DOI:10.19651/j.cnki.emt.2209351

基于改进的 YOLOv5 网络的异常行为检测算法研究*

王 雪[1] 程换新[1] 骆晓玲[2] 高 宇[1]
(1.青岛科技大学自动化与电子工程学院 青岛 266061; 2.青岛科技大学机电工程学院 青岛 266061)

摘 要: 各行各业安全问题尤为重要,对人员的异常行为须及时检测并采取相应的措施才能有效预防安全事故的发生。因此本文提出基于改进的 YOLOv5 网络的异常行为识别算法,通过实时处理视频监控中人员的异常行为,从而保证企业的安全运行。首先,对输入数据集进行特征提取处理,本文使用 YOLOv5 的 backbone 特征提取网络提取视频特征,能够在不同图像细粒度上聚合并形成图像特征;其次,送入到时间注意块,因为不同时刻特征的贡献值并不相同,因此加入此模块赋予特征不同的贡献值;最后,送入特征预测网络,该网络由 LSTM 搭建,对历史特征序列进行解码,以预测当前的特征。以玩手机和吸烟为例对所提出的网络进行验证,训练集准确率高达 96.42%,测试集准确率高达 95.21%。

关键词: 异常行为;YOLOv5; transformer;时间注意块
中图分类号: TP183 **文献标识码:** A **国家标准学科分类代码:** 520.60

Research on abnormal behavior detection algorithm based on improved YOLOv5 network

Wang Xue[1] Cheng Huanxin[1] Luo Xiaoling[2] Gao Yu[1]
(1. College of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266061, China;
2. College of Mechanical and Electrical Engineering, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: Safety problems in all walks of life are particularly important. Abnormal behaviors of personnel must be detected in time and corresponding measures must be taken to effectively prevent safety accidents. Therefore, this paper proposes an abnormal behavior recognition algorithm based on the improved yolov5 network, which can ensure the safe operation of the enterprise by dealing with the abnormal behavior of personnel in video monitoring in real time. Firstly, feature processing is carried out on the input data set. In this paper, the backbone feature extraction network of yolov5 is used to extract video features, which can aggregate and form image features on different image granularity; Secondly, it is sent to the time attention block. Because the contribution values of the features at different times are different, this module is added to give different contribution values to the features; Finally, it is sent to the feature prediction network, which is built by LSTM to decode the historical feature sequence to predict the current feature. Taking playing mobile phone and smoking as examples, the accuracy of the proposed network is as high as 96.42% in the training set and 95.21% in the test set.

Keywords: abnormal behavior; YOLOv5; transformer; time attention block

0 引 言

当今,在石油、化工等相关行业中严禁人员使用明火、玩手机和抽烟等危险行为。与传统人工分析视频的方法相比较,利用算法对视频进行智能监控会更快捕捉到这些危险行为并及时预警[1]。

随着社会保障体系的不断完善,公共摄像头已经形成了一个庞大的监控网络。利用监控视频进行异常行为检测已经成为计算机视觉的一个重要研究领域。异常行为检测可以应用于学校、街道和医院等公共监控系统[2-14]。传统的人工异常行为检测涉及视频采集、异常行为检测等多个步骤。在检测阶段需要巨大的人力物力,而且很容易因为监控人员的疏忽而漏检。因此,越来越多的研究人员对基于不同计算机视觉方法的异常行为自动检测进行了大量研究。

收稿日期:2022-03-21
* 基金项目:国家海洋局重大专项(国海科字[2016]494 号)资助

目前，传统的异常行为检测方法主要分为动态贝叶斯网络（DBNs）[15-18]、概率主题模型（PTMs）[19]、聚类模型[20-21]和稀疏表示方法[6,10,22-25]。隐马尔可夫模型及其变体是所有 DBN 中最常用的异常检测方法。这些方法一般将人类行为表示为一组状态向量，概率被用作行为模式与测试序列之间的相似度。期望最大化算法和最大概率估计和最大后验估计分类器用于检测异常行为。为了降低 DBN 巨大的计算成本，许多专家学者提出使用 PTM 进行异常行为检测。这两种方法最初是为文本挖掘而提出的。在异常行为检测过程中，一般通过推断隐藏变量的潜在分布来寻找视觉文档的主题分布，将那些出现概率较低的视觉文档视为异常行为：聚类模型，包括 K-means、Markov 集群（MCL）和其他方法。

随着深度学习在计算机视觉领域的不断发展，学者们开始用深度学习进行异常行为的检测。Yi 等[26]将全局特征与局部特征融合同时使用卷积稀疏编码对异常行为进行分类，取得了比较好的检测效果。Ji 等[27]提出了 3D CNN，通过 3D 卷积捕获沿空间和时间维度的特征，从而获得相邻帧之间的信息。以上算法均是有监督学习，会导致复杂度增加，因此大量学者开始转向研究无监督式深度学习算法。Tian 等[28]将多尺度直方图光流和显著性信息结合视频帧并采用 PCANet 提取异常事件检测的高级特征。Jian 等[19]提出了基于深度学习的人-物交互行为时便算法，使用 YOLOv3 检测感兴趣物体，提出决策融合策略得到最终的行为识别结果。Xia 等[29]提出了具有新的时间注意机制的特征预测框架，但该网络的特征提取部分使用的是 Vgg16 网络。由于 VGG 是双阶段神经网络，检测速度相对较慢，且随着神经网络的进一步发展，单阶段神经网络 YOLO 系列在准确率表现很好，且 Xia 等提出的框架准确率较低，为了提高目标检测的效率，本文在 Xia 等提出的网络基础上，选用 YOLOv5 网络并改进对异常行为进行检测。

1　模型介绍

本文所使用的模型为改进的 YOLOv5 模型，包括 3 个结构：特征提取网络，时间注意块[28]和特征预测网络，具体结构如图 1 所示。

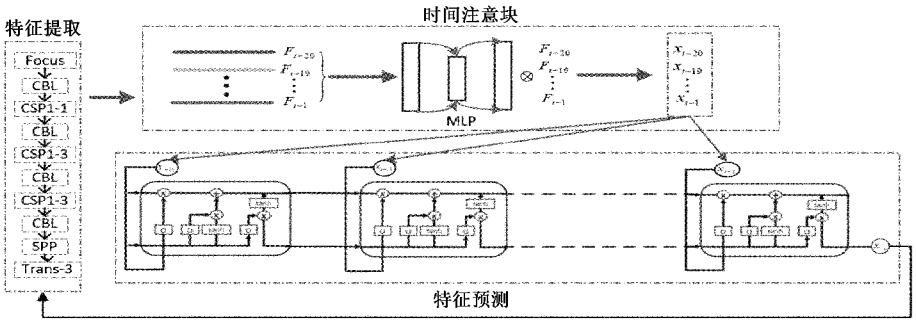

图 1　改进的 YOLOv5 结构

1.1　特征提取网络

特征提取网络保留了原 YOLOv5 网络的基准网络，一般都采用性能良好的分类器网络，主要由 Focus、CBL、CSP、SPP 和 Transformer[24] 3 种结构组成，其中 CBL、CSP 和 SPP 的结构组成如图 2 所示。

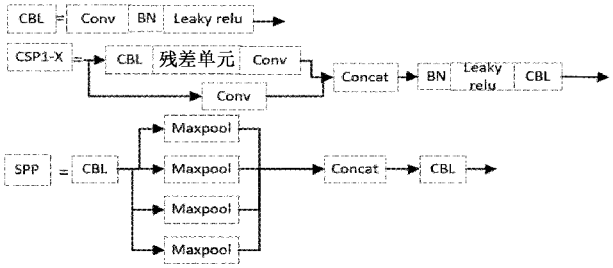

图 2　CBL、CSP 和 SPP 网络

Transformer 主要分为 4 个部分：输入，编码结构，解码结构和输出。输入：一个序列的数据，一般该结构常用于语言识别中。编码结构：由 6 个编码堆叠组成即 $N_x=6$，也就是上图中框内的结构，解码结构：与编码结构类似。输出：经过一次线性变化再通过 Softmax 得到输出的概率分布。

1.2　时间注意块

特征提取网络对每个训练视频中正常行为的特征进行提取。为了捕捉正常行为的运动特征，本文使用 LSTM 网络来提取运动模式。考虑到不同时间的特征对当前特征的贡献是不同的。因此，本文使用了时间注意块来探索不同时刻的不同贡献。

首先，将特征提取网络提取的时间特征序列按照时间轴 $L_t=\{F_{t-20},F_{t-19},\cdots,F_{t-1}\}$ 进行拼接。Zhou 等[30]建议使用平均池来学习目标对象的范围。Hu 等[31]在他们的注意力块中采用它来计算空间特征。对于本文得到的张量，在每个时间通道上使用最大池化和平均池化得到两个向量 $\boldsymbol{F}_{avg},\boldsymbol{F}_{max}\in R_{20}$。在获得两个向量 $\boldsymbol{F}_{avg}$ 和 $\boldsymbol{F}_{max}$ 后，它们被发送到具有隐藏层的共享多层感知器。然后，使用逐元素求和来合并输出向量以获得时间注意向量 $\boldsymbol{M}_t$。并将其与原始时间特征序列 L_t 相乘，得到具有时间注意力的序列 $T_t=\{x_{t\,20},x_{t\,19},\cdots,x_{t\,1}\}$。时间注意力块可以抽象如下：

$$\boldsymbol{M}_t(L_t)=\sigma(\boldsymbol{MLP}(\boldsymbol{F}_{avg}+\boldsymbol{MLP}(\boldsymbol{F}_{\max})))=\sigma(W_1(W_0(\boldsymbol{F}_{avg}))+W_1(W_0(\boldsymbol{F}_{\max}))) \tag{1}$$

其中,σ 是 sigmoid 激活函数,$\boldsymbol{W}_0\in R^{10}$ 和 $\boldsymbol{W}_1\in R^{20}$ 是 MLP 的隐藏层和输出层的权重向量。

1.3　特征预测块

经过时间注意力块后,得到时间序列 T_t。序列 T_t 暗示了同一空间位置的历史时刻特征对当前时间特征的贡献。对当前特征贡献较大的特征将得到较高的关注。为了捕捉历史时刻和当前时刻的特征之间的关系,本文引入了长短期记忆(LSTM)[32] 网络进行特征预测。该预测网络将具有时间注意力的历史特征作为输入,并将当前时刻的特征作为特征预测输出。

LSTM 单元的网络结构如图 3 所示,每个门的输出如式(2)~(7)所示。

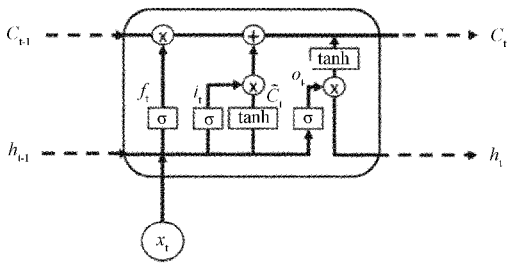

图 3　LSTM 单元的网络结构

$$f_t=\sigma(W_f\cdot[h_{t-1},x_t]+b_f) \tag{2}$$

$$i_t=\sigma(\boldsymbol{W}_i\cdot[h_{t-1},x_t]+b_i) \tag{3}$$

$$\tilde{C}_t=\tanh(\boldsymbol{W}_c\cdot[h_{t-1},x_t]+b_c) \tag{4}$$

$$C_t=\sigma(f_t\times C_{t-1}+i_t\times\tilde{C}_t) \tag{5}$$

$$o_t=\sigma(\boldsymbol{W}_o\cdot[h_{t-1},x_t]+b_o) \tag{6}$$

$$h_t=\sigma_t\times\tanh(C_t) \tag{7}$$

其中,$\boldsymbol{W}_*$ 表示为门对应的权重向量,b_* 表示为门对应的偏差。本文使用 LSTM 单元的隐藏状态 h_t 作为最终输出。

2　实验分析

2.1　实验环境及参数设置

实验环境为 Windows10 操作系统;CPU 为 Intel(R) Core(TM) i7-9750H;GPU 为 GeForce GTX 1650。

参数设置:学习率设为 0.000 1,batch-size 设为 32,epoch 为 3 000。

2.2　数据集

本文研究的基于改进 YOLOv5 的异常行为检测算法研究,主要检测人员玩手机和吸烟来验证所提出的网络,通用的数据集不适合用来验证本文所提出的方法。因此,我们自建视频数据集。图 4 为该数据集的部分示例:包括在办公室玩手机,在精馏塔附近(复杂环境)玩手机,办公室吸烟及边玩手机边吸烟等。

按照一定比例将自制数据集随机划分为训练集、验证集和测试集。训练集和验证集共有 300 个片段,每个片段 20 帧;测试集有 30 个片段,每个片段 20 帧。

图 4　数据集部分帧展示

2.3　实验结果与分析

1)由于硬件性能影响,共训练了 3 000 代。此时训练集准确率为 96.42%,测试集准确率为 95.21%,网络的损失值曲线如图 5 所示,可以看出在训练过程中损失值呈收敛趋势,说明改进是可行的。

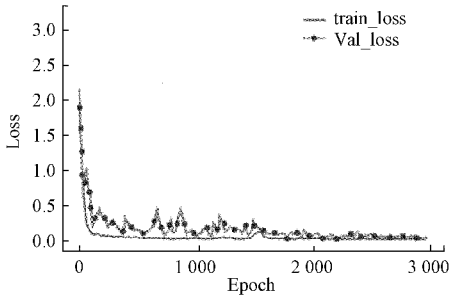

图 5　损失值曲线

2)为了直观表明本文的实验结果,图 6 为部分测试集输出的实验结果,图中数字为置信度,可以看出改进的网络识别比较准确,平均置信度可以达到 0.8,能够达到预期的效果。

图 6　测试集实验结果

3)最后,将改进的网络模型与其他模型相比较,采用mAP(平均精度)和检测速度 FPS 作为评估指标。与其他方法比较,本文所提出的方法在两个指标上都有提高,可以看出本文的改进在速度可准确率上都得到了提升,充分验证了本文所提出的网络达到了预期的目标。具体数据如表 1 所示。

表 1　改进的 YOLOv5l 与其他模型对比

模型	mAP/%	FPS
YOLOv5	59.87	13.02
机器视觉[31]	60.65	22.87
改进的 YOLOv3[32]	62.95	31.52
LSTM with TAM[33]	64.21	37.53
本文	**65.79**	**39.64**

3　结　　论

本文提出了一种基于改进的 YOLOv5 网络,通过引入时间注意块为了探索不同时刻的不同贡献和 LSTM 结构为了捕捉历史时刻和当前时刻的特征之间的关系,能够及时的对人员异常行为检测,有效的预防企业安全事故的发生。实验结果表明,本文方法的识别精度和检测速度均优于其他算法,可以对玩手机和吸烟等异常行为的准确识别。但在此过程中仍有一些误识别的情况发生,因此在未来的研究中,会继续改进模型。

参考文献

[1] XIA L, Li Z. A new method of abnormal behavior detection using LSTM network with temporal attention mechanism [J]. The Journal of Supercomputing, 2021, 77(4): 3223-3241.

[2] XIE S, ZHANG X, CAI J. Video crowd detection and abnormal behavior model detection based on machine learning method [J]. Neural Computing & Applications, 2018,31: 175-184.

[3] SHEN M, JIANG X, SUN T. Anomaly detection based on nearest neighbor search with locality-sensitive B-tree[J]. Neurocomputing, 2018, 289(10): 55-67.

[4] XING H, HUANG Y, DUAN Q, et al. Abnormal event detection in crowded scenes using histogram of oriented contextual gradient descriptor[J]. EURASIP Journal on Advances in Signal Processing, 2018,DOI: 10.1186/s13634-018-0574-4.

[5] XU K, JIANG X, SUN T. Anomaly detection based on stacked sparse coding with intraframe classification strategy [J]. IEEE Transactions on Multimedia, 2018: 1062-1074.

[6] SABOKROU M, FAYYAZ M, FATHY M, et al. Deep-cascade: Cascading 3D deep neural networks for fast anomaly detection and localization in crowded scenes[J]. IEEE Transactions on Image Processing, 2017, 26(4): 1992-2004.

[7] COAR S, DONATIELLO G, BOGORNY V, et al. Toward abnormal trajectory and event detection in video surveillance[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2017, 27(3): 683-695.

[8] YE O, DENG J, YU Z, et al. Abnormal event detection via feature expectation subgraph calibrating classification in video surveillance scenes[J]. IEEE Access, 2020, (99): 1-1,DOI:10.1109/ACCESS.2020.2997357.

[9] YU B, LIU Y, SUN Q. A content-adaptively sparse reconstruction method for abnormal events detection with low-rank property [J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017.

[10] MEHRAN R, OYAMA A, SHAH M. Abnormal crowd behavior detection using social force model [C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[11] FERNANDO T, DENMAN S, SRIDHARAN S, et al. Soft + hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection[J]. Neural networks : The official journal of the International Neural Network Society, 2018,DOI:10.1016/j.neunet.2018.09.002.

[12] ULLAH A, MUHAMMAD K, SER J D, et al. Activity recognition using temporal optical flow convolutional features and multilayer LSTM[J]. IEEE Transactions on Industrial Electronics, 2019, 66(12): 9692-9702.

[13] MARTINEL N, MICHELONI C, PICIARELLI C, et al. Camera selection for adaptive human-computer interface[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2014, 44(5): 653-664.

[14] SABOKROU M, FATHY M, HOSSEINI M. Real-time anomaly detection and localization in crowded scenes[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015: 56-62.

[15] B L L A, C S W, D B H, et al. Learning structures of interval-based Bayesian networks in probabilistic generative model for human complex activity recognition[J]. Pattern Recognition, 2018, 81: 545-561.

[16] EPAILLARD E, BOUGUILA N. Variational bayesian learning of generalized dirichlet-based hidden markov models applied to unusual events detection[J].

Neural Networks and Learning Systems, 2018, 30(4): 1034-1047.

[17] KAN O, GHARTI S, DAILEY M N. Incremental behavior modeling and suspicious activity detection[J]. Pattern Recognition, 2013, 46(3): 671-680.

[18] KRATZ L, NISHINO K. Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models[C]. IEEE Conference on Computer Vision & Pattern Recognition, IEEE, 2016: 1446-1453.

[19] JIAN L, GONG S, TAO X. Global behaviour inference using probabilistic latent semantic analysis[C]. Proceedings of the British Machine Vision Conference DBLP, 2008.

[20] ISUPOVA O, KUZIN D, MIHAYLOVA L. Learning methods for dynamic topic modeling in automated behavior analysis[J]. IEEE Trans Neural Netw Learn Syst, 2018, 29(9): 3980-3993.

[21] HOSPEDALES T, GONG S, XIANG T. Video behavior mining using a dynamic topic model[J]. Int J Comput Vis, 2012, 98(3): 303-323.

[22] TANG X, ZHANG S, YAO H. Sparse coding based motion attention for abnormal event detection[C]. 2013 IEEE International Conference on Image Processing. IEEE, 2013: 3602-3606.

[23] REN H, MOESLUND T B. Abnormal Event Detection Using Local Sparse Representation[C]. AVSS, 2014: 125-130.

[24] HE C, SHAO J, SUN J. An anomaly-introduced learning method for abnormal event detection[J]. Multimed Tools Appl, 2018, 77(22): 29573-29588.

[25] SUN J, WANG X, XIONG N, et al. Learning sparse representation with variational autoencoder for anomaly detection[J]. IEEE Access, 2018, 6: 33353-33361.

[26] YI W, LIM J, YANG M. Online object tracking: A benchmark supplemental material[J]. IEEE, 2013, DOI:10.1109/CVPR.2013.312.

[27] JI S, XU W, YANG M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2013, 35(1):221-231.

[28] TIAN Y L, BROWN L, HAMPAPUR A, et al. IBM smart surveillance system (S3): Event based video surveillance system with an open and extensible framework[J]. Machine Vision & Applications, 2017, 19(5-6): 315-327.

[29] XIA L, LI Z. A new method of abnormal behavior detection using LSTM network with temporal attention mechanism[J]. The Journal of Supercomputing, 2021, 77(4): 3223-3241.

[30] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[J]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 2921-2929.

[31] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[J]. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[32] 杨斌，云霄，董锴文，等. 基于机器视觉的石化场景人员危险行为识别[J]. 激光与光电子学进展，2021，58(22):11,DOI:10.3788/LOP202158.2215001.

[33] FANG M T, PRZYSTUPA K, CHEN Z J, et al. Examination of Abnormal Behavior Detection Based on Improved YOLOv3[J]. Electronics, 2021, 10(2): 197,DOI:10.3390/ELECTRONICS10020197.

作者简介

王雪，研究生，主要研究方向为人工智能、图像识别等。
E-mail: 2388861238@qq.com

程换新，工学博士，教授，主要研究方向为控制科学与工程、人工智能、图像识别等。
E-mail: 2635510239@qq.com

骆晓玲(通信作者)，工学博士，教授，主要研究方向为过程设备及控制的设计研究等。
E-mail: xiaolingluo@126.com

高宇，本科生，主要研究方向为图像识别。
E-mail: 1622844945@qq.com