文章编号:1007-1423(2019)27-0046-06 DOI:10.3969/j.issn.1007-1423.2019.27.009

利用姿态信息实现异常行为检测

郑爽,张轶

(四川大学计算机学院,成都 610065)

摘要:

异常行为检测广泛应用于安防、智能交通、机场监视、监考等领域,但异常行为数据难以获取,算法准确率较低。为了应对上述问题,提出一个基于对抗自编码思想的两路异常检测网络。其中,一路子网络利用像素信息,关注行为发生的整体环境。另一路子网络则利用姿态信息,关注人体行为。然后对两个子网络的结果进行混合,得到异常行为检测的结果。最后,在 CUHK Avenue 和 UCSD Ped 数据集上验证结果。

关键词:

异常行为检测; 对抗自编码器; 人体姿态

0 引言

异常行为检测是计算机视觉领域具有重要应用价值的一个领域。异常行为检测在公共安全防护,家庭老人防护中都有重要的意义。随着时间的发展,公共场所,如 ATM 机附近,交通要道的安全越来越受重视,监控随处可见,但要 24 小时监控,预防或阻止异常时间的发生需要投入极大的人力和物力[1]。若监控设备能够检测异常事件并报警,则能有效地阻止异常事件的发生。如在家庭中,如果独居老人摔倒或者晕倒,监控可以及时报警,也能使老人得到及时的救治;在智能交通中,如能及时发现撞人逃逸则能让交管部门采取主动行动。

在深度学习兴起以后,将其应用到异常行为检测领域中,使得这个领域有了很大的进步。但要做到准确、及时地检测到异常行为依然存在巨大的困难,其主要体现在:①不同场景下的异常行为类型不同,很难直接定义。②在某个场景下正常行为发生的次数多,异常行为发生的次数少,数据非常不平衡。本文采用了如文献[2]的异常检测方法直接使用正常数据进行训练,通过重构误差来检测异常行为、应对数据不平衡问题,并提出结合姿态特征来提高网络的准确率。

1 相关工作

1.1 异常行为检测

异常行为检测是行为识别的子领域[1],近年来在基于深度学习的行为识别领域,最突出的两种网络就是双流网络和 C3D 网络[3],其他的方法大多数是两种网络的改进。

双流网络利用了两个二维深度卷积网络,分别处理视频帧和帧间的光流。视频帧中包含了行为的空间信息,而光流包含了视频的运动信息。在分别提取两种信息特征以后进行混合,得以同时利用视频帧的空间和运动信息[4]。最初的双流网络缺点是能处理的行为时间比较短,不能提取长时间行为的特征,且准确率有待提高[3]。于是有了许多改进的网络架构:如文献[5-6]中提到的方法,通过在训练过程中混合特征来提高准确率。文献[7]提出分段提取行为特征并进行混合使网络可以处理更长时间段的行为。上述网络都依赖光流,但光流的提取也耗费时间。Tran D 等人[8]提出了一个三维网络,可向网络直接输入连续的帧。端到端的训练,使之快于双流网络,但是输入帧的数量是有限的,依旧限制了行为的时长。文献[9-10]中又提出通过扩展输入大小来提升网络。

上述网络都采用了监督学习,要进行异常行为检

测，只需要将需要检测的行为设置为异常类，就可以识别检测，但训练这样的网络通常需要大量的异常类数据，而真实情况下异常行为的数据非常稀少，且难以收集[11]。例如在银行 ATM 机附近的监控中，很难收集到抢劫、偷窃的视频，一是因为这些行为发生的次数较少，二是因为监控数据量非常大，要从大量的数据中找出这些行为非常困难[12]。因为有大量的正常数据，Akcay 等人[2]提出了仅使用正常数据做训练的异常检测方法。

本文受文献[2]的启发，用同样的方法来应对数据不平衡问题。但是这种方法更关注整个视频帧的分布，其中存在许多冗余信息。受文献[13-14]启发，本文利用了人体姿态信息，使网络更关注视频中人的行为，提高异常行为检测的准确率。为提取人体姿态信息，本文采用了基于深度学习的姿态估计网络。

1.2 姿态估计

姿态估计算法又分单人姿估计和多人姿态估计[15]，在异常行为检测的场景中一般会有多人存在，所以使用多人姿态估计。多人姿态估计分为两种，自顶向下的或者自底向上的[16]。自顶向下的方法是指先检测出视频帧中的人，然后检测每个人的关节点，估计人体的姿态[17]。自底向上的方法则是，向检测出视频帧中人体的关节点，然后聚类[16]。

Insafutdinov 等人[16]提出的就是自底向上的方法，先找人体出关节点，然后对关节点进行聚类。文献[18]中的方法也是自底向上的，并将姿态估计应用到了视频追踪中。Cao 等人[19]改进了自底向上的方法，使之速度更快。但自底向上的方法准确率并不高。文献[17]使用了自顶向下的方法，通过 YOLO、SSD 等网络先检测人体，然后对检测到对人体进行姿态估计，虽然速度有所下降，但是准确率得到了提高。为了得到较高准确率的姿态信息，本文采用文献[17]中的网络提取视频帧中行人的姿态信息。

2 实现

针对异常行为数据难以获取的问题，本文采用了基于生成对抗网络思想的半监督的异常检测算法[2-3]。生成对抗网络最初在文献[20]中提出，其主要目的是生成足够真实的图片。一般生成对抗网络包含两个子网络，一个生成器子网络，一个判别器自网络。生成器生成图片，判别器判断输入图片的真假，经过对抗训练，使生成器最终可以生成足够真实的图片。Makhzani 等人[21]提出了对抗自编码器与生成对抗网络有同样的训练思想。它有一个编码器和一个解码器，编码器从图片中提取特征向量，解码器通过特征向量重构图片。鉴别器的作用与生成对抗网络一致。

本文采用与文献[2]一致的网络结构，在对抗自编码器的基础上再添加一个编码器，对比真实视频帧和生成视频帧特征向量的差异。网络结构如图 1。

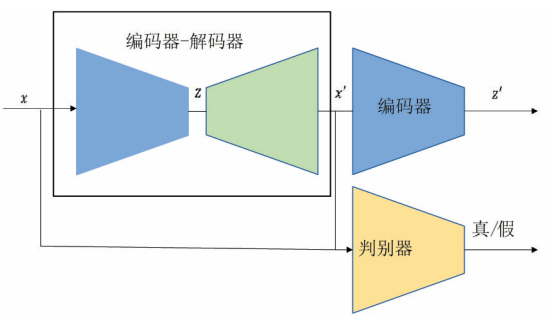

图 1 子网络结构

故实验可以只使用含有正常行为的数据对网络进行训练。向网络输入含有正常行为的视频帧，对抗自编码器提取特征向量并重构视频帧，判别器判断图片的真假，使重构的视频帧与输入视频帧一致，然后提取重构视频帧的特征向量与输入图片的特征向量对比。

在训练结束后，仅使用生成器加编码器部分进行推测。如图 1 所示，假设输入帧为 x，编码器提取特征向量 z：

$$z = f(x) \tag{1}$$

解码器通过特征向量 z 重构帧 x'：

$$x' = g(z) \tag{2}$$

编码器提取重构帧 x' 的特征向量 z'：

$$z' = f(x') \tag{3}$$

然后对比 z 和 z' 之间的差异。因为训练的时候仅使用正常数据，网络只学习到了正常数据的分布，所以可以很好地重构正常视频帧，不能很好地重构异常帧。如果输入帧是正常帧，z 和 z' 的差异就很小，如果是异常帧，则差异较大。

模型训练好后，进行测试，网络[2]在测试时其性能会随着 batch size 的减小而退化，因为文献[2]提出的网络使用了 Batch Norm 层。Batch Norm 是 Ioffe 等人[22]在

2015 年提出的，其使用具有重大的意义。Batch Norm 层对每层数据进行批量归一化，加速了网络的训练，使以前许多难以训练的网络可以进行训练。Batch Norm 通过估计数据的方差和均值对数据进行归一化，但是其方差和均值的估计依赖 batch size，所以测试时如果减小 batch size 就会使方差和均值的估计不准确导致性能的下降，所以我们修改文献[2]中网络的 Batch Norm 层，将其替换为 Group Norm 层。Group Norm[23]是在 2018 年提出的，这种归一化方式是基于分组思想的，将输入数据的通道分组，通过组内数据进行方差和均值的估计，这样数据的归一化就不再依赖 batch size，即使在测试时减小 batch size 也不会导致模型性能的下降。Group Norm 分组的多少对于模型的性能也有一定的影响，越少越接近 Batch Norm，所以我们使每层的分组数不定，但保证每组内有两个通道。

现在大多数异常检测网络都只使用了视频帧的像素信息，对于异常行为检测任务，像素信息中存在许多复杂而冗余的信息，而异常检测网络会关注整个帧的信息，也会学习整个帧的分布，而不仅仅是行为的分布，所以我们希望网络能更加关注视频帧中的人体。但因为人的行为和环境有关系，我们又不能完全放弃环境信息，所以希望能够利用环境信息的同时更加关注视频帧中的人类行为。受文献[13-14]的启发，本文采用了姿态信息，姿态信息包含了人体的关节点和关节之间的关系，它是结构化的信息，可以过滤掉冗余信息，更容易表达语义。人体的姿态的变化更能表现人的行为，如果出现异常，其特征更为显著，使得网络更加关注人体的行为。也使网络可以更好地学习行为的分布。文献[4]提出的算法同时利用了像素信息和光流信息，受此启发，本文也采用两个子网络来分别处理像素信息和姿态信息。

由于像素信息的分布和姿态信息的分布完全不同，所以本文没有选择在训练过程中对两个网络提取的特征进行混合，而是在最后对网络预测结果进行加权平均。因为在不同的环境下，异常行为不同，所以环境对结果的影响也不同，故针对不同的数据集应该有不同的权值。故提出如图 2 的双路网络。

对于像素信息处理子网络，采用上述改进的网络，使网络能够很好地学习整个帧的分布，并能生成足够真实的帧。对于处理姿态的子网络，本文提出在上述改进的网络前面加上姿态提取处理过程，本文采用 Alphapose[17]提取人体的姿态，Alphapose 是一个自顶向下的姿态估计网络，在估计人体姿态后使用非极大抑制来提高估计的准确率，本文采用姿态信息包含了 17 个关节点，分别是双眼、双耳、鼻子、肩、手肘、手腕、左右臀、膝盖、脚踝。这些关节点和其关系能够很好地表达一个人的姿态。在提取姿态信息后，对其进行二值化处理，进一步去掉冗余信息。然后 GAN 的输入为姿态图，使得上述改进的 GAN 可以很好地关注人的正常行为，并能够重构正常行为的姿态。

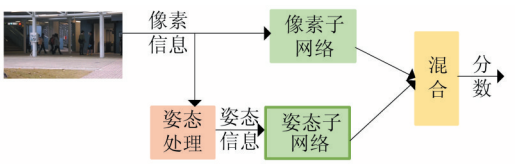

图 2 双路网络结构

最后对两种信息进行混合，混合方式如下。最后得到一个重构差异分数来判断帧中是否存在异常，分数计算如下：

$$S = w_1 score_r + w_2 score_p \tag{4}$$

$$S' = \frac{S - S_{\max}}{S_{\max} - S_{\min}} \tag{5}$$

S' 是最后的分数，S 是两路网络的混合结果。w_1 是像素子网络输出结果的权重，w_2 是姿态子网络输出结果的权重。$score_r$ 是像素子网络输出的分数，$score_p$ 是姿态子网络输出的分数，其计算公式如下：

$$score = mean((z' - z)^2) \tag{6}$$

3 实验

本文采用 PyTorch 框架实现。学习率为 0.0002，使用 Adam 优化，设置其动量参数为 0.5，输入帧大小为 256。所有的参数都参照了文献[2]。本文使用公开数据集 Avenue ped1 ped2 来验证模型效果。

Avenue 数据集：这是香港中文大学大道的监控视频，其中包含 16 个训练视频，21 个测试视频，一共 30652 帧，其中 15328 个训练帧，15324 个测试帧，训练帧中仅包含正常行走的视频，而测试帧中包含正常行走的正常帧和跑步，扔垃圾等异常行为的异常帧。

UCSD 数据集：整个数据集是一条人行道的监控，

又分为 ped1 和 ped2 两个部分。Ped1 共有 70 个视频，其中包含了 34 个训练视频，36 个测试视频，其中训练视频只包含正常行走的帧，而测试视频中包含行走和骑自行车等异常帧。Ped2 中共有 28 个视频，其中 16 个训练视频。12 个测试视频。训练视频如以上两个数据集一样只包含行走视频，而测试视频同时包含正常和异常视频帧。

常用于评价异常检测模型优劣的是受试者工作特征曲线(ROC)。曲线的横轴为假正例率，即测试为真，但实际为假的样本占负样本的比例；纵轴为真正例率，即，测试为正，且实际为正的样本占所有正样本的比例。每个点代表不同阈值下的真正率和假正率。图 3 是模型在 CUHK Avenue 数据集上测试的 ROC 曲线。

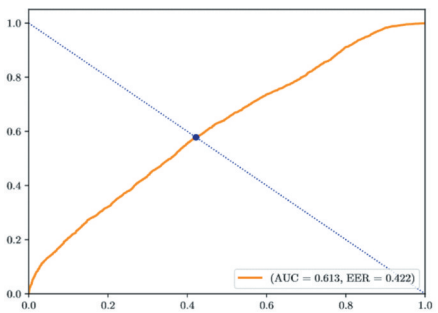

图3 在CUHK Avenue 数据集上测试的ROC曲线

AUC 则是 ROC 曲线下的面积，作为评价的标量，AUC 值大的模型优于 AUC 值小的模型。为了验证双路网络的有效性，本文对比了双路网络和其他算法在不同数据集上的 AUC 值，为评估不同归一化层带来的影响，本文对比了使用不同归一化层在不同数据集上的 AUC 值，结果如表 1 所示。

表 1 算法在不同数据集上的 AUC 值

数据集	CUHK Avenue	UCSD Ped1	UCSD Ped2
MPPCA[24]		0.59	0.69
Ganomaly[2]	0.56	0.63	0.66
Ganomaly+GN	0.55	0.63	0.68
本文	0.61	0.63	0.66

从表 1 可以看出其中 Avenue 数据集的效果优于其他两个数据集，通过分析可知，Avenue 数据集中的图像更清晰，人物并不密集，距离摄像头较近，能够提取到的姿态信息更明确，而 UCSD 数据集中行人比较密集，相对 Avenue 中的人更小，提取到的姿态信息显得比较模糊，所以导致 Avenue 的效果优于其他数据集。由此可知姿态信息对模型的性能有明显的影响。从表中结果可以看出，使用 Group Norm 层并不会太大地影响网络的性能，且使用 Group Norm 层后，模型性能也不会随着 batch size 的下降而降低。测试视频帧结果如图 4-图 5。

图4 提取到的视频序列中的姿态图

图5 子网络重构视频序列姿态图

图 4 是从视频帧中提取出来并进行二值化处理后的姿态信息，图 5 是网络中间生成的姿态信息。从图中可以看出，网络可以很好地学习姿态图的分布，并重构。

图 6 展示了包含正常行为的视频帧，而图 7 是子网络重构的视频帧可以看出两者之间的差别并不大。

图6 包含正常行为的视频帧序列

图7 子网络重构包含正常行为的视频帧序

图 8 包含异常行为的视频帧序列

图 8 为包含异常行为的视频帧序列，其中异常行为由红框标出。图 9 为子网络中间重构的视频帧，可以看出，网络并不能很好地重构包含异常行为的帧。其差异较大。通过对比其特征向量的差异，则能检测出是否存在异常行为。

图9 子网络重构包含异常行为的视频帧序列

4 结语

本文采用了半监督的方法来进行异常行为检测以应对数据不平衡问题，并因异常检测方法并不只关注人类行文而提出了一个双路网络来加强异常检测网络对人体行为的关注，以增加异常行为的准确率。从本文的实验可以看出，将姿态应用于其中是非常有效的，姿态越准确对网络性能的提升帮助越大。在后续的工作中将会使用更加准确的姿态信息，并针对群体异常行为和单人异常行的姿态特征做研究。

参考文献：

[1]杨锐，罗宾，郝叶林，常津津. 一种基于深度学习的异常行为识别方法[J]. 五邑大学学报(自然科学版)，2018:27-34.

[2]Akcay S，Atapour-Abarghouei A，P. Breckon T. GANomaly:Semi-Supervised Anomaly Detection via Adversarial Training[J]. ACCV，2018.

[3]Herath S，Harandi M，Porikli F. Going Deeper into Action Recognition: A Survey[J]. Image and Vision Computing，2017:4-21.

[4]Simonyan K，Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos[J]. Advances in Neural Information Processing Systems，2014:568-576.

[5]Feichtenhofer C，Pinz A，Wildes R. P. Spatiotemporal Residual Networks for Video Action Recognition[J]. CVPR，2016.

[6]Feichtenhofer C，Pinz A，Wildes R. P. Spatiotemporal Multiplier Networks for Video Action Recognition[J]. CVPR，2017.

[7]Limin W，Yuanjun X，Zhe W，Yu Q，Dahua L，Xiaoou T，et al. Temporal Segment Networks:Towards Good Practices for Deep Action Recognition[J]. European Conference on Computer Vision，2016:20-36.

[8]Tran D，Bourdev L，Fergus R，Torresani L，Paluri M. Learning Spatiotemporal Features with 3D Convolutional Networks[J]. ICCV，2015.

[9]Varol G，Laptev I，Schmid C. Long-term Temporal Convolutions for Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017:1-1.

[10]Xu H，Das A，Saenko K. R-C3D:Region Convolutional 3D Network for Temporal Activity Detection[J]. ICCV，2017.

[11]Wen L，Luo W，Lian D，Gao S. Future Frame Prediction for Anomaly Detection-A New Baseline[J]. IEEE Conference on Computer Vision and Pattern Recognition，2018.

[12]Kiran B，Thomas D，Ranjith P. An Overview of Deep Learning Based Methods for Unsupervised and Semi-Supervised Anomaly Detection in Videos[J]. Journal of Imaging，2018:36.

[13]王恬，李庆武，等. 利用姿势估计实现人体异常行为识别[J]. 仪器仪表学报，2016:2366-2372.

[14]Morais1 R，Le V，Tran T，Saha B，Mansour M，Venkatesh S. Learning Regularity in Skeleton Trajectories for Anomaly Detection in Videos[Z]. unpublished.

[15]Sun K，Xiao B，Liu D，Jingdong W. Deep High-Resolution Representation Learning for Human Pose Estimation[Z]. unpublished.

[16]Insafutdinov E，Pishchulin L，Andres B，Andriluka M，Schiele B. DeeperCut:A Deeper，Stronger，and Faster Multi-Person Pose Estimation Model[J]. ECCV，2016.

[17]Hao-Shu F，Shuqin X，Yu-Wing T，Cewu L. RMPE:Regional Multi-person Pose Estimation[J]. ICCV，2017.

[18]Insafutdinov E，Andriluka M，Pishchulin L，Tang S，Levinkov E，Andres B，Schiele B. ArtTrack:Articulated Multi-Person Tracking in the Wild[J]. CVPR，2017.

[19]Cao Z,Simon T,Wei S E,Sheikh Y. Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields[J]. CVPR,2016.
[20]Goodfellow I J,Pouget-Abadie J,Mirza M,Xu B,Warde-Farley D,Ozair S,et al. Generative Adversarial Nets[Z]. MIT Press,2014.
[21]Makhzani A,Shlens J,Jaitly N,Goodfellow I,Frey B. Adversarial Autoencoders[J],2015.
[22]Ioffe S,Szegedy C. Batch Normalization:Accelerating Deep Network Training by Reducing Internal Covariate Shift[C]. International Conference on Machine Learning,2015.
[23]YuXin W,Kaiming H. Group Normalization[J],2018.
[24]Kim J,Grauman K. Observe Locally,Infer Globally:a Space-Time MRF for Detecting Abnormal Activities with Incremental Updates[J]. CVPR,2009.

作者简介：
郑爽(1993-),女,四川绵阳人,在读硕士,研究方向为计算机视觉
通信作者:张轶(1981-),男,四川成都人,博士,副教授,研究方向为计算机视觉、机器学习,E-mail: yizhang@scu. edu. com
收稿日期:2019-08-16　　修稿日期:2019-08-30

Using Pose Information for Anomaly Detection

ZHENG Shuang, ZHANG Yi

(College of Computer Science, Sichuan University, Chengdu 610065)

Abstract:
Anomaly detection has been widely used in security, intelligent, invigilation, etc. But the abnormal data is difficult to obtain and the accuracies of the current algorithms are not high. To address these problems, proposes a new model, which consisting of two adversarial autoencoders-like (AAE-like) sub-networks. One sub-network focuses on the environment by processing appearance information. Another sub-network focuses on the human behavior by utilizing the human pose information in the frames. Then the results of the two sub-networks are combined to obtain the result. Finally, t evaluates the proposed model on CUHK Avenue and UCSD Ped datasets.
Keywords:
Anomaly Detection; Adversarial Autoencoder; Human Pose