基于YOLO网络模型的异常行为检测方法研究

刘雪奇[1,2,3,4],孙胜利[1,4]
(1.中国科学院上海技术物理研究所 上海 200083;2.中国科学院大学 北京 100049;
3.上海科技大学 上海 201210;4.中国科学院红外探测与成像技术重点实验室 上海 200083)

摘要: 针对监控视频中人体异常行为的复杂多样难检测问题,提出了基于YOLO网络模型的异常行为检测方法。根据对监控场景的异常行为定义需求,将标定的异常行为通过YOLO网络模型进行训练,不进行人体目标的提取而将其放到神经网络中,直接实现端到端的异常行为分类,从而实现对具体应用场景的异常行为检测。实验结果表明,该方法召回率接近100%并且平均精确率达到96%以上,同时通过GPU加速对于视频流的检测速度可以达到30FPS左右,实现对监控视频异常行为的实时检测。

关键词: YOLO;异常行为;识别框;IOU;召回率;精确率
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 1674-6236(2018)20-0154-05

Research on abnormal behavior detection based YOLO network

LIU Xue-qi[1,2,3,4],SUN Sheng-li[1,4]
(1.*Shanghai Institute of Technical Physics of the Chinese Academy of Sciences*,*Shanghai* 200083,*China*;
2. *University of Chinese Academy of Sciences*,*Beijing* 100049,*China*;3. *ShanghaiTech University*,
Shanghai 201210,*China*;4. *Key Laboratory of Infrared System Detection and Imaging Technology of the Chinese Academy of Sciences*, *Shanghai* 200083, *China*)

Abstract: Aiming at the problem of complex and difficult detection of human abnormal behavior in surveillance video, an abnormal behavior detection method based on YOLO network model was proposed. According to the demand of the definition of abnormal behavior for monitoring scene, the labeled abnormal behaviors were trained by YOLO network; without the extraction of the human target, put it into the neural network and achieved end-to-end abnormal behavior classification directly, so as to realize the abnormal behavior detection of a specific application scenario. The experimental results show that the recall rate of this method is close to 100% and the recognition rate is reached, and real-time monitoring of monitoring video can be realized through GPU acceleration. Experimental results show that the recall of this method is close to 100% and the average precision is over 96%. In addition, the video stream detection speed can reach about 30FPS and real-time detection of abnormal behavior for video surveillance can be realized by GPU acceleration.

Key words: YOLO; abnormal behavior; bounding box; IOU; recall; precision

与传统的人工监控相比,智能视频监控可有效地降低误检和漏检概率,但目前监控视频中对各种异常的排查和报警大多依赖于人工,智能化的人体异常行为检测技术还处于发展中[1]。同时由于监控中人体和背景的复杂多样,对于特定场景下的检测就变的尤为困难。

对于某一特定监控场景,涉及到的人体异常行为通常较复杂,包括目标远近大小、重叠遮挡,背景纷乱复杂等[2]。这些都会对异常行为的检测产生很大影响,通常采用的方法是先将目标通过轮廓信息从视频序列中分割出来,然后进行特征提取,将提取

收稿日期: 2018-02-05 **稿件编号:** 201802026
基金项目: 中国科学院上海技术物理研究所2015年所创新专项资助项目(No.CX-63)
作者简介: 刘雪奇(1991—),男,山东潍坊人,硕士研究生。研究方向:深度学习与信息图像处理。

到的特征与标准异常行为样本进行比对，最后利用分类器判断是否为异常行为[3]；包括基于区域光流能量的检测方法[4]也是先对目标进行提取然后作分析。然而对于特定场景下，人体异常行为存在复杂多样、较难明确定义的问题，对此，一种简单的解决方式是将人体行为分为两类，忽略中间的过渡行为分为正常行为和异常行为。因为对于如实验室这样的特定场景，异常行为识别的需求可能只是是否穿戴工作服和工作帽这类简单的分类问题，所以分为两类可以很好的解决异常行为定义的问题。

传统智能视频分析技术由于采用人工选择特征，存在准确率低、浅层学习无法解析大数据等问题，而深度学习可以很好地克服这些问题，使视频分析过程中识别准确率更高、鲁棒性更好、识别种类更丰富[5]。本文将深度学习应用到监控中的人体异常行为检测，通过将复杂的异常行为输入到深度神经网络YOLO中进行自动特征提取和分类，将目标提取这一步交给神经网络，与之后的目标分类同时放到一个网络中，利用神经网络的深层次特征提取、高精度检测分类特性，可以将定义的异常行为准确的检测出来，实现从输入数据到输出检测结果的端到端的异常行为检测。通过GPU对检测过程进行加速，可以实现对监控视频的实时检测。

1　方法描述

文中选取了实验室监控场景进行相关的实验研究工作，具体的方法流程如图1所示。

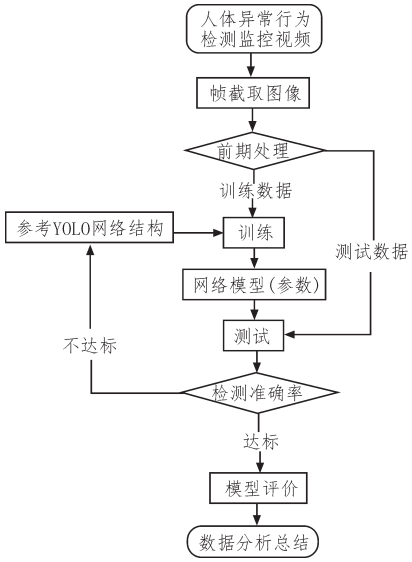

图1　实验方法流程图

首先要获得监控场景的数据视频，通过帧截取图像、筛选；然后根据需求对异常行为的定义用LabelImg软件进行标定，此处标定的标签不是为了目标提取，而是直接标定为是否异常的两类，从而让目标提取和分类都交给网络去做，获得网络训练和测试的图像数据；参考YOLO的网络结构，将训练数据输入到YOLO网络中进行训练，从而直接训练出可以对输入图像或视频数据进行判断是否存在异常行为的网络模型。

1.1　数据处理

实验所采用的数据是实验室上网工作室中的监控数据视频，对于不规范上网的异常行为进行检测。在输入网络进行训练之前，数据需要进行一系列处理，包括视频的帧截取，图像筛选，标签标定。实际工作中用到的处理软件有Smart Player和LabelImg，Smart Player软件用来将视频数据进行转换和截取图像；LabelImg软件用来标定图像标签。

其中图像的标签标定涉及到对异常行为的定义，由于异常行为多样，较难泛化为一类或两类，因此实验中将除了正常的上网行为动作以外的其他行为都定义为异常行为。例如除了正常行走和端坐，其他包括弯腰、下蹲、伸展肢体、低头玩手机等行为都定义为异常行为。实验中标签分为两类：normal和abnormal；一共标定了1 146张图像数据，1 000张用作训练数据，其余的作为测试数据。

1.2　采用的网络结构

YOLO是一种新的目标检测方法[6]，由Joseph Redmon[7]等人提出，这种方法的突出优势在于目标检测速度很快，同时又能保持较高的识别准确率，而且背景误检率低。其主要思路是将目标检测问题看作一个回归问题，只采用单个神经网络来回归从整张图的输入到目标边界和类别概率的输出，实现端到端的直接预测，即端到端的目标检测。本文借鉴这一回归思想，将目标检测网络结构YOLO应用到分类问题，将目标区域的类别作为最终结果，把异常行为的检测当做回归问题来解决。此外，YOLO的检测速度非常快，可以达到45帧/s的实时检测。

从R-CNN网络[8]、Fast R-CNN网络[9]到Faster R-CNN网络[10]采用的思路都是proposal+分类（proposal提供目标位置信息，分类提供类别信息）[11-13]，在VOC2007数据集上mAP能达到73.2%，精度已经很高，但是识别速度还不够快。YOLO采用了更为

直接的思路：将整张图作为网络的输入，直接在输出层回归bounding box（识别框，记作bbox）的位置和bounding box所属的类别，从而实现把目标识别当作回归问题来解决。基于这种回归思想，本文将异常行为的检测也作为回归问题，将整张图作为输入，检测出标定有是否为异常行为的bbox的图像作为输出。

YOLO借鉴了Google-Net分类网络结构，有24个卷积层，2个全链接层，开始的卷积层用来提取图像特征，全连接层用来预测输出概率。作者还修改训练了YOLO的快速版本（Fast YOLO），Fast YOLO模型卷积层和卷积核相对较少，它只有9个卷积层和2个全连接层，最终输出为7×7×（5×2+20）的张量（因为有20类）[7]。这样的网络结构对于人体行为特征的提取和对是否异常的分类同样适用，对于实验中的两类标签，输出就是7×7×（5×2+2）的张量。

1.3 模型训练

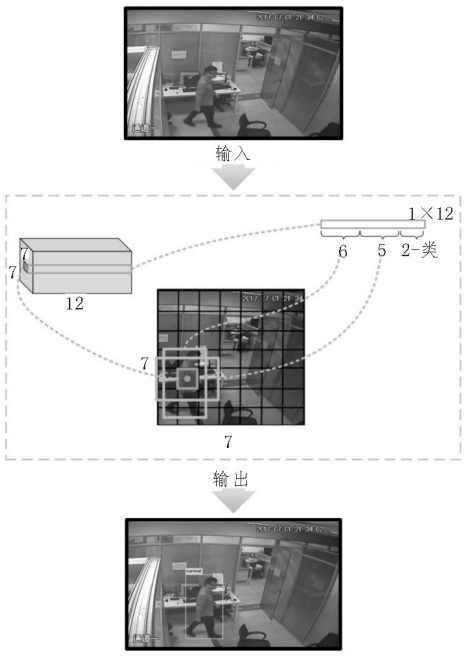

图2 实验训练过程的基本原理图

实验网络训练中的具体流程如下：

1）将处理过的图像进行resize处理，调整图像尺寸为448×448作为整个神经网络的输入。

2）通过运行神经网络得到一些bbox坐标、bbox中包含的人体目标（Object）的置信度和类别概率3种信息：

将输入的一幅图像分成S×S个网格，如图2虚框内将图像分成7×7个网格，当某Object的中心落在这个网格中，那么这个网格就负责预测这个Object。

每个网格都要预测一个类别信息记为C类，那么对于S×S个网格，每个网格既要预测B个bbox，同时还要预测C个类别，所以输出就是S×S×（5×B+C）的一个张量，对应图2中输出的张量就是7×7×（5×2+2）。

每个网格要预测B个bbox（图2中为2个），而每个bbox又要预测x，y，w，h和confidence共5个值。其中x，y是bbox中心位置的坐标，并且其值被归一化到[0，1]；w，h是bbox的宽度和高度，同样归一化到[0，1]；每个bbox除了要回归自身的位置之外，还要附带预测一个confidence值。这个confidence代表了所预测的bbox中含有Object的置信度和这个bbox预测的有多准两种信息，计算方式如下：

$$confidence=Pr(Object)\times IOU_{pred}^{truth}$$

其中如果有Object落在一个网格里，等式右边第一项取1，否则取0。第二项是预测的bbox和实际的标签框之间的IOU值[7]。

3）在测试的时候，将每个网格预测的类别信息和bbox预测的confidence信息相乘，就得到每个bbox的具体类别置信分数：

$$Pr(Class_i|Object)\times Pr(Object)\times IOU_{pred}^{truth}=Pr(Class_i)\times IOU_{pred}^{truth}$$

最后通过设置阈值，滤掉得分低的bbox，对保留的bbox进行非极大值抑制处理，就得到最终的检测结果。

图2为实验训练过程的基本原理图，区别于传统异常行为检测方法先进行目标识别然后进行分类，进而对异常行为进行检测，本文直接将两个步骤都交给YOLO网络去做，从而实现了输入一张图像直接输出一张经过网络预测有检测结果的图像，即端到端的检测过程。而通过GPU的加速，对于输入的视频流可以实时的显示异常行为的检测框。

1.4 实验条件

硬件条件：DELL PowerEdge T630服务器（32G内存）；两个CPU-E5-2620 v4；4个GTX1080Ti；

软件条件：Ubuntu14.04系统；CUDA 8.0；Python 2.7.6；OpenCV 2.4.13。

2 实验结果及分析

2.1 训练及测试结果

YOLO使用均方和误差作为loss函数来优化模型参数，即网络输出的S×S×（5×B+C）维向量与真实图像

的对应 S×S×(5×B+C) 维向量的均方和误差。实验中训练的网络的loss变化情况如图3所示，可以看到随着batches(训练批次)的增加，average loss(平均损失)在不断减小，逐渐趋于0，即整个网络结果趋于收敛。

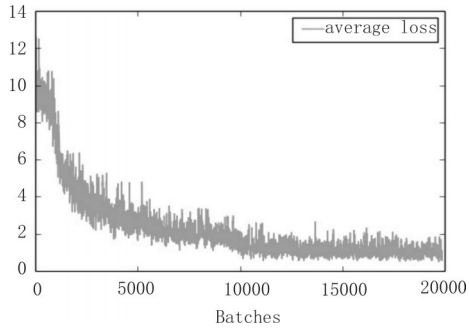

图3　训练网络的平均损失变化

文中在4个GTX1080Ti显卡加速的实验条件下，对网络进行训练，最终获得网络权重文件，进而对图像和视频数据进行测试。实验中采用了1 000张标定图像作为训练图像数据，146张作为测试数据，同时也对YOLO网络结构的简化版Fast YOLO和升级版YOLOv2[14]进行了训练和测试，部分图像测试对比结果如图4所示。图中3行图像分别对应Fast YOLO、YOLO和YOLOv2的检测结果。实验中红框的标签是normal，即正常行为；绿框的标签是abnormal，即异常行为。图4中第1列图中为红框，第2列、第3列为绿框，第4列中，左边为绿框，右边为红框，从第1、2列能明显的看出YOLO和YOLOv2网络模型框出的检测边界更准确。同时结果也充分说明了，主要用作目标检测的YOLO网络结构对于直接解决人体异常行为这种细分的分类问题同样适用。

图4　Fast YOLO、YOLO和YOLOv2检测结果对比图

2.2　评价与分析

文中采用了IOU[15-16]、召回率(Recall)和精确率(Precision)[17]3个评价指标来对实验结果进行评价。

在目标检测的评价体系中，有一个检测评价函数叫做intersection-over-union(IOU)，在实验中简单来讲就是模型产生的目标窗口和原来标记窗口的交叠率。即检测边框与真实边框(Ground Truth)的交集比上它们的并集，即为检测的准确率IOU：

$$IOU=\frac{检测边框\cap 实际边框}{检测边框\cup 实际边框}$$

召回率和精确度则采用如下方式计算：

$$精确度=\frac{TP}{TP+FP}$$

$$召回率=\frac{TP}{TP+FN}$$

TP —— True Positive (真正，TP)被模型预测为正的正样本，可以称作判断为真的正确率；

FP ——False Positive (假正，FP)被模型预测为正的负样本，可以称作误报率；

FN——False Negative(假负，FN)被模型预测为负的正样本，可以称作漏报率。

表1　3种网络结构检测结果对比

网络结构	IOU	召回率	精确率	秒/图
Fast YOLO	60.27%	83.60%	89.28%	0.003
YOLO	78.36%	99.88%	97.23%	0.009
YOLOv2	84.61%	100.00%	96.45%	0.010

表1中列出了Fast YOLO、YOLO和YOLOv2 3种网络结构的检测对比结果。可以看出，简化版本的Fast YOLO网络，各项指标都最低，因为网络结构相对简化，在牺牲了检测精度的情况下，检测速度相对于其他两种要稍快些。而从IOU参数来看，其他两种网络结构比简化版本有很大提高，所以相对应图4中的检测框也能更准确的框定检测目标。YOLO和YOLOv2两种网络模型召回率接近100%，精确率可

以达到96%以上。此外，在本文实验采用显卡GTX 1080Ti加速的条件下，对于视频流的检测速度可以达到30FPS左右。

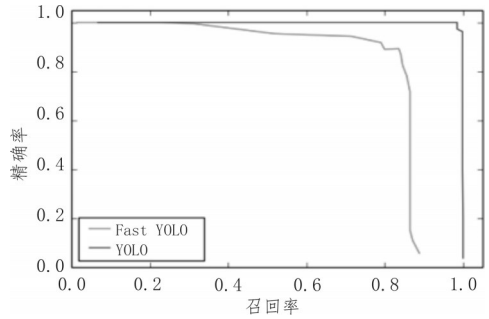

图5 Precision-Recall(P-R)曲线图

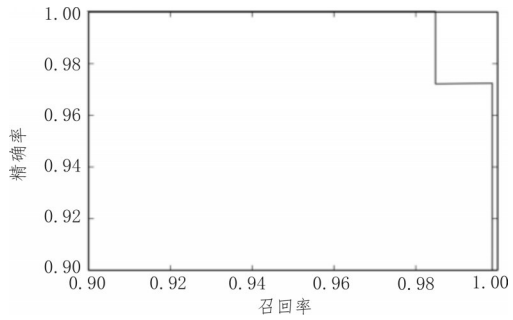

图6 YOLO模型P-R曲线细节图

通过调整阈值可以让网络检测出更多结果(检测框)，进而改变准确率或召回率的值，而评估一个分类器的性能，比较好的方法就是：观察当阈值变化时，Precision与Recall值的变化情况。如果一个分类器的性能比较好，那么它应该有如下的表现：让Recall值增长的同时保持Precision的值在一个很高的水平。从图5中可以看到，Fast YOLO网络模型的P-R曲线随着Recall值增长，Precision的值呈缓慢下降趋势，直到接近90%的召回率时下降为0；而YOLO网络模型当Recall值增长的时候，从图6可以看出，Precision的值一直处于96%以上；YOLOv2网络模型的P-R曲线跟YOLO网络模型的曲线很接近，没有在图中画出，精确率一直处于97%以上。因此，训练出的网络模型分类器可以很好地对异常行为进行检测。

3 结束语

本文通过对特定监控场景下对异常行为的定义需求标定数据，基于YOLO网络模型将标定的异常行为数据直接输入网络中进行训练，不将目标检测作为输出而直接得到是否异常的分类结果，从而让网络自动提取特征并分类，实现端到端的检测系统。最终实验结果表明这种方法可以很好地检测出监控视频中复杂的人体异常行为，并能够达到较高的检测精度。此外，这种方法可以迁移到不同监控场景，针对于特定场景、特定需求，可以达到很好地检测效果，对于满足不同行业的个性化需求方面具有重要意义。

参考文献：

[1] 桑海峰，郭昊，徐超. 基于运动特征的人体异常行为识别[J]. 中国科技论文，2014，9(7):812-816.

[2] 温向兵，满君丰，李倩倩，等. 视频监控中针对拥挤人群的人体分割与跟踪[J]. 小型微型计算机系统，2012(4): 891-895.

[3] 周宜波，何小海，张生军，等. 一种新的异常行为检测算法[J]. 计算机工程与应用，2012，48(3): 192-194.

[4] 罗超宇，李小曼，李浩. 基于光流能量的人体异常行为检测研究[J]. 计算机与网络，2014，40(21): 71-73.

[5] 胡晓燕. 视频监控迈入深度智能时代[EB/OL].(2017-05-08) [2017-12-15].http://security.asmag.com.cn/news/201705/91195.html.

[6] Redmon J. Darknet: Open source neural networks in c. Pjreddie. com.[EB/OL]. Available: https://pjreddie.com/darknet/. [Accessed: 21-Jun-2017], 2016.

[7] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 779-788.

[8] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2014: 580-587.

[9] Girshick R. Fast R-CNN[C]// IEEE International Conference on Computer Vision. IEEE, 2015: 1440-1448.

[10]Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]// International Conference on

(下转第164页)

步展开[15],包括:

1)可视化效果可扩展。考虑到与现有系统的结合使用,本文在说明气象时空数据可视化方法时,仅选取了等值线场和色彩场方法。对于时空数据可视化的呈现方式本身,也是值得研究的问题。

2)方法验证和改进。论文采用的实验数据类型有限,还需要采用不同类型不同规模的真实数据,对方法进行测试,更好地确定方法的适应性、有效性。

参考文献:

[1] 白雪莹.气象数据可视化表达研究与分析[J].科技传播,2017,9(12):58-59.

[2] 贾朋群,王淞秋.气象可视化——用可视化方法诠释[J].气象知识,2016(4):63-67.

[3] Mckinley S, Levine M. Cubic Spline Interpolation [M] Methods of Shape-Preserving Spline Approximation, 2011:37-59.

[4] Geng A C. The Application of Matlab in Teaching of Cubic Spline Interpolation Function[J]. Value Engineering, 2016(18):181-182.

[5] 朱立勋,魏萍.三次样条插值的收敛性[J].长春理工大学学报:自然科学版,2006,29(4):131-133.

[6] 李杰.地理观测数据时空可视化方法研究[D].天津:天津大学,2015.

[7] Sapiro G, Randall G. Morphing Active Contours [M]. IEEE Computer Society, 2000.

[8] 曾丽娜,周德云,潘潜,等.SAR图像最佳欧式空间距离矩阵匹配方法[J].系统工程与电子技术,2017,39(5):1002-1006.

[9] 张勇,王元珍,曹忠升.基于形态拟合的时间序列距离计算[J].华中科技大学学报(自然科学版),2012,40(8):72-76.

[10]陈高琳.图像缩放算法中常见插值方法比较[J].福建电脑,2017,33(9):98-99.

[11]许小勇,钟太勇.三次样条插值函数的构造与Matlab实现[J].兵工自动化,2006,25(11):76-78.

[12]张佳静,白红利,丁立平.一种基于线性插值的气象云图剖面投影方法,CN 103455715 B[P].2016.

[13]徐鹏飞.双线性插值和三次卷积在图像缩放中的应用及实现[J].网络安全技术与应用,2017(12):51-52.

[14]陈为.数据可视化的基本原理与方法[M].北京:科学出版社,2013.

[15]张澄钺.气象信息可视化技术现状及发展趋势[J].科研,2016(9):118.

(上接第158页)

Neural Information Processing Systems. MIT Press, 2015:91-99.

[11]Szegedy C, Toshev A, Erhan D. Deep neural networks for object detection[C]//Advances in neural information processing systems, 2013: 2553-2561.

[12]Girshick R, Iandola F, Darrell T, et al. Deformable part models are convolutional neural networks[C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition,2015: 437-446.

[13]Sande K E A V D, Uijlings J R R, Gevers T, et al. Segmentation as selective search for object recognition[C]// IEEE International Conference on Computer Vision. IEEE, 2012:1879-1886.

[14]Redmon J, Farhadi A. YOLO9000: Better, Faster, Stronger[R]2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, Hawaii, USA, 2017:6517-6525.

[15]Rahman M A, Wang Y. Optimizing Intersection-Over-Union in Deep Neural Networks for Image Segmentation[C]//International Symposium on Visual Computing. Springer, Cham, 2016: 234-244.

[16]Nowozin S. Optimal Decisions from Probabilistic Models: The Intersection- over- Union Case[C]// Computer Vision and Pattern Recognition. IEEE, 2014:548-555.

[17]Davis J, Goadrich M. The relationship between Precision- Recall and ROC curves[C]// International Conference on Machine Learning. ACM, 2006: 233-240.