基于改进 C3D 的视频监控异常行为检测算法

郑凯东[1] 江 怡[1]
ZHENG Kaidong JIANG Yi

摘 要 随着科技的发展，视频监控技术已经在各种场景得到广泛应用，如城市安防、交通管理、工业监控等。然而，传统的视频监控系统通常依靠人工监控来发现异常行为，但这种方式效率低下且容易遗漏，因此需要借助计算机视觉和深度学习技术实现自动化的异常行为检测。针对视频监控下异常行为检测的问题，提出了一种异常检测算法 SE-C3D。首先，将传统的二维卷积和池化操作扩展到了三维；接着，利用 C3D 网络来提取视频的时空特征；然后，采用残差思想，设计了一种 3D 残差模块，增强泛化能力，使其在处理视频数据时更为有效；最后，为了进一步提高准确率，将 SENet 扩展到三维，并嵌入到残差 C3D 模块上，使用 Softmax 输出结果。实验结果表明，SE-C3D 相较于其他模型在多个性能指标上均有显著提升，提出的算法在异常行为检测任务中有着广泛的应用前景。

关键词 深度学习；异常行为检测；C3D；SENet；3D 残差结构

doi：10.3969/j.issn.1672-9528.2024.06.028

0 引言

异常行为检测是一项具有挑战性的任务，通常涉及识别视频中的异常行为或事件，这些行为与正常行为模式不同或具有潜在的威胁性。人体动作识别是异常行为检测中的一个重要方面，因为大多数的异常行为都涉及人的动作或行为。监控系统 [1]、智能场景建模 [2] 以及视频注释和检索 [3] 都是视频中行为识别的一些实际应用。视频异常行为检测主要利用监控系统，通过对视频数据进行实时的分析和处理，及时分析和识别可能存在的异常或异常行为。这些分析主要包括目标检测与跟踪、特征提取、模型训练、异常检测与判断四个步骤。当前异常行为检测主要有两种方法，即基于深度学习的方法和基于传统机器学习的方法 [4]。

基于传统机器学习的方法主要依赖于手工特征和分类器。其特征提取包括基于局部特征和全局特征。基于局部特征的方法是使用 3D Harris[5] 角点检测器或 DoG[6] 等不同技术从视频时空兴趣点（STIP）[7] 中提取局部时空描述符，比如方向梯度直方图（HOG）[8] 和光流直方图（HOF）[9]。在拥挤的视觉场景中，传统的基于局部特征的方法可能会由于大量移动对象导致产生许多无用的时空兴趣点，这会降低方法的准确性。因此，在拥挤的情况下，这些描述符可能不足以有效地捕捉场景的关键信息。而基于全局特征的方法在拥挤场景中具有一定优势，因为它们能够综合考虑整个场景的信息，而不是局限于局部区域。该方法是从一系列帧中提取特征，并通过光流幅度大小的变化和光流的方向信息来识别异常行为，比如 VIF 描述符 [10] 和 DiMOLIF[11] 描述符。目前传统机器学习方法在异常行为检测中仍存在一系列问题，包括特征提取的复杂性、准确率低、对光照和噪声的敏感性等。这些问题限制了传统方法在处理异常行为检测任务时的效果和适用性。

基于深度学习的方法能够利用神经网络作为特征提取器，构建端到端模型，能够自动提取特征并有强大的拟合能力，通过模拟人类大脑对数据的分析和学习，能对异常行为进行非线性描述，同时深度学习模型结构灵活，可以根据任务需求进行设计和调整，并且可以通过增加更多的层次或神经元来提高模型的表达能力和性能。

在基于深度学习的异常行为检测算法中，三维卷积神经网络具有较高的速度和准确率。C3D（convolutional 3D）是一种专门设计用于处理视频数据的卷积神经网络，能够有效地捕获视频数据中的时空信息。本文基于 C3D 网络模型，研究视频中的打架、盗窃、抢劫、纵火等异常行为，并且为了进一步增加检测的准确度，本文融合了 SENet（squeeze-and-excitation networks）[12] 与残差结构，提出了一个新型的 SE-C3D 模型。

1 基于改进 C3D 的视频监控异常行为检测算法

为了解决传统方法在视频监控异常行为检测中准确率低、复杂度大的问题，本文提出了基于 SENet 改进的 C3D

1. 西安石油大学 陕西西安 710065

视频监控异常行为检测算法，相对于原始的 C3D 模型，在精确度和评价指标上有显著提升，命名为 SE-C3D。图 1 为 SE-C3D 的模型结构，其中图 1 的红色虚线框为本文所提出的 3D 残差模块。

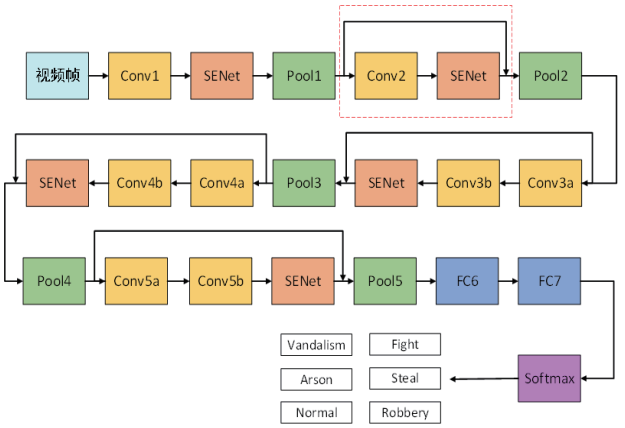

图 1 SE-C3D 模型结构图

算法步骤如下。

（1）输入数据：算法首先接收原始视频数据作为输入。

（2）C3D 特征提取：利用 C3D 的三维卷积操作，从视频数据中提取时空特征，不需要进行预处理。

（3）引入残差结构：为了增强网络的泛化能力，引入了残差结构。在神经网络中，残差结构可以帮助网络更容易地学习到恒等映射，从而降低训练难度，并且有助于防止过拟合。

（4）构造 3D 残差模块：通过构造 3D 残差模块，进一步增强了网络的泛化能力。这个模块通常由多个 3D 卷积层和激活函数组成，可以在保留主要特征的情况下减少信息的丢失。

（5）嵌入三维 SENet：为了提高准确率，将三维 SENet 嵌入到残差 C3D 网络中。SENet 能够学习每个特征通道的重要性，并动态地调整这些通道的权重，从而增强了网络对于有用特征的提取能力。

（6）Softmax 输出预测结果：使用 Softmax 函数对网络输出进行归一化处理，得到视频异常行为的预测结果。

1.1 3D 残差模块

本文的创新点之一是在 C3D 上引入残差思想，构建 3D 残差模块。该模块通过 shortcut connection 方式实现了 x 与 $H(x)$ 的叠加，提高了模型的性能。表达公式为：

$$H(x)=F(x)+x \tag{1}$$

式中：$H(x)$ 表示残差模块的输出，$F(x)$ 表示残差学习的函数，x 表示模块的输入。通过将输入和学习的残差相加，模块可以有效地学习输入数据的表示，并且可以通过 shortcut connection 传播梯度，减轻了训练深度网络时的梯度消失问题。这种设计能够提高网络的性能和泛化能力，有助于更好地识别异常行为。3D 残差模块如图 2 所示。

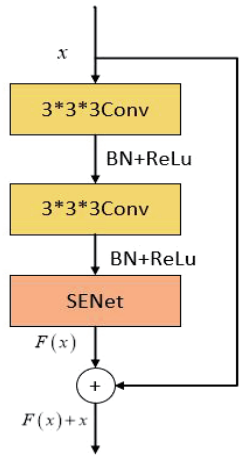

图 2 3D 残差结构

1.2 SENet

SENet 的引入是本文的另一个创新点，它被扩展到三维并嵌入到 3D 残差模块中。它主要由 Squeeze 和 Excitation 两个部分组成。原理如图 3 所示。

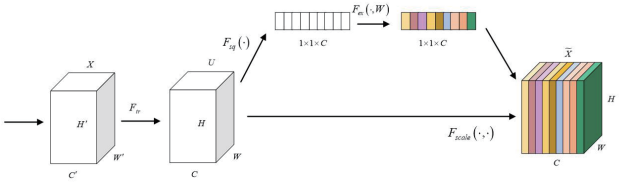

图 3 SENet 模型原理图

SENet 通过对卷积特征通道之间的关系进行建模，提高网络的表示能力。其关键在于根据各个特征通道的重要性程度对权重进行分配，从而使网络能够更有效地学习并利用不同特征通道之间的信息。这种机制有助于增强网络的表征能力，提高网络在各种视觉任务中的性能。而这种过程类似于 Attention 机制，在卷积特征通道上动态地调整重要性。SENet 的第一步是利用卷积操作提取图像特征图，其公式为：

$$F_{tr}: X \to U, X \in \mathbf{R}^{H' \times W' \times D' \times C'}, U \in \mathbf{R}^{H \times W \times D \times C} \tag{2}$$

$$F_{tr}: u_c = v_c * x = \sum_{s=1}^{c'} v_c^s * x^s \tag{3}$$

F_{tr} 这一步是转换操作，把输入的原始图像 X 映射到特征图 U 上。其中，H 和 W 表示特征图的高度和宽度，C 和 D 表示通道数和时间深度。v_c^s 表示第 c 个卷积核的参数，x^s 表示第 s 个输入，$V=[v_1,v_2,\ldots,v_c]$ 为学习到的卷积核集合，$U=[u_1,u_2,\ldots,u_c]$ 是输出。

在 Squeeze 阶段，输入特征 U 的大小为 $H\times W\times C$，通过全局平均池化操作，将每个空间位置的特征汇总成一个值，即将每个 $H\times W$ 的区域压缩为一个值，得到 $1\times 1\times C$ 的向量，其公式为：

$$z_c = F_{sq}(u_c) = \frac{1}{W \times H \times D} \sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{r=1}^{D} u_c(i,j,z) \quad (4)$$

通过 Squeeze 操作获得了特征通道的全局特征，接下来进行 Excitation 操作。Excitation 操作公式为：

$$s = F_{ex}(z,w) = \sigma(\vartheta(z,w)) = \sigma(w_2 \delta(w_1 z)) \quad (5)$$

式中：δ 为 ReLU 函数，σ 为 Sigmoid 函数，w_1^* 和 w_2^* 表示全连接（FC）操作。Excitation 操作首先通过全连接操作降维，其次在降维后应用 ReLU 函数进行非线性激活，然后通过全连接操作升维，最后利用 Sigmoid 函数进行权重归一化。

在重标定阶段，为每个特征通道分配相应的权重。$\hat{x}_c$为输出的特征映射，u_c 为输入原始特征映射、s_c 为权重向量。其公式为：

$$\hat{x}_c = F_{scale}(u_c, s_c) = s_c \cdot u_c \quad (6)$$

1.3 C3D 网络

C3D（Convolutional 3D）网络是一种用于视频分析的深度学习模型。它是在 2D 卷积神经网络（CNN）的基础上扩展，以便处理视频数据的时间维度。与传统的 2D CNN 不同，C3D 网络可以直接从视频数据中学习时空特征，而无需额外的预处理步骤。

C3D 网络通常由卷积层、池化层和全连接层组成，其中卷积层在时间和空间上同时操作，以捕获视频中的时空特征。通过在视频序列中移动 3D 卷积核，C3D 网络可以有效地提取视频中的运动信息和空间结构。此外，对于第一层的池化层，采用了较小的池化核和步长，以减少时间信息的丢失。在训练过程中，C3D 网络通过反向传播算法优化其权重参数，以最小化预测值与实际标签之间的误差。

C3D 网络在视频分类、动作识别、行为检测等任务中取得了显著的成果，成为视频分析领域的重要模型之一。C3D 网络架构如图 4 所示。

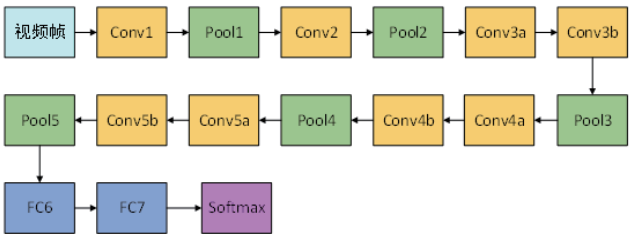

图 4 C3D 网络结构

2 实验结果分析

2.1 所用数据集

本文实验所用数据集为 UCF-Crime 数据集 [13] 和自制异常行为数据集，UCF-Crime 数据集中包含了 1900 个监控视频，视频的尺寸大部分为 240×320，涵盖了真实场景中的打架斗殴、纵火、抢劫等异常行为。数据集如图 5 所示。

图 5 UCF-Crime 与自制异常行为数据集展示

2.2 实验过程

首先，在 UCF-Crime 数据集上进行训练，以获得稳定且准确的异常行为检测模型。接着，将这个经过训练的模型应用到自制的异常行为数据集上进行验证，并将验证结果与模型在 UCF-Crime 数据集上的性能进行对比分析。

本实验在数据集划分时，采用 5 折交叉验证方法，将数据集划分为训练集和测试集，训练集包含 800 个视频，测试集包含 200 个视频。训练集和测试集分别包含了相同数量的异常行为和正常行为视频片段。在算法上，采用了小批量随机梯度下降法（SGD）进行模型优化，设置了权重衰减为 0.001，动量为 0.5。在学习率调整策略上，初始学习率设置为 0.001，在训练周期的 1/4 时再进行调整，分别调整为 0.000 1 和 0.000 01。使用交叉熵函数作为损失函数。

2.3 评价标准

在本次实验中使用精确度（Precision）、召回率（Recall）和准确率（Accuracy）的平均值来评估 SE-C3D 网络模型在视频异常行为识别的性能。其定义式为公式（7）～（9）。

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (7)$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (8)$$

$$\text{Accurary} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Negative} + \text{False Positive}} \quad (9)$$

2.4 实验结果和分析

2.4.1 对比分析

为了防止实验的结果出现偶然性，本实验将原始模型与 SE-C3D 各自在相同数据集上训练 100 轮，并进行比对。表 1 是展示了 C3D 网络和 SE-C3D 网络在 UCF-Crime 数据集上进行 100 轮实验的对比分析，从表 1 中可以看出 SE-C3D 模型要优于 C3D 模型。

表 1 UCF-Crime 数据集上模型的评估指标

模型	Avg.Accurary	Avg.Precision	Avg.Recall
C3D	93.50%	93.30%	95.23%
SE-C3D	98.20%	97.25%	99.15%

为了进一步验证本文提出的算法的有效性，将其与其他现有算法对比，如表 2 所示。其中，Radon Transform、STIFV 和 ViF+OViF 是传统机器学习的方法，其余为深度学习的方法。从表 2 可看出，本文提出的算法准确率明显高于其他算法。

表 2 不同算法在两个数据集的分类精度

方法	UCF-Crime	自制异常行为数据集
Radon Transform	90.10%	98.70%
STIFV	93.45%	99.10%
ViF+OViF	87.45%	—
ConvLSTM	97.10%	99.99%
3D CNN	91%	—
3D CNN+SVM	92.10%	98.95%
C3D	93.50%	—
C3D+KNN	92.30%	97.25%
SE-C3D	98.20%	100%

实验结果如图 6 所示。

图 6 SE-C3D 网络识别结果

3 结论

针对传统机器学习方法在异常行为检测上复杂度高、准确率低的问题，提出了一种新型的异常行为检测算法，名为 SE-C3D。该算法融合了 SENet 与 C3D，并引入了残差思想，以提高检测的性能。通过这些改进和融合，SE-C3D 在行为检测方面比原始模型有明显提升。实验结果表明，在 UCF-Crime 数据集上平均精确度上升了 3.95%，平均召回率上升了 3.92%，平均准确率上升了 4.70%，验证了本文算法的有效性。

参考文献：

[1]BEN M A, ZAGROUBA E.Abnormal behavior recognition for intelligent video surveillance systems:a review[J].Expert systems with application,2018,91(1):480-491.

[2]ULLAH H, ISLAM I U, ULLAH M, et al.Multi-feature-based crowd video modeling for visual event detection[J].Multimedia systems,2021,27(4):589-597.

[3]ROSSETTO L, GASSER R, LOKOC J, et al.Interactive video retrieval in the age of deep learning-detailed evaluation of VBS 2019[J].IEEE transactions on multimedia,2021,23:243-256.

[4] 朱煜 , 赵江坤 , 王逸宁 , 等 . 基于深度学习的人体行为识别算法综述 [J]. 自动化学报 ,2016,42(6):848-857.

[5]CHEN D, WACTLAR H, CHEN M, et al.Recognition of aggressive human behavior using binary local motion descriptors[C]//embc'08,[v.14].Piscataway:IEEE,2008:5238-5241.

[6]LOWE D G.Distinctive image features from scale-invariant keypoints[J].International journal of computer vision,2004,60(2):91-110.

[7]NIEVAS E B, SUAREZ O D, GARCÍA G B, et al.Violence detection in video using computer vision techniques[C]// Computer analysis of images and patterns,Part II.Heidelberg: Springer, 2011:332-339.

[8]DALAL N, TRIGGS B.Histograms of oriented gradients for human detection[C]//CVPR 2005,V.1.Los Alamitos:IEEE Computer Society,2005:886-893.

[9]DALAL N, TRIGGS B, SCHMID C.Human detection using oriented histograms of flow and appearance[C]//Computer Vision-ECCV 2006 pt.2.Berlin:Springer,2006:428-441.

[10]HASSNER T, ITCHER Y, KLIPER-GROSS O.Violent flows:-real-time detection of violent crowd behavior[C]//2012 IEEE computer society conference on computer vision and pattern recognition workshops,part 2 of 2.Piscataway:IEEE,2012:1-6.

[11]MABROUK A B, ZAGROUBA E.Spatio-temporal feature using optical flow based distribution for violence detection[J]. Pattern recognition letters,2017,92(Jun.1):62-67.

[12]HU J, SHEN LI, SAMUEL A, et al.Squeeze-and-excitation networks[J].IEEE transactions on pattern analysis and machine intelligence, 2020,42(8):2011-2023.

[13]SULTANI W, CHEN C, SHAH M.Real-world anomaly detection in surveillance videos[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,[Volume 9 of 13]. Piscataway:IEEE, 2018:6479-6488.

【作者简介】

郑凯东（1964—），男，广东汕头人，副教授，硕士生导师，研究方向：图形学与虚拟现实、深度学习与计算机视觉、程序设计。

江怡（2000—），女，福建漳州人，硕士研究生，研究方向：计算机视觉、深度学习。

（收稿日期：2024-04-04）