c-b_{ijk}^c\right)^2\right)\end{aligned} \tag{6}$$

这一部分计算与 ground truth 匹配的预测框的坐标损失, 置信度损失以及分类损失之和.

4 实验与分析

本次实验自建围栏跨越数据集, 共采集视频 70 段, 将每段视频按帧截取并分别保存到不同文件夹, 共包含图片 7000 余张. 使用 LabelMe 软件标注包含此动作的一系列帧生成相应 JSON 文件, 编写程序实现将多个 JSON 文件转化为训练所需的 txt 文件格式, 并汇总到 trainlist 文件中以开始训练.

实验显卡配置 NVIDIA GeForce RTX 2080Ti, 处理器为 Intel i7. 学习率初始化为 0.0001, 并在 30 k, 40 k, 50 k 和 60 k 次迭代后分别降低 0.5 倍.

4.1 效果展示

训练 12 个 epoch 后选取视频测试, 实验选取不同场景下的围栏跨越违规动作模拟视频以验证该方法的泛化性, 实际测试效果如图 6 所示.

从图中可以看出, 在不同的场景下使用此方法可以较为准确的检测出视频中的围栏跨越行为, 具有一定的泛化能力. 当处理实时监控时, 使用 OpenCV 不断获取视频监控截图, 将连续帧组成的剪辑作为输入, 经格式化处理后输入到训练好的模型中, 若检测到当前帧存在违规动作, 使用红色方框标记并预警, 避免事故的发生, 从而达到智能化管理.

4.2 对比实验

本次实验使用 Frame-AP 作为评价指标. 对所有包含预测框的帧, 计算每一个预测框与真实框间的 *IOU* 值, 若超过阈值 (预先设置为 0.5), 则记为 *TP*, 否则为 *FP*, 漏检记为 *FN*; 当出现多个预测框同时匹配一个真实框的情况时, 则只保留 *IOU* 值最大的预测框, 记为 *TP*, 其余均为 *FP*. 相应得出准确率 (*Precision*) 及召回率 (*Recall*).

$$Precision=\frac{TP}{TP+FP} \tag{7}$$

$$Recall=\frac{TP}{TP+FN} \tag{8}$$

为探究输入剪辑长度以及下采样率对围栏跨越检测模型的影响, 选取剪辑长度为 8 帧和 16 帧, 下采样率 d=1,2,3 进行对比实验. 具体结果如表 2 所示. 比较后, 本次实验选取输入剪辑长为 16, 下采样率 d 取 1.

探究不同 3D backbones 对结果的影响 (16-frames, d=1). 如表 3 所示. 从表 3 中可以看出, 在 3D-ResNext-101 基础上加入 SE 模块后, 达到最好效果, 因此将其作为本文三维卷积 backbone 使用.

探究本文 (3D-SE-ResNext-101+Darknet-19) 方法与其他方法在围栏跨越行为检测上的效果, 测试结果如表 4 所示.

从结果可以看出, 在保证 Frame-AP 的情况下, 本文方法在实际测试围栏跨越违规行为时处理速度可以达到 43 fps, 实时性更强.

5 结论与展望

针对围栏跨越违规行为检测问题, 本文从计算机视觉角度出发, 提出一种基于视频的智能化检测与识别算法, 使用三维卷积提取时序特征, 同时在二维卷积上提取空间特征, 解决定位问题. 通过设置对比试验以寻找最优方法. 实验测试结果表明, 该方法可以较为准确的检测出视频中的跨越行为, 具有较高的准确性和

鲁棒性, 大大提升了监管效率, 实现智能化管理. 未来将会考虑在此基础上加入目标检测模块, 重点检测围栏区域范围内的动作, 以消除无关区域动作干扰, 使围栏跨越违规检测与识别方法更加成熟.

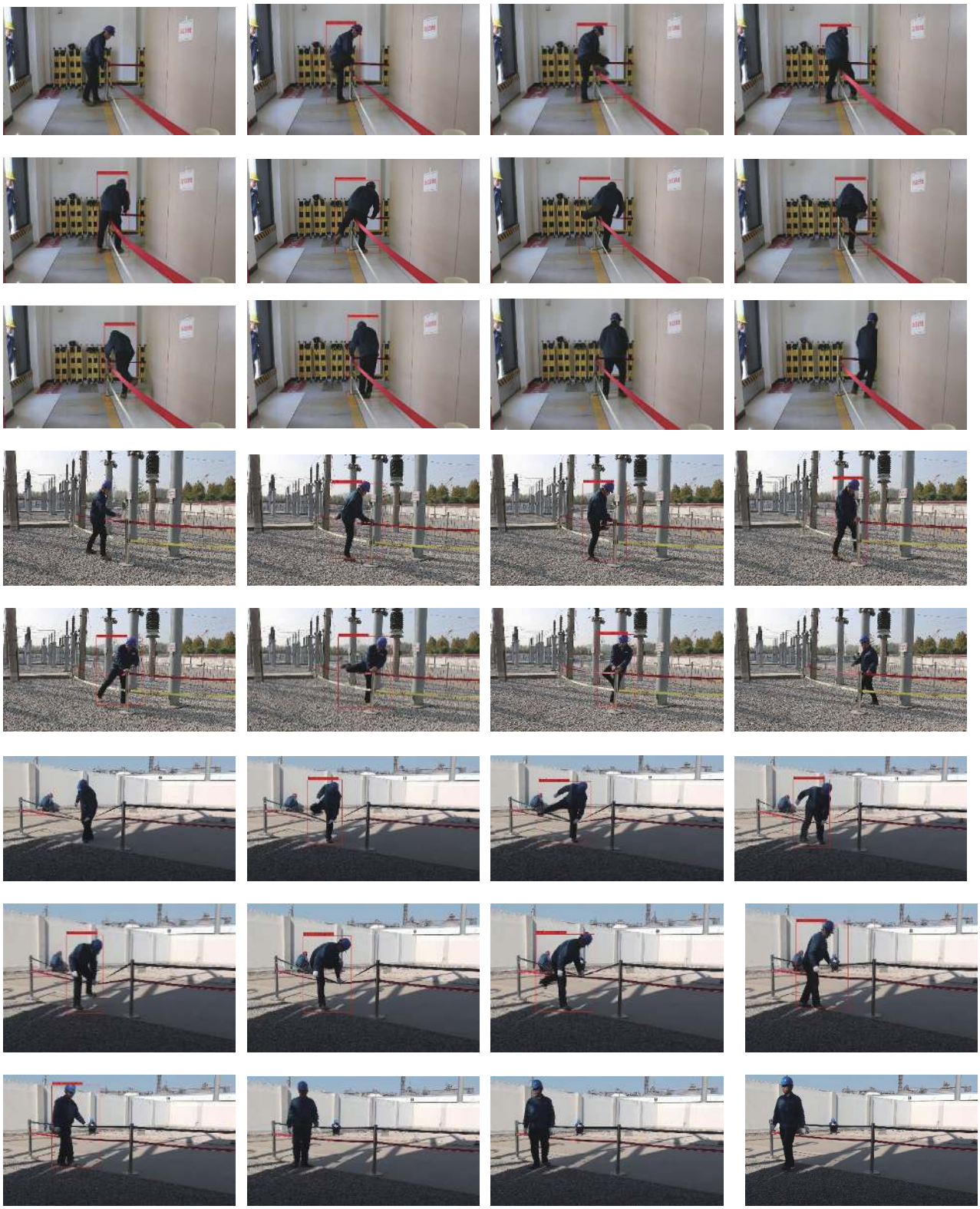

图 6 不同场景下的实际测试效果

表 2 输入长度及下采样率对结果的影响 (*IOU*=0.5)

输入	Frame-AP	输入	Frame-AP
8-frames (*d*=1)	87.6	16-frames (*d*=1)	88.3
8-frames (*d*=2)	86.5	16-frames (*d*=2)	87.1
8-frames (*d*=3)	83.1	16-frames (*d*=3)	84.3

表 3 3D Backbone 对结果的影响 (*IOU*=0.5)

Backbone	Frame-AP	Backbone	Frame-AP
3D-ResNet-18	76.7	3D-SE-ResNext-101	88.3
3D-ResNet-50	80.8	3D-ShuffleNetV1 2.0x	59.3
3D-ResNet-101	85.7	3D-MobileNetV1 2.0x	53.0
3D-ResNext-101	87.5		

表 4 不同方法测试效果 (*IOU*=0.5)

方法	速度 (fps)	Frame-AP
T-CNN	20	69.4
VideoCapsuleNet	32	83.5
本文	43	88.3

参考文献

1 罗会兰, 王婵娟, 卢飞. 视频行为识别综述. 通信学报, 2018, 39(6): 169–180. [doi: 10.11959/j.issn.1000-436x.2018107]

2 杨建全, 梁华, 王成友. 视频监控技术的发展与现状. 现代电子技术, 2006, 29(21): 84–88, 91. [doi: 10.3969/j.issn.1004-373X.2006.21.030]

3 朱英群. 基于移动办公的安全管控平台建设. 中小企业管理与科技 (中旬刊), 2018, (6): 155–156.

4 郭丽丽, 丁世飞. 深度学习研究进展. 计算机科学, 2015, 42(5): 28–33. [doi: 10.11896/j.issn.1002-137X.2015.05.006]

5 Zhang Z, Tao DC. Slow feature analysis for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3): 436–450. [doi: 10.1109/TPAMI.2011.157]

6 Tran D, Ray J, Shou Z, *et al*. Convnet architecture search for spatiotemporal feature learning. arXiv: 1708.05038, 2017.

7 Zhao R, Xu WR, Su H, *et al*. Bayesian hierarchical dynamic model for human action recognition. Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA. 2019. 7733–7742.

8 Zolfaghari M, Singh K, Brox T. ECO: Efficient convolutional network for online video understanding. arXiv: 1804.09066, 2018.

9 He DL, Li F, Zhao QJ, *et al*. Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition. arXiv: 1806.10319, 2018.

10 Guo GD, Lai A. A survey on still image based human action recognition. Pattern Recognition, 2014, 47(10): 3343–3361. [doi: 10.1016/j.patcog.2014.04.018]

11 杨国亮, 王志良, 牟世堂, 等. 一种改进的光流算法. 计算机工程, 2006, 32(15): 187–188, 226. [doi: 10.3969/j.issn.1000-3428.2006.15.066]

12 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, QC, Canada. 2014. 568–576.

13 Sevilla-Lara L, Liao YY, Guney F, *et al*. On the integration of optical flow and action recognition. arXiv: 1712.08416, 2017.

14 Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 1933–1941.

15 Feichtenhofer C, Pinz A, Wildes RP. Spatiotemporal residual networks for video action recognition. Proceedings of 29th Conference on Neural Information Processing Systems (NIPS 2016). Barcelona, Spain. 2016. 3468–3476.

16 Feichtenhofer C, Pinz A, Wildes RP. Spatiotemporal multiplier networks for video action recognition. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA. 2017. 4768–4777.

17 Ji SW, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221–231. [doi: 10.1109/TPAMI.2012.59]

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

19 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. 2017. 6299–6308.

20 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, Italy. 2017. 5533–5541.