本文引用格式：赵彦龙,钟震宇.基于注意力机制的异常行为识别方法[J].自动化与信息工程,2023,44(3):17-22.

ZHAO Yanlong, ZHONG Zhenyu. Abnormal behavior recognition method based on attention mechanism[J]. Automation & Information Engineering,2023,44(3):17-22.

基于注意力机制的异常行为识别方法

赵彦龙[1]　钟震宇[2]

（1.内蒙古军区数据信息室，内蒙古 呼和浩特 010051

2.广东省科学院智能制造研究所，广东 广州 5100700）

摘要：针对行为识别领域中异常行为相似度高、特征关联度强等因素，导致相似性异常行为难识别的问题，提出一种基于注意力机制的异常行为识别方法。首先，将基于解耦结构的预测模块嵌入到基于 3D 卷积的双流行为识别网络中，改善网络对视觉细粒度特征的表达能力；然后，分别构建空间注意力模块和时间注意力模块，提高模型对空间关键特征区域的提取和时间维度上长期依赖关系的捕捉；最后，经过测试，6 种异常行为的识别精度均达到 97.6%，验证了该方法的有效性。

关键词：异常行为；3D 卷积；解耦结构；空间注意力；时间注意力

中图分类号：TP391.4　　**文献标志码：**A　　**文章编号：**1674-2605(2023)03-0003-06

DOI：10.3969/j.issn.1674-2605.2023.03.003

Abnormal Behavior Recognition Method Based on Attention Mechanism

ZHAO Yanlong[1]　ZHONG Zhenyu[2]

(1.Data Information Office, Inner Mongolia Military Region, Hohhot 010051, China

2.Institute of Intelligent Manufacturing, Guangdong Academy of Sciences, Guangzhou 510070, China)

Abstract: An abnormal behavior recognition method based on attention mechanism is proposed to address the issue of difficulty in identifying similar abnormal behaviors due to factors such as high similarity and strong feature correlation in the field of behavior recognition. Firstly, the prediction module based on decoupling structure is embedded into the dual flow behavior recognition network based on 3D convolution to improve the network's ability to express visual fine-grained features; Then, construct a spatial attention module and a temporal attention module respectively to improve the model's ability to extract key spatial feature regions and capture long-term dependencies in the temporal dimension; Finally, after testing, the recognition accuracy of six abnormal behaviors reached 97.6%, verifying the effectiveness of this method.

Keywords: abnormal behavior; 3D convolution; decoupling structure; spatial attention; temporal attention

0　引言

随着我国智慧城市的加速建设，视频监控作为一种辅助管理手段已大规模覆盖到各领域，如健康监护、工业生产以及公共安全等[1-2]。在当前社会高速发展的背景下，如何消除安全隐患、防控安全事故、保障人民生命健康已成为民生关注的热点和重要课题[3]。人是社会活动的主体，其行为涉及生产生活的各个领域。通过判定视频中人们活动是否存在异常行为，并采取必要的措施进行干预，对提升远程管控能力和维护社会秩序稳定具有重要的现实意义。

近年来，硬件设备的快速迭代和算力的跨越式提升，为基于视频的人体异常行为自动识别提供了可行性。视频数据下的人体异常行为自动识别技术利用高性能计算设备和计算机视觉技术，对采集的视频画面进行逻辑推理与科学决策，对其中存在的异常行为快速地定位并识别，从而通过捕捉真实场景下行为的动

态变化，自动完成视频监控任务。

目前，国内外学者对视频数据下的人体异常行为自动识别技术开展了大量研究。游青山等[4]设计一套基于机器视觉的矿井作业人员行为检测及违章识别系统，用于矿井作业人员违章操作的自动识别。OUYANG 等[5]通过多任务学习构架将三维卷积神经网络与长短期记忆网络相结合，通过多个视频段的特征提取，更有效地在不同类别间共享不同视频段下的视觉特征。YAN 等[6]通过图卷积网络与时间卷积网络交替结合的方式，同时捕获骨骼序列中的时空特征，提高了骨骼序列下的行为识别率。林创鲁等[7]通过 YOLOv4 网络和 DeepSORT 算法，实现自动扶梯出口拥堵、长时间滞留等乘客异常行为的识别。

尽管学者们通过行为识别技术在视频监控领域取得了显著的进步，但当前技术仍存在的两个问题限制了其在实际场景中的应用：1） 相似动作误判，针对特定场景中的行为识别任务，不同动作类别间的差异性小，类间特征关联性强，经过网络的多层特征提取后，动作细节丢失，致使模型对相似动作做出误判或误报；2） 时空特征离散化，在时空场景中，行为在时间触发上具有随机性、在类别上具有不确定性，而模型在特征提取过程中将时域中每一帧、空域中每个像素都同等化处理，易引入干扰信息，同时缺乏对关键特征信息的关注，导致模型识别精度下降。

针对上述问题，本文提出一种基于注意力机制的异常行为识别方法。首先，构建一个基于 3D 卷积的双流行为识别网络；然后，为基于 3D 卷积的双流行为识别网络设计一种基于解耦结构的预测模块，提高网络对相似性动作的识别精度；接着，分别构建空间注意力模块和时间注意力模块，对输入特征进行建模分析，提升网络对时空重要特征的关注，从而提高异常行为的识别精度；最后，基于注意力机制的异常行为识别模型在数据集上训练并测试。

1　数据集建立

1.1　数据采集

本文采集人体常见的背痛、胸腹痛、颈痛（咳嗽）、跌倒、头痛、久坐等 6 种异常行为的视频数据。为了确保视频数据的多样性，增强模型在实际场景中的鲁棒性，在视频采集过程中，采用了多角度、多位置、多时间段和多视角策略进行异常行为的录制，每个动作视频均由位于不同位置的 2 个摄像头进行拍摄。动作视频录制分辨率为 1 920×1 080 像素，每个视频时长为 0~5 s，并以 avi 的格式进行保存。最终收集了 4 746 个视频，随机选择其中的 4 271 个视频作为训练集，剩余 475 个视频作为测试集，并利用文档分类的方式对视频进行标注。

1.2　数据增强

为扩充人体异常行为的数据量，使行为识别模型在真实场景中更具实用性与鲁棒性，对原始视频的训练集采用裁剪与放缩、椒盐噪声与水平翻转、旋转与模糊、旋转与颜色抖动、平移与亮度调整 5 种数据增强方法进行预处理。数据增强后的效果图如图 1 所示。

(a)　原图　(b)　裁剪与放缩

(c)　椒盐噪声与水平翻转　(d)　旋转与模糊

(e)　旋转与颜色抖动　(f)　平移与亮度调整

图 1　5 种数据增强方法的效果图

2　基于注意力机制的异常行为识别模型

针对以往异常行为识别方法对相似性异常行为识别效果差、忽视异常行为在时空维度上的特征关联

等问题，分别设计基于解耦结构的预测模块、空间注意力模块、时间注意力模块，并通过上述模块重构基于 3D 卷积的双流行为识别网络，提高网络对相似性异常行为的识别能力和跨域场景的鲁棒性，改进的基于 3D 卷积的双流行为识别网络结构图如图 2 所示。

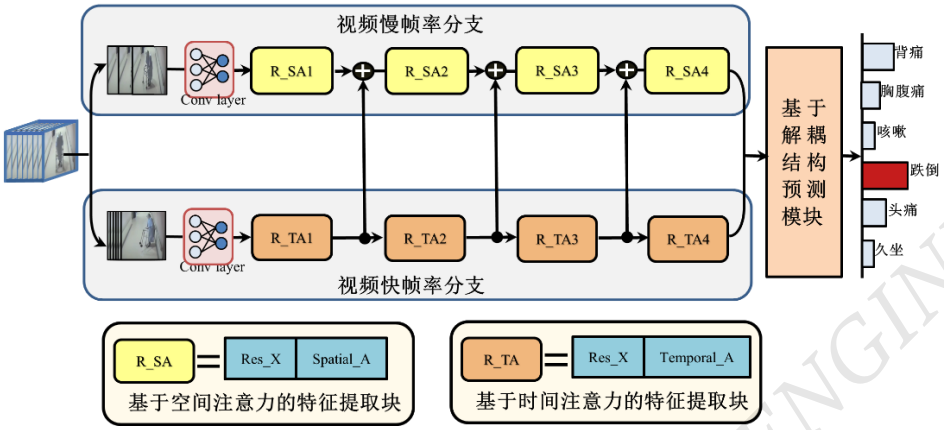

图 2　改进的基于 3D 卷积的双流行为识别网络结构图

改进的基于 3D 卷积的双流行为识别网络由视频慢帧率分支和视频快帧率分支构成，分别提取空间特征和时间特征。在空间维度上，为使网络对单帧图像关键区域赋予更大的权重，在慢帧率分支中，每个阶段的3D卷积模块Res后添加空间注意力模块。同理，在时间维度上，为提高网络对重要视频帧的关注，在快帧率分支中，每个阶段的 3D 卷积模块 Res 后添加时间注意力模块。

2.1　基于解耦结构的预测模块

考虑到在以往的异常行为识别模型中，网络通常先利用全连接层对特征数据进行维度整合，再推理异常行为的类别和发生时刻。这种将分类任务和回归任务混淆的处理方式，导致时序特征表达模糊，不利于模型对异常行为的精准判定。

解耦结构在目标检测与目标分割等领域已取得显著的效果[8-9]。通过解耦结构能够有针对性地对有用特征进行约束，使网络各分支更专注于学习自身有用的特征。本文将解耦结构引入异常行为识别模型中，通过解耦合的方式对全连接层进行结构优化，设计双分支结构使各子任务更加关注于自身任务的特征分布，在提取与子任务匹配的特征后，再进行总体特征的融合，从而实现行为识别效果最优化。基于解耦结构的预测模块结构如图 3 所示。

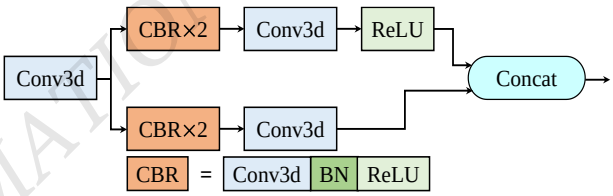

图 3　基于解耦结构的预测模块结构

在基于解耦结构的预测模块中，首先，将融合的双流特征通过 1×1 的卷积核进行降维，降低模型的参数与计算开销；然后，将特征分别融入分类分支与回归分支进行相应的特征提取；最后，将双分支结构的特征进行拼接，拼接后的结果通过全连接层确定异常行为的类别和发生时刻。

2.2　空间注意力模块

考虑到异常行为识别模型的输入是视频数据，而每一个视频帧中都存在杂乱的背景干扰信息，且不同视频帧的相似性动作的细粒度特征经过多层卷积后易被忽略。因此，通过引导模型关注人体区域的信息，有助于网络提取重要区域的特征并保留更多的细粒度特征。

本文将空间注意力模块嵌入慢帧率分支中，通过对视频帧空间特征重组的方式，使网络更加关注视频帧中人体区域，从而提高视频数据的人体异常行为识别率。空间注意力模块结构如图 4 所示。

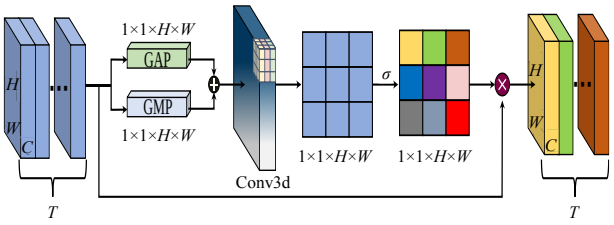

图 4　空间注意力模块模块结构

空间注意力模块的输入可视为多帧特征图的集合$U=[u_1,u_2,\ldots,u_t]$，其中 t 表示输入特征图的帧数，采用两个并行的池化分支（全局平均池化和全局最大池化）对输入模块的数据进行降维，同时获得维度相同的输出数据 $1\times1\times H\times W$。通过池化操作可有效提高网络的表达能力，同时滤除特征图中无用的信息。空间注意力模块处理过程如下：

1） 池化操作

$$\mathrm{GAP}_{3D}(U)=\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}U(c,t) \tag{1}$$

$$\mathrm{GMP}_{3D}(U)=\mathrm{Max}(U(c,t)) \tag{2}$$

式中：c 表示单个视频帧 u_i 的通道个数，$c\in(1,2,\ldots,C)$，Max表示从多帧特征图集合$U(c,t)$中取最大值，$\mathrm{GAP}_{3D}(U)$和$\mathrm{GMP}_{3D}(U)$分别表示经过全局平均池化和全局最大池化的输出特征图；

2） 经过全局平均池化和全局最大池化 2 个分支处理后，得到 2 个维度相同的输出数据$1\times1\times H\times W$，将 2 个输出数据相加，得到空间特征描述器 S：

$$S=\mathrm{GAP}_{3D}(U)+\mathrm{GMP}_{3D}(U) \tag{3}$$

3） 采用一个 7×7 的卷积核对空间特征描述器 S 进行特征筛选，一方面对相加后特征图中的冗余信息进行滤除，另一方面增强空间特征描述器 S 的表达能力，确定特征图中有效的关键区域；

4） 利用 Sigmoid 激活函数获得空间注意力的权重参数 M_S，三维卷积激活过程公式为

$$M_S=\sigma(S\times W_{7\times7}+b) \tag{4}$$

式中：σ表示 Sigmoid 激活函数，$W_{7\times7}$表示尺寸为 7×7 的三维卷积核，b表示卷积层的偏置量。

2.3　时间注意力模块

在 RGB 视频模式下的异常行为识别任务中，异常行为往往发生在长期视频序列的特定时间段，仅与数百个视频帧具有强相关性，而与其他时间段的视频帧弱相关甚至无关。若仅依赖三维卷积的方式将视频数据的不同帧进行整合与时序特征的盲目提取，易引发无关信息对模型的干扰，同时产生异常行为检测的滞后。因此，在三维卷积网络中引入时间注意力模块，有助于提升模型对视频中特定且信息丰富帧的关注，从而降低无关帧对模型的干扰。考虑到前文构建的异常行为识别模型中的快帧率分支用于捕捉时序序列的相关性，因此将时间注意力模块嵌入到快帧率分支中，提升模型对视频帧中有效片段的关注，时间注意力模块结构如图 5 所示。

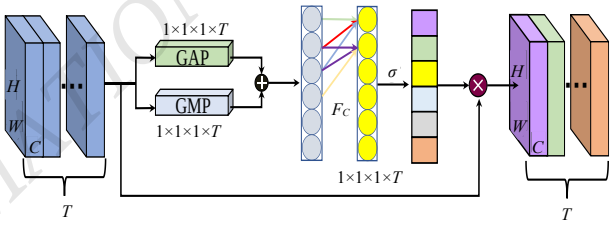

图 5　时间注意力模块结构

时间注意力模块的输入可视为多帧特征图的集合$U'=[u_1',u_2',\ldots,u_t']$，采用 2 个并行的池化分支（全局平均池化和全局最大池化）对输入模块的数据进行降维，同时获得维度相同的输出数据$1\times1\times1\times T$。时间注意力模块处理过程如下：

1） 池化操作

$$\mathrm{GAP}_{3D}(U')=\frac{1}{CHW}\sum_{c=1}^{C}\sum_{i=1}^{H}\sum_{j=1}^{W}U'(c,i,j) \tag{5}$$

$$\mathrm{GMP}_{3D}(U')=\mathrm{Max}(U'(c,i,j)) \tag{6}$$

式中：H和W分别表示输入特征图的高度与宽度；C表示单个视频帧的通道个数；

2） 经过池化处理后得到 2 个维度相同的输出数据$1\times1\times1\times T$，将 2 个输出数据相加，得到时间特征描述器T_d：

$$T_d=\mathrm{GAP}_{3D}(U')+\mathrm{GMP}_{3D}(U') \tag{7}$$

3） 采用全连接层对时间特征描述器T_d进行特征筛选，对发生异常行为的视频帧赋予更高的评分，增强网络对关键视频帧的关注，对未出现异常行为的视频帧赋予较少的关注，避免无用帧对模型推理的影响；

4） 利用 Sigmoid 激活函数获得时间注意力的权重参数M_T，三维卷积激活过程公式为

$$M_T=\sigma(F_c(T_d)) \tag{8}$$

式中：F_c表示时间注意力模块中的全连接层，σ表示 Sigmoid 激活函数。

3 实验

3.1 模型训练

实验基于 Ubuntu16.04 操作系统，采用 Python 语言编程和 PyTorch 深度学习框架完成。利用迁移训练的方式通过大型行为识别数据集 UCF101[10]进行预训练，并使用随机梯度下降算法[11]进行优化。实验平台硬件配置：英特尔 i7-11800H 处理器、Nvidia GeForce GTX 1080Ti。本文提出的基于注意力机制的异常行为识别模型的超参数设置如表 1 所示。

表 1 基于注意力机制的异常行为识别模型的超参数设置

参数名称	参数值
迭代次数	120
批处理个数	32
学习率	0.001
输入图像尺寸	(224, 224)
帧率间隔	2
输入帧数量	32

3.2 评价指标

为衡量本文提出的模型性能及分析模型对每个异常行为类别的检测效果，采用评价指标精确度（accuracy, Ac）对识别实验结果进行综合性评估。此外，考虑到人体存在多种行为同时发生的可能性，还需要检索概率值最高的3个预测结果中是否有真实的标签。Top-1 表示最大概率值的预测结果为正样本的准确性，Top-3 表示在模型输出的前 3 个最大概率中存在正样本的准确性，评价指标精确度$\mathrm{Ac}_{\mathrm{Top}\text{-}x}$计算公式为

$$\mathrm{Ac}_{\mathrm{Top}\text{-}x}=\frac{TP}{TP+FP} \tag{9}$$

式中：*TP*表示模型能正确识别出异常行为的数量，*FP*表示模型错误的预测结果数量。

3.3 实验结果

为了验证本文方法的优越性，采用先进的行为识别方法 I3D[12]、TSM[13]、Slowfast[14]、TANet[15]、TPN[16]与本文方法进行性能比较。针对测试集中 475 个测试视频，不同行为识别方法的测试结果对比如表 2 所示。

表 2 不同行为识别方法的测试结果对比

模型	模型大小	*TP*	*FP*	Top-1	Top-3
I3D	208.1 MB	400	20	0.952	0.991
TSM	179.8 MB	401	19	0.955	0.997
Slowfast	469.7 MB	404	16	0.961	1.00
TANet	189.5 MB	405	15	0.965	1.00
TPN	689.2 MB	408	12	0.971	1.00
本文方法	483.8 MB	410	10	0.976	1.00

由表 2 可知：本文方法在模型大小及性能上均达到最优；在精确度方面，相比于效果最好的 TPN 行为识别网络精度提升 0.5%，达到 97.6%；且本文提出方法具有较低的误检率，可满足现实场景中异常行为的检测需求。

4 结论

本文提出一种基于注意力机制的异常行为识别方法，通过采用多种数据增强方法对采集的数据进行数据增强，增加数据的多样性和模型在复杂场景下的鲁棒能力；在基于 3D 卷积的双流行为识别网络中嵌入基于解耦结构的预测模块，改善网络对于视觉细粒度特征的表达能力，提高相似性行为的识别精度；使用空间注意力模块和时间注意力模块，提高模型对空间中关键特征区域的重视和时间维度上长期依赖关系的捕捉。经测试，异常行为的识别精确度达到 97.6%，验证了本文方法的有效性及实用性。

参考文献

[1] 胡艳君,温强,朱晓妹,等.智慧城市背景下产业智慧化管理系统的构建与应用[J].智能建筑与智慧城市,2022(2):152-155.

[2] 何炜,周保林,王皓.视频监控技术在智慧城市中的应用[J].电子技术,2022,51(1):40-41.

[3] 李雪峰.提高公共安全治理水平的战略意涵与实现路径[J].中国应急管理科学,2022(11):13-26.

[4] 游青山,冉霞.基于机器视觉的矿井作业人员行为监测及违章识别系统[J].自动化与信息工程,2021,42(4):20-24.

[5] OUYANG X, XU S, ZHANG C, et al. A 3D-CNN and LSTM based multi-task learning architecture for action recognition[J]. IEEE Access, 2019,7:40757-40770.

[6] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition[C]// Proceedings of the AAAI conference on artificial intelligence, 2018.

[7] 林创鲁,叶亮,李刚,等.基于深度学习的自动扶梯乘客异常行为识别方法研究[J].自动化与信息工程,2022,43(6):1-6.

[8] GE Z, LIU S, WANG F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv: 2107.08430, 2021.

[9] ZHANG H, WANG M, LIU Y, et al. FDN: Feature decoupling network for head pose estimation[C]//Proceedings of the AAAI conference on artificial intelligence, 2020,34(7):12789-12796.

[10] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild[J]. arXiv preprint arXiv:1212.0402, 2012.

[11] BOTTOU L. Stochastic gradient descent tricks[J]. Neural Networks: Tricks of the Trade: Second Edition, 2012:421-436.

[12] CARREIRA J, ZISSERMAN A. Quo Vadis, action recognition? a new model and the kinetics dataset[C]// proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017:6299-6308.

[13] LIN J, GAN C, HAN S. TSM: temporal shift module for efficient video understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7083-7093.

[14] FEICHTENHOFER C, FAN H, MALIK J, et al. Slowfast networks for video recognition[C]//Proceedings of the IEEE/CVF international Conference on Computer Vision, 2019: 6202-6211.

[15] LIU Z, WANG L, WU W, et al. Tam: temporal adaptive module for video recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 13708-13718.

[16] YANG C, XU Y, SHI J, et al. Temporal pyramid network for action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 591-600.

作者简介：

赵彦龙（通信作者），男，1984 年生，硕士，工程师，主要研究方向：计算机网络、人工智能、大数据。E-mail: 757477184@qq.com

钟震宇，男，1971 年生，博士，研究员，主要研究方向：深度学习、人工智能、大数据。E-mail: zy.zhong@giim.ac.cn

（上接第 16 页）

纪艺杭，男，2001 年生，本科，主要研究方向：图形图像处理。E-mail: atticusji@163.com

韩耀荣，男，2002 年生，本科，主要研究方向：三维重构。E-mail: han_yaorong@163.com

李震，男，1981 年生，博士，教授，主要研究方向：基于计算机图像技术和无线多媒体物联网的果园病虫害检测方法研究与装备设计。E-mail: lizhen@scau.edu.cn

吕石磊，男，1984 年生，博士，教授，主要研究方向：山地果园运送、植保装备设计及智能信息化。E-mail: lvshilei@scau.edu.cn

宋淑然，女，1965 年生，博士，教授，主要研究方向：检测及测控技术在农业中的应用。E-mail: songshuran@scau.edu.cn

薛秀云，女，1980 年生，博士，高级实验师，主要研究方向：智能检测与控制、图形图像处理和三维重构。E-mail: xuexiuyun@scau.edu.cn