

文章编号:1671-251X(2020)04-0075-06

DOI:10.13272/j.issn.1671-251x.2019080074

基于改进双流法的井下配电室巡检行为识别

党伟超¹, 张泽杰¹, 白尚旺¹, 龚大力², 吴喆峰²

(1. 太原科技大学 计算机科学与技术学院, 山西 太原 030024;
2. 精英数智科技股份有限公司, 山西 太原 030006)



扫码移动阅读

摘要: 井下配电室监控视频持续时间较长且行为类型复杂,传统双流卷积神经网络(CNN)法对此类行为识别效果较差。针对该问题,对双流 CNN 法进行改进,提出了一种基于改进双流法的井下配电室巡检行为识别方法。通过场景分析,将巡检行为分为站立检测、下蹲检测、走动、站立记录、坐下记录 5 种类型,并制作了巡检行为数据集 IBDS5。将每个巡检行为视频等分为 3 个部分,分别对应巡检开始、巡检中和巡检结束;对 3 个部分视频分别随机采样,获取代表空间特征的 RGB 图像和代表运动特征的连续光流图像,并分别输入空间流网络和时间流网络进行特征提取;对 2 个网络的预测特征进行加权融合,获取巡检行为识别结果。实验结果表明,以 ResNet152 网络结构为基础,且权重比例为 1:2 的空间流和时间流双流融合网络具有较高的识别准确度,Top-1 准确度达到 98.92%;本文方法在 IBDS5 数据集和公共数据集 UCF101 上的识别准确率均优于 3D-CNN、传统双流 CNN 等现有方法。

关键词: 井下配电室; 巡检行为识别; 双流卷积神经网络; 视频分段; 特征融合
中图分类号: TD67 **文献标志码:** A

Inspection behavior recognition of underground power distribution room based on improved two-stream CNN method

DANG Weichao¹, ZHANG Zejie¹, BAI Shangwang¹, GONG Dali², WU Zhefeng²

(1. School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China; 2. Jingying Shuzhi Technology Co., Ltd., Taiyuan 030006, China)

Abstract: The monitoring video of underground power distribution room has a long duration and complex behavior types, and the traditional two-stream convolutional neural network (CNN) has poor recognition effect on such behaviors. In view of the problem, the two-stream CNN method was improved, and a method of inspection behavior recognition of underground power distribution room based on improved two-stream CNN was proposed. Through scene analysis, the inspection behaviors are divided into five types: standing detection, squatting detection, walking, standing record, and sitting down record, and the inspection behavior dataset IBDS5 is produced. Each inspection behavior video is divided into three parts, corresponding to the start of inspection, middle inspection and end of inspection; RGB images representing spatial features and continuous optical flow images representing motion features are obtained by random sample from three parts of the video, and the images are input to spatial flow network and time flow network respectively for feature extraction; weighted fusion of predicted features of the two networks are performed to obtain inspection behavior recognition results. The experimental results show that the spatial-temporal and dual-stream fusion network based on ResNet152 network structure with a

收稿日期: 2019-08-25; **修回日期:** 2020-03-26; **责任编辑:** 胡娴。
基金项目: “十二五”山西科技重大专项资助项目(20121101001); 山西省中科院科技合作资助项目(20141101001); 山西省科技攻关资助项目(20141039); 山西省重点研发计划项目(201703D121042-1, 201803D121048)。
作者简介: 党伟超(1974—), 男, 山西运城人, 副教授, 博士, 主要研究方向为智能计算、软件可靠性, E-mail: dangweichao@tyust.edu.cn。
引用格式: 党伟超, 张泽杰, 白尚旺, 等. 基于改进双流法的井下配电室巡检行为识别[J]. 工矿自动化, 2020, 46(4): 75-80.
DANG Weichao, ZHANG Zejie, BAI Shangwang, et al. Inspection behavior recognition of underground power distribution room based on improved two-stream CNN method[J]. Industry and Mine Automation, 2020, 46(4): 75-80.

weight ratio of 1 : 2 has high recognition accuracy, and Top-1 accuracy reaches 98.92%; the recognition accuracy of the proposed method on the IBDS5 dataset and the public dataset UCF101 are better than existing methods such as 3D-CNN and traditional two-stream CNN.

Key words: underground power distribution room; inspection behavior recognition; two-stream CNN; video segmentation; feature fusion

0 引言

稳定的电力供应是保障煤矿安全生产的前提^[1]。井下配电室是煤矿电力供应系统的重要一环,因此,需要对井下配电室进行定期巡检,及时发现隐患并进行处理^[2]。当前判断人员是否按规定完成巡检任务,主要是通过检查纸质记录及监控室人工监视,存在巡检人员作弊、人工监视工作量大等问题。因此,有必要研究配电室巡检行为智能识别方法,以确保巡检人员按照规定完成巡检,保障煤矿电力系统安全。

对于煤矿井下配电室巡检行为检测来说,其主导因素是人,对人体行为的视觉分析显得尤为重要。卷积神经网络(Convolution Neural Network, CNN)^[3]由于其特殊的网络结构,特别适用于计算机视觉任务。当前主流的基于 CNN 的人体行为识别方法主要有以下 3 种:① 三维卷积神经网络(3D CNN)方法。Ji Shuiwang 等^[4]提出了 3D CNN 模型,将传统的二维卷积神经网络(2D CNN)模型拓展到 3D CNN 模型并提取空间和时间 2 个维度上的特征。D. Tran 等^[5]在文献[4]的基础上提出了可以基于视频提取特征的 C3D 网络。但是,3D CNN 模型的计算量很大,且当网络深度增加时,在数据样本不够大的情况下,容易产生过拟合,人体行为识别效果不佳。② CNN 与循环神经网络结合的方法。J. Donahue 等^[6]提出了长期循环卷积神经网络(Long-term Recurrent Convolutional Network, LRCN)。LRCN 的结构包括长短期记忆网络(Long Short-Term Memory, LSTM)和 CNN 两部分。LRCN 充分利用了视频空间维度和时间维度上的信息,但是只能避免梯度消失问题,未解决梯度爆炸问题。③ 双流 CNN 方法^[7]。所谓双流,即网络由分别处理空间(RGB 图像帧)和时间(堆叠光流帧)维度的 2 个 CNN 组成。该方法采用多任务训练方法将 2 个网络进行融合,获取行为特征,但是对长视频中的行为识别效果不佳。

井下配电室监控视频持续时间较长且行为类型复杂,传统双流 CNN 方法对此类行为识别效果较差。针对该问题,本文对双流 CNN 方法进行改进,提出了一种基于改进双流法的井下配电室巡检行为

识别方法。首先,将每个巡检视频等分为 3 个部分,分别对应巡检开始、巡检中和巡检结束;采用双流 CNN 方法对各部分视频进行分别处理,即随机采样获取代表空间特征的 RGB 图像及代表运动特征的 10 帧 X 方向和 10 帧 Y 方向的连续光流图像;将 RGB 图像和光流图像分别输入空间流 CNN 和时间流 CNN 进行特征提取,在各自流内对等分网络的输出特征进行融合;对 2 个流的预测特征进行加权融合,获取巡检行为识别结果。

1 巡检行为数据集

配电室巡检行为动作由一系列简单动作组合而成,为保证每类动作分类准确,类型间差别明显,对巡检动作进行了详细分解。根据煤矿配电室巡检有关规定,巡检人员应定时认真检查各种仪表、线路、进线柜、出线柜等,以发现电力设备运行隐患,保证配电设备安全运行。巡检过程中,巡检人员除站立查看、检测外,还需下蹲检测线路、设备等,确保完成规定检测;配电室范围较大,完成整个配电室的巡检需要多次走动;完成巡检后,需按规定进行记录,一般有站立记录和坐下记录 2 种情况。因此,将巡检行为分为 5 类,分别为站立检测、下蹲检测、走动、站立记录、坐下记录。完成这 5 类中的 4 类即可认为完成了巡检行为。

巡检行为发生在背景较为固定的井下配电室,且由多个简单动作组成,完成一次巡检行为需要数分钟。使用现有公开数据集进行模型训练,巡检动作识别效果较差,准确率不高,因此,专门制作了用于配电室巡检行为识别的数据集 IBDS5。IBDS5 数据集采集于现有的煤矿井下配电室视频监控系统。摄像头安装于配电室入口处正上方,距地面 2.5 m,光源对准配电室内部,确保能清晰地拍摄人物动作。共采集 21 段视频(共计 10.5 h),经过人工挑选从每段视频中提取包含 5 类动作的样本,共计提取 400 个视频样本,每类动作样本数约为 80 个,每个视频样本长度约为 5 s。巡检动作分类如图 1 所示。

2 巡检行为识别总体框架

站立检测和下蹲检测行为都可分为 3 个部分:与设备刚接触、进行设备检测和完成检测后准备离



图1 巡检动作分类

Fig. 1 Inspection action classification

开设备。如果对整个视频进行随机帧采样,就不能很好地学习此类动作的特征。因此,将视频分割成3个等长且不重叠的视频段,对每段视频分别进行RGB图像帧和堆叠光流帧提取。将获取到的图像帧分别输入对应的空间流网络和时间流网络进行特征提取,然后在各自流内对网络输出特征进行融合,最后对2个流的预测特征进行加权融合,获取巡检行为识别结果。巡检行为识别总体框架如图2所示。

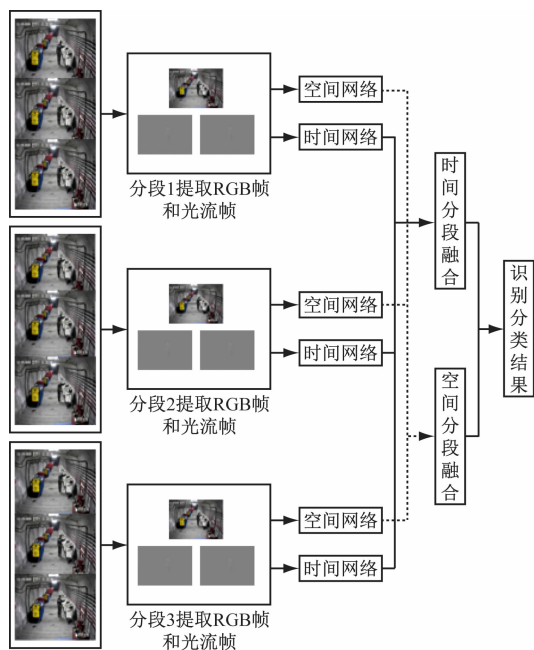


图2 巡检行为识别总体框架

Fig. 2 Overall framework for inspection behavior identification

将代表行为类别的视频分割为3个等长的视频段 V_1, V_2, V_3 , 基于分段采样的时空双通道 CNN 表

达式为

$$Y(V_1, V_2, V_3) = H(g(F(T_1; W), F(T_2; W), F(T_3; W))) \quad (1)$$

式中: H 为输出函数,用于对识别结果进行分类,得到每个行为类别的概率值,本文选用 softmax 函数; g 为聚合函数,用于对3个分段特征以均值方法进行融合,得到空间流或者时间流的特征; T_j 表示视频第 j 个分段的随机采样, $j=1, 2, 3$; $F(T_j; W)$ 表示用参数为 W 的 CNN 对 T_j 进行特征提取,时间流和空间流网络在3个视频片段上分别共享各自网络的一套参数 W 。

softmax 函数表达式为

$$H(G_i) = \frac{\exp(G_i)}{\sum_i \exp(G_i)} \quad (2)$$

式中 G_i 为分段共识函数, $G_i = g(F_i(T_1), F_i(T_2), F_i(T_3))$, $i=1, 2, \dots, C$, C 为行为分类的总类别数, $C=5$ 。

通过 softmax 函数预测得到整段视频被识别分类为每一类行为的概率。为避免训练过程太慢,结合标准分类交叉熵损失函数,得到关于分段共识函数 G_i 的损失函数为

$$L(y_i, G_i) = - \sum_{i=1}^C y_i (G_i - \ln \sum_{i=1}^C \exp(G_i)) \quad (3)$$

式中 y_i 为第 i 类行为的真实标签值。

模型是有梯度的,使用标准的反向传播算法,利用多个片段来联合优化参数 W 。在反向传播过程中,损失函数 $L(y_i, G_i)$ 对于模型参数 W 的梯度为

$$\frac{\partial L(y_i, G_i)}{\partial W} = \frac{\partial L}{\partial G_i} \sum_{k=1}^K \frac{\partial g}{\partial F(T_k)} \frac{\partial F(T_k)}{\partial W} \quad (4)$$

式中 K 为视频分段数, $K=3$ 。

网络是从3个视频段即完整的视频中学习模型参数,而不是从单一的短片段中学习模型参数。

3 巡检行为识别具体实现

3.1 空间流网络

通过空间流网络对视频中随机采样的静态 RGB 图像进行训练,提取巡检行为的空间特征。在静态图片识别过程中,目标物体的姿态与背景起着至关重要的作用。IBDS5 数据集采集于煤矿井下,背景固定且变化较少,能更好地实现空间特征提取。

卷积网络分类器以端到端的多层方式进行集成,通过网络层数量的叠加丰富图像特征。第一代经典的 LeNet 网络模型集成了5层卷积层,后来逐渐发展出8层的 AlexNet 模型、19层的 VggNet 模型及22层的 GoogLeNet 模型。大量实验表明,卷积层数的增加可以增强网络学习能力,提高图像分

类准确率。但是增加网络层数也会带来随机梯度消失问题,网络准确率达到饱和状态后会迅速下降,因此,CNN 的层数最多为 20。HE Kaiming 等^[8]提出了 ResNet 模型,并利用残差网络将 CNN 的层数增加到 152,将错误率降低了 3.75%。残差网络大大提高了图像识别率。ResNet 网络不仅可以加深 CNN 的层数,而且有效解决了因层数叠加导致的训练误差增大的问题。本文采用 ResNet152 提取图像的空间特征。

3.2 时间流网络

视频中的运动信息对于行为识别至关重要,光流因其简单实用并能表达图像序列运动信息被广泛用于提取行为运动特征。B. K. P. Horn 等^[9]推导出了图像序列光流的计算公式,因光流数值接近 0 且有正有负,为了能够作为时间流网络通道的输入,需要对其进行线性变换,最终将 X,Y 方向的光流保存为 2 张灰度图像。本文使用 TV-L1^[10] 方法提取视频的光流帧,提取结果如图 3 所示。

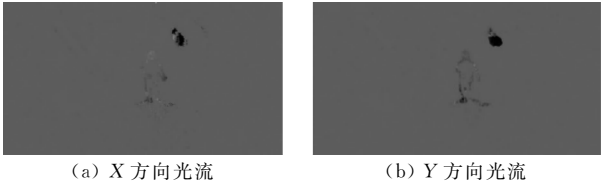


图 3 视频光流帧

Fig. 3 Video optical flow frames

时间流网络和空间流网络都采用 ResNet152 网络进行巡检行为识别。因为时间流和空间流网络都在 ImageNet^[11] 上进行预训练,所以第 1 个卷积层输入的通道数为 3。空间流网络输入的是 RGB 图像,不需要进行调整。而对于时间流来说,采用 10 帧 X 方向的连续堆叠光流帧和 10 帧 Y 方向的连续堆叠光流帧进行运动特征提取,相当于向网络输入 20 幅光流图像,与第 1 个卷积层的通道数不匹配。采用跨模态交叉预训练的方法解决该问题,即获取第 1 个卷积层的 3 个通道的权值后,取其平均值,再复制 20 份作为时间流网络第 1 个卷积层 20 个通道的权值,而时域网络其他层的权值与空域对应层的权值参数相同。

3.3 双通道特征融合

由式(1)可得整个空间流网络的行为识别结果 S_{spatial} 和时间流网络的行为识别结果 S_{temporal} 分别为

$$S_{\text{spatial}} = g(F(T_1;W),F(T_2;W),F(T_3;W)) \tag{5}$$

$$S_{\text{temporal}} = g(F(T_1;W),F(T_2;W),F(T_3;W)) \tag{6}$$

对 S_{spatial} 和 S_{temporal} 进行加权求和,得到最后的视频分类结果 S_{labels} :

$$S_{\text{labels}} = k_1 S_{\text{spatial}} + k_2 S_{\text{temporal}} \tag{7}$$

式中 k_1, k_2 分别为空间流网络和时间流网络权值,均为正整数。

3.4 数据增强和迁移学习

为了得到一个好的分类模型,机器学习方法需要足够的训练样本用于学习。IBDS5 数据集训练样本相对较少,为避免因训练样本数据不足造成过拟合情况,使用了数据增强和迁移学习技术。

通过数据增强技术,可扩大输入数据的规模,增加样本的差异性,并增强网络模型的泛化能力。对 RGB 图像帧和光流图像帧使用了角度翻转、平移变换、边角剪裁、尺度抖动^[12] 等数据增强方法。边角剪裁是指对输入图像的 1 个中心区域和边缘 4 个角落区域进行裁剪,裁剪后的图像与原图像差异性较大,网络输入的变化增加,因此,能有效减小过拟合的影响。尺度抖动是指将输入图像裁剪为固定尺寸,本文将输入图像从 $1\,920 \times 1\,080$ 修正为 256×340 ,然后裁剪宽和高,宽和高的尺寸在 $\{256, 224, 196, 168\}$ 中随机采样,再将裁剪区域尺寸修正为 224×224 后输入对应网络进行训练。

迁移学习是指利用已有的知识来解决不同但相似的问题,即使用预训练模型来克服目标任务数据不足的缺点。对于新目标任务,需要将预训练网络模型中最后一个用于分类的全连接层替换成新的针对目标任务类别数目的全连接层。本文采用的源域数据集为 UCF101 行为识别数据集,目标数据集为 IBDS5 数据集。首先对源域数据集进行处理,构建训练集和测试集,并进行预处理;将预处理后的数据输入 ImageNet 网络中进行预训练,保存参数;将预训练保存的参数加载到目标域神经网络模型(本文采用 VGG16, ResNet18/34/50/101/152)中,并将模型中最后一个用于分类的全连接层设置为对应的 5 类输出;最后对 IBDS5 数据集进行处理,并输入迁移后的神经网络模型中进行训练,得到行为识别结果。

4 实验分析

4.1 参数设置

基于 Pytorch1.1.0 深度学习框架进行实验,实验所使用的硬件及软件配置见表 1。

在实验过程中,将 IBDS5 数据集分为训练集、测试集和验证集 3 个部分。这 3 个部分之间没有交集,数据量比例为 3:1:1。

4.2 不同网络结构性能对比

4.2.1 空间流网络训练

将每个包含巡检行为的视频等分为 3 段短视频,从每段视频中随机抽取 RGB 图像,输入空间流网络进行特征提取,然后融合 3 段视频的识别结果,获得空间流网络的识别结果。

表 1 实验软硬件配置

Table 1 Experiment hardware and software configuration

项目	参数
CPU	Intel Xeon W-2012
内存	16 GB
GPU	NVIDIA RTX 2080Ti
操作系统	Windows 10
GPU 加速工具	CUDA10.0

空间流网络基本参数设置:初始学习率为 0.000 1,批尺寸 Batch-size 为 16,训练 2 000 代,每 50 轮测试 1 次,超参数 momentum 为 0.9。学习率是非常重要的一个超参数,甚至能左右模型性能,本文采用随机梯度下降优化算法,根据学习结果自动更新学习率。

为加快网络收敛速度并抑制神经网络过拟合现象^[13],在全连接层后增加一个丢包层,对网络进行优化。在训练过程中,随机丢弃神经网络单元及其连接,抑制过拟合现象。设置空间流网络的丢失率为 0.9。表 2 给出了网络结构为 VGG16,ResNet18,ResNet34,ResNet50,ResNet101,ResNet152 的空间流网络在 IBDS5 数据集上的准确率。其中 Top-1 准确率是指预测排名第一的类别与实际结果相符的概率。

表 2 空间流网络在 IBDS5 数据集上的准确率

Table 2 Accuracy of spatial flow network on dataset IBDS5

网络结构	Top-1 准确率/%	网络结构	Top-1 准确率/%
VGG16	90.87	ResNet50	92.39
ResNet18	91.53	ResNet101	93.04
ResNet34	92.82	ResNet152	94.47

从表 2 可看出,与其他网络结构相比,ResNet152 结构的空间流网络取得了最高的行为识别准确率,Top-1 准确率达到 94.47%。

4.2.2 时间流网络训练

时间流网络训练与空间流网络训练类似。将每个包含巡检行为的视频等分为 3 段短视频,从每段视频中随机抽取 10 帧 X 方向和 10 帧 Y 方向的连续堆叠光流帧,输入时间流网络进行学习,然后融合 3 段视频的识别结果,获得时间流网络的识别结果。

相对较大的初始学习率有利于网络快速收敛,因此,设置初始学习率为 0.001。丢失率设置为 0.8。其他参数与空间流网络参数相同。时间流网络在 IBDS5 数据集上的准确率见表 3。

从表 3 可看出,ResNet152 结构的时间流网络取得了最高的行为识别准确率,Top-1 准确率达到 96.22%。

表 3 时间流网络在 IBDS5 数据集上的准确率

Table 3 Accuracy of time flow network on dataset IBDS5

网络结构	Top-1 准确率/%	网络结构	Top-1 准确率/%
VGG16	90.20	ResNet50	94.49
ResNet18	91.25	ResNet101	95.18
ResNet34	93.45	ResNet152	96.22

4.3 权重比例确定

ResNet152 结构的时空流网络和时间流网络性能最好,因此,对其进行进一步加权融合。设置多种权重比例进行实验分析,结果见表 4。

表 4 双流网络在 IBDS5 数据集上的准确率

Table 4 Accuracy of two-stream network on IBDS5 dataset

$k_1:k_2$	Top-1 准确率/%	$k_1:k_2$	Top-1 准确率/%
1:1	97.60	2:3	98.04
1:2	98.92	3:1	95.90
1:3	98.71	3:2	97.54
2:1	97.04		

从表 4 可看出,当空间流和时间流的权重比例逐渐减小时,Top-1 准确率逐渐上升,这说明时间流网络提取的运动特征对巡检行为识别有更重要的作用。当 k_1 和 k_2 的比例为 1:2 时,Top-1 准确率最高,达到了 98.92%。而单独空间流网络使用 ResNet152 网络的 Top-1 准确率为 94.47%,单独时间流网络使用 ResNet152 网络的 Top-1 准确率为 96.22%,这说明集成双流网络的特征可以有效提升巡检行为识别性能。

4.4 不同方法性能对比

4.4.1 IBDS5 数据集实验

不同方法在 IBDS5 数据集上的准确率见表 5。3D-CNN 方法的 Top-1 准确率为 92.48%,这是由于巡检行为数据集样本较少,出现了过拟合问题,导致识别效果不佳。传统双流 CNN 方法的 Top-1 准确率为 94.27%,这是由于配电室巡检监控视频持续时间较长,而传统双流 CNN 方法对含有复杂动作的长视频识别效果不佳。双流 CNN+LSTM 方法^[14]的 Top-1 准确率为 95.86%,亦低于本文方法。

表 5 不同方法在 IBDS5 数据集上的准确率

Table 5 Accuracy of different methods on dataset IBDS5

方法	Top-1 准确率/%
3D-CNN	92.48
传统双流 CNN	94.27
双流 CNN+LSTM	95.86
本文方法	98.92

4.4.2 UCF101 数据集实验

为了进一步验证本文方法的性能,使用

UCF101 数据集进行实验。UCF101 数据集是目前行为识别领域应用最广泛的数据集,共包含 101 类动作,由 25 个人每人做 4~7 组动作,每个视频时长为 2~10 s,共 13 320 个视频,共 6.5 GB。

使用 UCF101 数据集训练时,采用 Split1 训练/测试分割方案,空间流网络和时间流网络的权重比例为 1 : 2,实验结果见表 6。

表 6 不同方法在 UCF101 数据集上的准确率
Table 6 Accuracy of different methods on dataset UCF101

方法	Top-1 准确率/%
3D-CNN	85.23
双流法	88.07
双流法+LSTM	88.64
本文方法	92.83

对比表 5 和表 6 可知,各种方法在 IBDS5 数据集上的识别准确率均高于在 UCF101 数据集上的识别准确率。其原因在于 IBDS5 数据集的背景较固定,而 UCF101 数据集的背景变化较大,且人员遮挡问题较严重。

5 结论

(1) 提出了一种基于改进双流法的井下配电室巡检行为识别方法。自制配电室巡检行为数据集 IBDS5,通过实验讨论并分析了不同网络结构和不同权重比例集成策略对识别准确率的影响。

(2) 实验结果表明,以 ResNet152 网络结构为基础,且权重比例为 1 : 2 的空间流和时间流双流融合网络具有较高的识别准确率,Top-1 准确率达到 98.92%;本文方法在 IBDS5 数据集和公共数据集 UCF101 上的识别准确率均优于 3D-CNN、传统双流 CNN 等现有方法。

(3) 本文方法仍存在以下不足:当多个巡检人员同时巡检时,巡检人员之间可能出现相互遮挡的情况,造成误识别和未识别问题。下一步工作将在多人巡检行为识别方面进行研究,以满足实际需求。

参考文献 (References):

[1] 孙继平,靳春海,曹玉超. 基于视频图像的矿井水灾识别及趋势预测方法研究[J]. 工矿自动化,2019,45(7):1-4.
SUN Jiping, JIN Chunhai, CAO Yuchao. Research on mine flood identification and trend prediction method based on video image [J]. Industry and Mine Automation, 2019, 45(7): 1-4.

[2] 孙继平,钱晓红. 2004—2015 年全国煤矿事故分析[J]. 工矿自动化,2016,42(11):1-5.
SUN Jiping, QIAN Xiaohong. Analysis of coal mine

accidents in China during 2004–2015[J]. Industry and Mine Automation, 2016, 42(11): 1-5.

[3] 李岳云,许悦雷,马时平,等. 深度卷积神经网络的显著性检测[J]. 中国图象图形学报,2016,21(1): 53-59.
LI Yueyun, XU Yuelei, MA Shiping, et al. Saliency detection based on deep convolutional neural network [J]. Journal of Image and Graphics, 2016, 21(1): 53-59.

[4] JI Shuiwang, YANG Ming, YU Kai. 3D convolutional neural networks for human action recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[5] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//Proceedings of IEEE International Conference on Computer Vision, Santiago, 2015.

[6] DONAHUE J, HENDRICKS L A, ROHRBACH M, et al. Long-term recurrent convolutional networks for visual recognition and description [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(4): 677-691.

[7] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [C]//Proceedings of the 2014 Conference on Neural Information Processing Systems, New York, 2014.

[8] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition [C]//IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016.

[9] HORN B K P, SCHUNCK B G. Determining optical flow [J]. Artificial Intelligence, 1981, 17(1-3): 185-203.

[10] WANG L, XIONG Y, WANG Z, et al. Towards good practices for very deep two-stream ConvNets [J]. Computer Science, 2015(7): 1-5.

[11] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database [C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, 2009.

[12] ZACH C, POCK T, BISCHOF H. A duality based approach for realtime TV-L1 optical flow [M]. Heidelberg: Springer, 2007.

[13] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[14] YUE-HEI N J, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: deep networks for video classification [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.