doi:10.3969/j.issn.1001-4616.2026.03.012

基于多模态特征融合的视频暴力行为识别方法

丁瑞祺,吉根林,巢　新,胡泽生,王占岭

(南京师范大学计算机与电子信息/人工智能学院,江苏 南京 210023)

[**摘要**]　视频暴力行为识别在智能安防领域发挥着重要作用. 现有方法大多使用单一模态,提供的信息有限,难以充分表达复杂的暴力行为,在实际场景下的适用性较低. 对此,本文提出了一种基于多模态特征融合的视频暴力行为识别方法,通过融合 RGB 视频帧提供的人体姿态信息、RGB 帧差提供的人体运动信息以及深度图提供的人物间相对位置信息,增强了暴力行为的特征表示,使其能够有效应对人群打斗中的遮挡现象和开放场景的光照变化. 在特征融合阶段利用注意力机制,模型自适应地学习不同模态特征的重要性,动态分配不同模态特征的权重,从而提高真实环境下暴力行为识别的准确性和鲁棒性. 所提方法在 RWF-2000、Hockey 和 Movies 数据集上进行实验,识别准确率分别为 88.75%、99%和 100%. 结果表明,所提方法能够进行有效的暴力行为识别.

[**关键词**]　视频暴力行为识别,视频理解,多模态特征融合,注意力机制
[**中图分类号**]TP391.41　[**文献标志码**]A　[**文章编号**]1001-4616(2026)03-0103-10

Video Violence Recognition Method Based on Multi-modal Feature Fusion

Ding Ruiqi,Ji Genlin,Chao Xin,Hu Zesheng,Wang Zhanling

(School of Computer and Electronic Information/School of Artificial Intelligence,Nanjing Normal University,Nanjing 210023,China)

Abstract:Video violence recognition plays a significant role in the field of intelligent security. Most existing methods rely on a single modality,providing limited information and making it difficult to fully represent complex violent behaviors,thus having relatively low applicability in real-world scenarios. This paper proposes a video violence recognition method based on multi-modal feature fusion,which enhances the feature representation of violent behavior by integrating the human pose information provided by RGB frames,the human motion information provided by RGB frame differences,and the relative position information between characters provided by depth maps,so that it can effectively cope with the occlusion phenomenon in crowd fights and the lighting changes in open scenes. During the feature fusion process,attention mechanism is applied to enable the model to adaptively learn the importance of multi-modal features and dynamically allocate weights for different features,thereby improving the accuracy and robustness of violence recognition in real-world environments. The proposed method was tested on the RWF-2000,Hockey,and Movies datasets,achieving recognition accuracies of 88.75%,99% and 100%,respectively. The results indicate that the proposed method can effectively identify violent behaviors.

Key words:video violence recognition,video understanding,multi-modal feature fusion,attention mechanism

近年来,基于视频的暴力行为识别受到广泛关注. 越来越多的监控摄像头被部署在各类公共场所,用来预防和阻止暴力行为并提供犯罪线索和证据. 然而,监控人员不可能实时准确地监控所有视频数据. 因此,从视频中自动识别暴力行为变得至关重要. 一个有效的暴力行为识别方法能够节约用于监控的人力资源,降低漏检风险,减少反应时间,从而保护公众的生命财产安全[1].

早期的视频暴力行为识别主要依赖于手工制作特征[2-4],即设计各种描述符用于捕捉视频中出现的暴力动作. 然而,这类方法只能针对特定数据集,泛化性较差. 随着深度学习的兴起,各类神经网络展示出强大的特征提取和泛化能力,端到端的深度学习方法逐渐成为视频暴力行为识别的主流方法. 按照所使用的模态[5],视频暴力行为识别方法大致可分为两类:第一类方法基于 RGB 模态[6-7]. 从 RGB 视频中提取外观、姿态特征,利用光流法[8]或帧差法[9]提取运动特征,使用深度学习模型对特征进行分类,判断视频中是否存在

收稿日期:2025-01-21.
基金项目:国家自然科学基金项目(41971343).
通讯作者:吉根林,博士,教授,博士生导师,研究方向:大数据挖掘与人工智能. E-mail:glji@njnu.edu.cn

暴力行为. 然而,在人群打斗过程中,个体经常会被其他对象遮挡,导致提取的人物信息不完整;开放场景的光照变化也影响着光流法的使用效果. 第二类方法基于骨骼模态,使用图神经网络进行建模[10]. 人体骨骼能够表征人体运动,且对于背景和光照变化不敏感. 然而,在拥挤的、存在大量遮挡现象的人群中进行姿态估计的精度较低,当人体骨骼不能被很好地检测到时,识别准确率将受到影响.

单一模态提供的信息有限,难以充分表达复杂的暴力行为,从不同模态中分别提取特征并进行融合则能增强特征表示,达到互补效果,从而提高暴力行为识别的准确性和鲁棒性. 在人群打斗过程中,各人物之间相对位置的变化是识别暴力所需补充的信息. 考虑到深度模态在部分遮挡的情况下依然可以保持对人物形态和距离的感知,能够提供人物之间的相对位置信息,并且独立于光照条件,在光照变化时可以提供稳定的空间信息,能够弥补 RGB 模态的缺陷. 因此,本文提出一种基于多模态特征融合的视频暴力行为识别方法,从 RGB 模态中提取人体姿态特征和人体运动特征,从深度模态中提取人物之间的相对位置特征,利用注意力机制[11]对 3 种特征进行充分融合,达到互补效果,实现暴力行为识别. 本文的主要创新点如下:

(1)提出了一种基于多模态特征融合的视频暴力行为识别方法. 提出了一个三通道融合网络(tri-stream fusion net,TSFN),3 个通道分别提取人体姿态特征、人体运动特征以及人物之间的相对位置特征,将 3 种特征进行融合,从而增强暴力行为的特征表示,使其能够有效应对人群打斗中的遮挡现象和开放场景的光照变化.

(2)提出了一种基于注意力机制的特征融合方法. 在特征融合阶段,模型能够自适应地学习不同模态特征的重要性,动态分配不同模态特征的权重,从而提高了在真实环境下进行暴力行为识别的准确性和鲁棒性.

1 相关工作

1.1 视频暴力行为识别研究现状

目前,基于深度学习的方法已经成为视频暴力行为识别领域的主流方法. Ding 等[12]使用三维卷积神经网络(3D convolutional neural network,3D CNN)来识别视频中的暴力行为. 该网络是二维卷积神经网络(2D convolutional neural network,2D CNN)的扩展,将二维滤波器扩展到时间轴,从而捕获时空特征. 然而,这种扩展需要更多的参数. 为了改进这一点,Li 等[13]提出了一种高效的 3D CNN 架构,通过减少内核大小和特征重用特性进一步减少参数的数量. 另一种方法是先处理空间信息,再处理时间信息,即首先使用 2D CNN 提取空间特征,然后使用长短期记忆网络(long short-term memory,LSTM)聚合这些空间特征中的时间信息. 然而,常规的 LSTM 仅接受一维输入,因此对二维空间特征进行投影可能会导致信息的丢失. 为了克服这一问题,Sudhakaran 和 Lanz[9]提出使用卷积长短期记忆网络(convolutional long short-term memory,ConvLSTM)跨时间聚合由预先训练的 CNN 提取的空间特征. 实验表明,ConvLSTM 在较少的参数下获得了比常规 LSTM 更好的结果. 此外,他们将相邻两帧的差而非原始帧输入模型,作为一种高效的变化检测器,代替计算昂贵的光流. Hanson 等[14]扩展了该架构,提出双向 ConvLSTM,它可以向前和向后聚合时间信息. 最近,Cheng 等[6]使用一个双流网络分别处理原始 RGB 视频和光流,然后结合两个通道中学习到的时空特征进行预测. Islam 等[7]也使用两个独立的通道,分别处理背景抑制帧和相邻帧的差值. 每个通道由一个 2D CNN 和一个分离卷积长短期记忆网络(separable convolutional long short-term memory,SepConvLSTM)[15]组成. Guillermo 等[16]沿用双流网络的架构,一个通道处理相邻帧的差值,另一个通道处理人群的骨架渲染图,对两个通道的特征进行融合,并使用 ConvLSTM 聚合时间信息. Su 等[10]提出另一种方法,通过研究人体各点之间的相互关系来识别暴力. 他们首先使用一个预先训练的网络来检测人体的骨骼点. 在此之后,使用自定义图神经网络对骨骼点之间的交互信息进行建模,他们称之为骨骼点交互学习(SPIL)模块.

上述方法大多使用单一模态,对暴力行为的特征表示不够充分,且难以应对人群打斗中频繁发生的遮挡现象,在实际场景下的适用性较低. 考虑到深度模态在部分遮挡的情况下依然可以保持对人物形状和距离的感知,能够提供识别暴力所需的人物间相对位置信息,本文从 RGB 模态中提取人体姿态特征和人体运动特征,从深度模态中提取人物之间的相对位置特征,对 3 种特征进行融合,优势互补,实现暴力行为识别.

1.2 深度估计技术

在视频暴力行为识别任务中,深度估计技术具有重要意义. 通过准确估计场景中的深度信息,能够更好地理解个体之间的相对位置,从而更精确地捕捉暴力行为的发生和发展. 深度指的是拍摄物体的表面到相机的距离,即图 1 中 $P^{(C)}$ 点到 $Z^{(C)}$ 点的坐标,深度图表示图中像素点对应像素的深度值,深度估计即将输入的图像转化为对应的深度图.

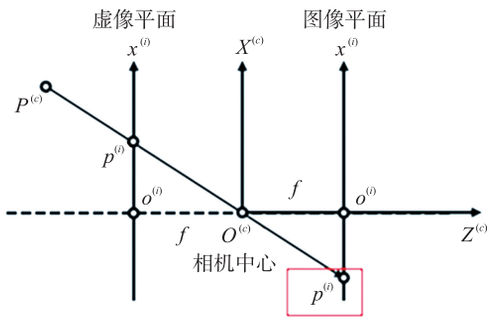


图 1 深度估计的基本原理

Fig. 1 Basic principle of depth estimation

近年来,基于深度学习的深度估计方法成为主流方法. 通过端到端训练神经网络,模型直接从单帧图像中预测深度图,显著降低了硬件依赖. 有监督的深度估计方法利用深度神经网络从图像中提取特征,并以带有真实深度信息的图像作为监督信号,从而学习图像信息与深度之间的映射关系. Eigen 等[17]首次将深度学习应用于单张图像的深度估计任务,他们设计了一种多尺度网络结构,能够提取不同精细程度的图像特征. 随后,Ramamonjisoa 等[18]提出了一种细化网络,通过预测 2D 位移场,对遮挡边界周围的像素重新采样,生成更清晰的重建结果,并通过预测残差深度图来优化初始深度图,进一步提升深度估计的精度. 自监督的深度估计方法在近些年也有较多的研究成果. Garg 等[19]首次提出了一种创新的自监督训练框架,该框架由深度估计网络和图像重建网络两部分构成. 深度估计网络负责从输入图像中提取目标视角的深度图,图像重建网络则基于深度图生成目标视角图像. 通过最小化重建图像与真实目标视角的差异,实现网络的训练. Godard 等[20]引入了多尺度特征匹配机制,以增强神经网络在遮挡场景下的鲁棒性,同时有效减少了边界伪影问题. 最近,Yang 等[21]通过创新性地利用大规模无标注数据,显著提升了模型的鲁棒性与通用性. 该研究不仅突破了传统方法依赖小规模标注数据的瓶颈,还通过自监督和半监督策略实现了对复杂场景的深度感知能力,成为当前深度估计领域里程碑式的成果.

2 本文方法

本文提出的基于多模态特征融合的视频暴力行为识别方法的整体框架如图 2 所示. 本文抓住识别暴力的重要因素,即人体姿态、人体运动的趋势和强度以及人物之间的相对位置关系,通过突出这 3 个要素,

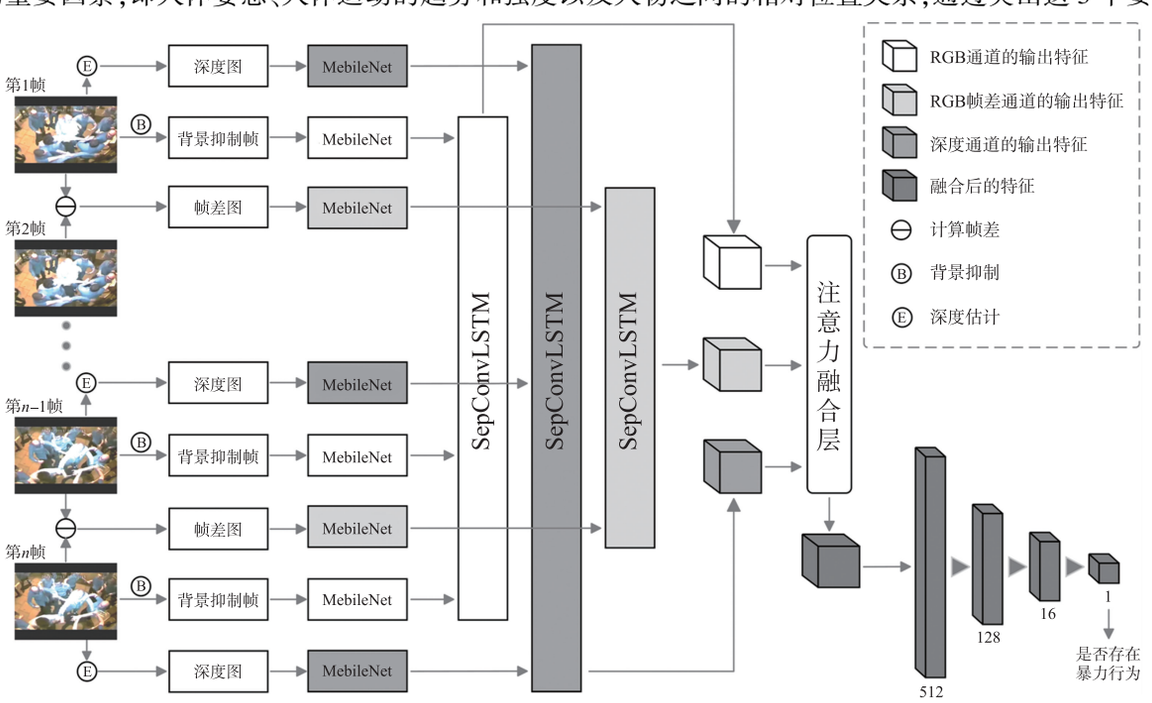


图 2 基于多模态特征融合的视频暴力行为识别框架

Fig. 2 Framework of video violence recognition based on multi-modal feature fusion

增强暴力行为与非暴力行为的区分性. 为此,本文提出一个三通道融合网络(TSFN),包括 RGB 通道、RGB 帧差通道和深度通道. 在原始 RGB 视频中采样若干帧作为模型的输入,使用 3 种不同的预处理方法分别得到背景抑制帧[7]、帧差图和深度图. 3 个通道分别从背景抑制帧、帧差图和深度图中提取人体姿态特征、人体运动特征和人物之间的相对位置特征. 然后,利用注意力机制对 3 个通道的输出特征进行融合,将融合后的特征送入分类器进行分类,判断视频中是否存在暴力行为.

2.1　RGB 通道

该通道的输入是背景抑制帧,侧重于提取人体的外观、姿态特征.

本文使用 Islam 等[7]提出的方法,对原始 RGB 帧进行背景抑制操作. 该预处理方法的目的是剥离背景,以减少背景对前景对象的干扰,同时降低计算开销. 首先,计算视频中所有帧的平均值,记作平均帧. 平均帧主要包含背景信息,因为背景在整段视频中几乎保持不变. 然后,从每一帧中减去平均帧,去除几乎不变的背景,保留运动变化的人体. 由于暴力行为主要是身体的运动,这种通过抑制背景信息来强调视频中移动对象的预处理方法,促使模型关注更相关的信息. 背景抑制帧的计算过程如下:

$$\mathrm{frame}_{avg}=\sum_{i=1}^{n}\frac{\mathrm{frame}_i}{n}, \tag{1}$$

$$\mathrm{frame}_{bs}=\mathrm{frame}_i-\mathrm{frame}_{avg}. \tag{2}$$

其中,n 表示视频的帧数,frame_i 表示视频第 i 帧,frame_{avg} 表示平均帧,frame_{bs} 表示背景抑制帧,如图 3 所示.

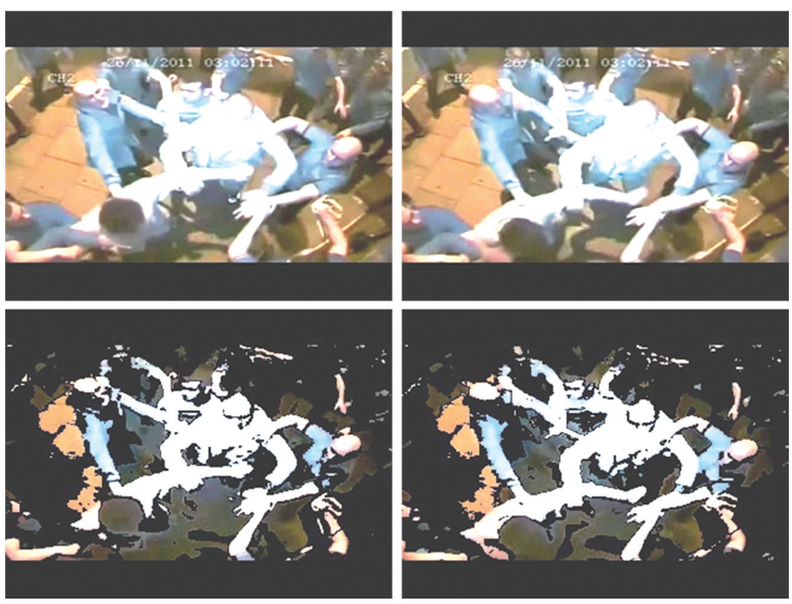


图 3　原始 RGB 帧和对应的背景抑制帧

Fig. 3　Original RGB frames and corresponding background-suppressed frames

对于经过背景抑制处理后的每一帧,使用在 ImageNet 数据集[22]上预训练的 MobileNet V2[23]作为 2D CNN 提取帧级空间特征,得到所有帧级特征后,输入 SepConvLSTM[15]聚合时间信息,最后输出 RGB 通道的视频级时空特征.

2.2　RGB 帧差通道

该通道的输入是 RGB 帧差,侧重于提取人体的运动特征.

暴力行为往往具有幅度大、速率快的特点,例如拳打、脚踢、抢夺、推搡等,这些行为比日常的身体互动更为激烈. 通过聚焦人体运动的幅度和速率,能够更好地区分暴力行为与其他行为. 帧差[9]即相邻两帧的差值,其计算过程如下:

$$\mathrm{frame_diff}_t=\mathrm{frame}_{t+1}-\mathrm{frame}_t. \tag{3}$$

其中,frame_t 表示视频第 t 帧,$\mathrm{frame_diff}_t$ 表示第(t+1)帧与第 t 帧的差值.帧差法是一种能够有效捕捉人体从前一帧到后一帧发生的运动变化的方法. 如图 4 所示,帧差图中非零像素区域的大小反映了人体运动的幅度. 由于相邻两帧的时间间隔固定,因此也在一定程度上表征着人体运动的速率、强度. 对连续多幅

帧差图进行分析,可以把握移动对象的运动趋势.

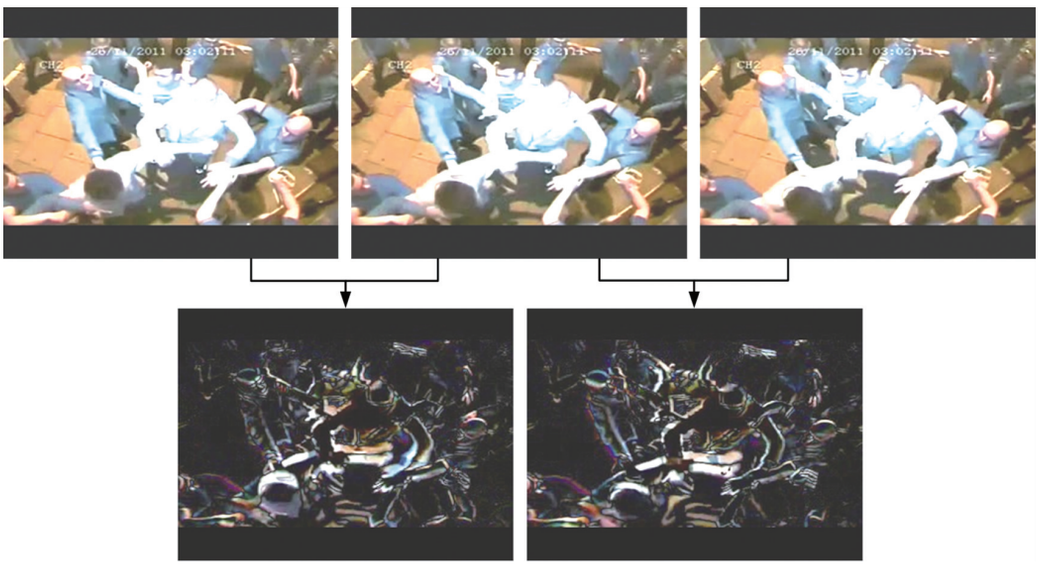


图 4　原始 RGB 帧和对应的帧差图

Fig. 4　Original RGB frames and corresponding frame difference maps

使用帧差法对原始 RGB 帧进行处理,得到所有相邻帧的帧差图. 与 RGB 通道相同,选择在 ImageNet 数据集上预训练的 MobileNet V2 作为 2D CNN 提取每一幅帧差图的空间特征,得到所有帧差图的空间特征后,输入 SepConvLSTM 编码时间信息,学习人体在帧间的运动变化,最后输出 RGB 帧差通道的时空特征.

2.3　深度通道

该通道的输入是深度图,侧重于提取人物之间的相对位置特征.

深度图是一种灰度图像,每个像素点的值对应于该点在三维空间中相对摄像机的距离,用于反映场景中物体到相机的距离信息. 其中较亮的像素表示物体离相机较近,较暗的像素表示物体离相机较远.

深度图通过提供三维空间位置信息,能够有效解析场景中的空间拓扑结构与动态交互模式,为复杂环境下的暴力识别提供关键支撑. 如图 5 所示,深度图的价值主要体现在以下几个方面:首先,在人群打斗过程中,个体时常会被其他对象遮挡,二维的 RGB 图像难以准确解析被遮挡的人物信息,而深度图通过提



图 5　原始 RGB 帧和对应的深度图

Fig. 5　Original RGB frames and corresponding depth maps

供额外的三维空间信息,可以定位被遮挡目标的近似位置,从而提升应对遮挡的能力;其次,深度图能够刻画攻击者与被攻击者的空间分布特征,为推断暴力意图提供依据. 例如,当多名攻击者围绕单一目标形成包围态势,或被攻击者分散逃离时,深度图能够清晰地反映这种动态变化,从而有效区分正常社交互动(缓慢接近)与暴力冲突(快速聚拢);最后,在光照变化或背景杂乱的监控场景中,深度图独立于颜色与纹理信息,仅依赖几何结构特征,能够稳定表征人物的空间关系,减少环境噪声对行为分析的干扰.

鉴于 Yang 等[21]提出的深度估计方法精度高、泛化能力强,并且推理速度能够达到 50 ms/帧,适合轻量级部署,本文使用该方法对原始 RGB 帧进行深度估计,得到连续的深度图. 对于每一幅深度图,选择 MobileNet V2 作为 2D CNN 提取空间特征,得到所有深度图的空间特征后,输入 SepConvLSTM 编码时间信息,学习人物之间相对位置的变化,最后输出深度通道的时空特征.

2.4　基于注意力机制的多模态特征融合

上述 3 个通道分别输出了 3 个对应的时空特征. RGB 通道的输出特征强调人体外观信息,RGB 帧差通道的输出特征强调运动信息,深度通道的输出特征强调人物间的相对位置关系. 由于这些特征所强调的信息不同,它们在最终决策中的重要性也各不相同. 因此,在特征融合阶段,需要对不同的特征赋予不同的权重,以确保融合结果能够充分利用每种特征的优势. 为了实现这一目标,本文提出一种基于注意力机制[11]的特征融合方法,框架如图 6 所示,展示了整个特征的融合过程.

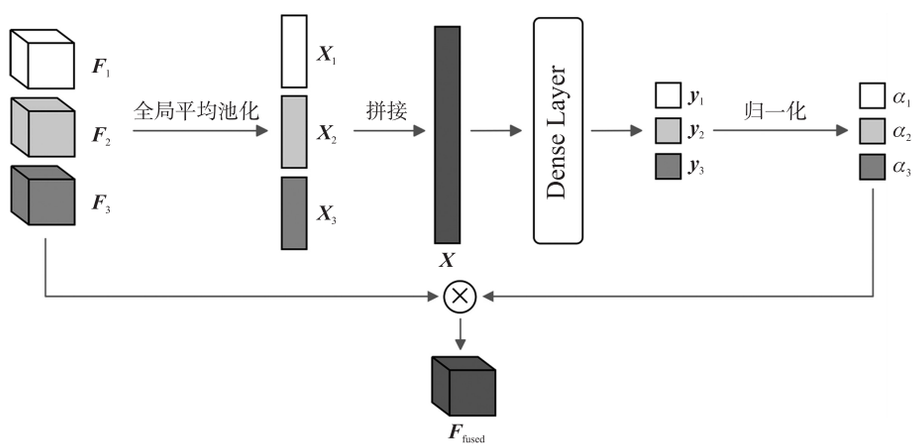


图 6　基于注意力机制的多模态特征融合方法

Fig. 6　Multi-modal feature fusion method based on attention mechanism

各通道的输出特征记为 F_1,F_2,F_3,均为形状相同的三维张量,即 $F_i\in\mathbf{R}^{H\times W\times C}$,其中 H 是高度,W 是宽度,C 是通道数. 首先,对每个特征进行全局平均池化[24],以减少特征维度:

$$\boldsymbol{X}_i=\frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W}F_i(h,w,c),\quad \forall c\in\{1,2,\cdots,C\}. \tag{4}$$

其中,$\boldsymbol{X}_i\in\mathbf{R}^C$ 表示经过全局平局池化后的特征向量. 然后,将上述 3 个特征向量进行拼接,得到一个综合特征向量 $\boldsymbol{X}$:

$$\boldsymbol{X}=[\boldsymbol{X}_1^{\mathrm{T}},\boldsymbol{X}_2^{\mathrm{T}},\boldsymbol{X}_3^{\mathrm{T}}]^{\mathrm{T}}, \tag{5}$$

其中,$\boldsymbol{X}\in\mathbf{R}^{3C}$表示拼接后的特征向量. 这一步骤使得模型能够同时评估所有类型的特征. 接下来,计算每种类型特征的注意力分数. 具体来说,使用一个全连接层(Dense Layer)对拼接后的特征向量 $\boldsymbol{X}$ 进行线性变换,并通过激活函数 ReLU 引入非线性:

$$\boldsymbol{Y}=\sigma(\boldsymbol{W}\boldsymbol{X}+\boldsymbol{b}), \tag{6}$$

其中,$\boldsymbol{W}\in\mathbf{R}^{3\times 3C}$是权重矩阵,$\boldsymbol{X}\in\mathbf{R}^{3C}$是拼接后的特征向量,$\boldsymbol{b}\in\mathbf{R}^3$ 是偏置向量,σ 是激活函数,$\boldsymbol{Y}\in\mathbf{R}^3$ 是包含 3 个注意力分数的向量,$\boldsymbol{Y}=(y_1,y_2,y_3)^{\mathrm{T}}$,每个元素分别对应一种特征的重要性. 使用 Softmax 函数对注意力分数进行归一化,得到注意力权重:

$$\boldsymbol{\alpha}=\mathrm{Softmax}(\boldsymbol{Y})=\left[\frac{e^{y_1}}{\sum_{j=1}^{3}e^{y_j}},\frac{e^{y_2}}{\sum_{j=1}^{3}e^{y_j}},\frac{e^{y_3}}{\sum_{j=1}^{3}e^{y_j}}\right]. \tag{7}$$

其中,$\boldsymbol{\alpha} \in \mathbf{R}^3$ 是包含 3 个注意力权重的向量,$\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,\boldsymbol{\alpha}_3)^{\mathrm{T}}$,每个权重都介于 0 和 1 之间,且所有权重之和等于 1. 然后,根据注意力权重对 3 个通道的输出特征进行加权求和,得到最终特征 F_{fused}:

$$F_{\text{fused}} = \sum_{i=1}^{3} \boldsymbol{\alpha}_i F_i. \tag{8}$$

其中,$F_i \in \mathbf{R}^{H\times W\times C}$是第 i 个通道的输出特征,$\boldsymbol{\alpha}_i$ 是对应的注意力权重,$F_{\text{fused}} \in \mathbf{R}^{H\times W\times C}$是融合后的特征.

最后,将融合后的特征展平为一维特征向量,输入全连接层进行分类,判断视频中是否存在暴力行为.

所提方法融合了 RGB 通道的人体姿态特征、RGB 帧差通道的人体运动特征,以及深度通道的人物间相对位置特征,各模态特征优势互补,增强了暴力行为的特征表示. 并且,模型能够自适应地学习不同模态特征的重要性,动态分配不同模态特征的权重,从而提高了在复杂的真实环境下进行暴力行为识别的准确性和鲁棒性.

3 实验

3.1 实验设置

本文实验基于 Tensorflow 框架,在 NVIDIA GeForce RTX 3090 GPU 上训练和测试模型.

为了降低计算复杂度、提高模型的训练效率,本文使用均匀采样的方法,从每个视频中采样 32 帧,并将每一帧的大小调整为 320×320. 在进入模型之前,对它们进行随机裁剪,将大小调整为 224×224,形状为 32×224×224×3 的 RGB 帧序列被输入模型.

在 RGB 通道中,对 RGB 帧序列进行背景抑制处理,得到形状为 32×224×224×3 的背景抑制帧序列. 然后,使用在 ImageNet 数据集上预训练的 MobileNet V2 提取得到形状为 32×7×7×56 的空间特征. 最后,输入 SepConvLSTM 聚合时间信息,得到形状为 7×7×64 的时空特征.

在 RGB 帧差通道中,对 RGB 帧序列使用帧差法进行处理,得到形状为 31×224×224×3 的帧差序列. 然后,使用在 ImageNet 数据集上预训练的 MobileNet V2 提取得到形状为 31×7×7×56 的空间特征. 接着,输入 SepConvLSTM 聚合时间信息,得到形状为 7×7×64 的时空特征.

在深度通道中,对 RGB 帧序列进行深度估计,得到形状为 32×224×224×3 的深度图序列. 然后,使用 MobileNet V2 提取得到形状为 32×7×7×56 的空间特征. 最后,输入 SepConvLSTM 聚合时间信息,得到形状为 7×7×64 的时空特征.

3 个通道的输出特征均被送入窗口大小为 2×2 的最大池化层进行池化,得到 3 个形状为 3×3×64 的特征. 根据注意力权重对这 3 个特征进行融合,得到形状为 3×3×64 的最终特征. 将该特征展平得到长度为 576 的一维特征向量,使用 512 → 128 → 16 → 1 的全连接层对其进行分类.

训练迭代次数设置为 150,批处理大小设置为 4. 使用 AMSGrad 作为优化器,初始学习率设置为 4×10^{-4},每 5 次迭代后学习率减半,达到 5×10^{-5}后保持不变. 使用二元交叉熵作为损失函数,评估模型预测值和真实值之间的误差.

3.2 数据集和评价指标

本文共使用 3 个公开数据集对所提方法进行评估,分别是 RWF-2000 数据集[6]、Hockey 数据集[25]以及 Movies 数据集[25].

RWF-2000 数据集由现实世界中的监控摄像机捕获的 2 000 个视频片段组成,每个视频的时长为 5 s,帧率为 30 帧. 该数据集规模庞大且场景真实,是研究真实监控场景下暴力行为识别的重要资源. Hockey 数据集包含 1 000 个从曲棍球比赛中收集的视频,每个视频 50 帧,所有视频都有着相似的背景. Movies 数据集是一个相对较小的数据集,包含 200 个来自不同电影的剪辑. 以上数据集均拥有视频级标签,即每个视频片段被标注为“暴力”或“非暴力”,并且正负样本的比例均为 1∶1,即一半的视频包含暴力行为,另一半属于非暴力活动.

对于每个数据集,选取其中 20%作为测试集,用于性能评估,其余 80%作为训练集,用于模型训练.

本文采用准确率作为评价指标. 在分类任务中,准确率是衡量模型性能最常用的指标. 它表示模型正确预测的样本数占总样本数的比例,计算公式如下:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}. \tag{9}$$

式中,TP 表示真正例,TN 表示真反例,FP 表示假正例,FN 表示假反例.

3.3 实验结果与分析

3.3.1 与其他方法的对比实验

为了评估本文方法,将其与主流的暴力行为识别方法进行比较. 在 3 个数据集上的对比结果如表 1 所示.

表 1 不同暴力行为识别方法在公开数据集上的准确率

Table 1 Accuracy of different violence recognition methods on public datasets

方法名称	RWF-2000	Hockey	Movies
Flow Gated Net[6]	87.25%	98%	100%
SPIL Convolution[10]	89.30%	96.8%	98.5%
Two Stream SepConvLSTM[7]*	88.50%	98.5%	100%
Lightweight 2D+3D CNN[26]	81.98%	94.71%	100%
U-Net+LSTM[27]	82.00%	96.1%	99.5%
Human Skeleton+Change Detection[16]	90.25%	94.5%	98.5%
Residual+ResNet50+POT[28]	84.00%	95%	97.5%
Our Method:Tri-Stream Fusion Net	88.75%	99%	100%

* 表示按照论文作者发布的代码和规范,在我们的设备上重新训练的结果.

与依赖于单一 RGB 数据的方法相比,本文方法利用多模态特征融合,不仅从 RGB 数据中提取人体外观特征和运动模式,还引入深度模态以提供三维空间信息,从深度数据中提取人物之间的相对位置特征,从而有效缓解了遮挡场景下的信息丢失问题. 在特征融合阶段,利用注意力机制动态分配不同模态特征的权重,使得外观、运动与空间位置特征充分融合,显著增强了暴力行为的特征表示. 尽管在 RWF-2000 数据集上,本文方法的表现略低于 Su 等[10]和 Guillermo 等[16]提出的方法,主要原因在于这两种方法均基于骨骼模态,在捕捉人体运动特征方面更具优势;而本文方法在不同的数据集上均达到较高的识别准确率,展现出更强的适应性和鲁棒性.

3.3.2 消融实验

(1)各通道的有效性

为了验证所提出的三通道融合网络中每个通道的有效性,在 RWF-2000 数据集上进行消融实验,依次使用单一通道、2 个通道的组合以及全部 3 个通道(对多通道的特征进行融合时,不使用注意力模块,不同通道的特征均按照同等权重进行融合),结果如表 2 所示. 可以看出,仅使用单一通道时,RGB 帧差通道的效果更为出色,这表明运动特征是识别暴力的主要特征;使用 2 个通道并进行融合的效果优于使用单一通道的效果;当使用全部 3 个通道并进行融合时,效果最佳. 这表明不同模态特征能够优势互补,从而提高识别准确率.

表 2 模型选用不同通道时在 RWF-2000 数据集上的准确率

Table 2 Accuracy of the model with different streams on the RWF-2000 dataset

RGB 通道	RGB 帧差通道	深度通道	准确率
√			82.75%
	√		85.25%
		√	78.50%
√	√		86.75%
√		√	83.25%
	√	√	86.25%
√	√	√	87.75%

(2)基于注意力机制的多模态特征融合方法的有效性

为了验证所提出的基于注意力机制的多模态特征融合方法的有效性,在 RWF-2000 数据集上进行消融实验,对 RGB 通道、RGB 帧差通道和深度通道的特征按照不同的权重进行融合,结果如表 3 所示. 首次实验不考虑不同模态特征重要性的差异,以 1∶1∶1 的权重对 3 种特征进行融合,准确率为 87.75%;第二次实验考虑到使用单一通道时,RGB 帧差通道的结果最优,因此,提高该通道的特征的权重,以 1∶2∶1 的权重进行融合,准确率为 88%,略有提升;第三次实验使用模型通过自适应学习得到的注意力权重对 3 种特征进行融合,准确率进一步提升,达到 88.75%. 这表明利用注意力机制对不同模态的特征分配相应的权重

是必要且有效的.

(3)选用不同特征提取网络对于实验结果的影响

本文对选用不同特征提取网络对于实验结果的影响进行了探究,结果如表 4 所示. 3D CNN 作为基础模型,直接通过三维卷积处理时空信息,但受限于时序依赖建模不足,准确率最低. 2D CNN 与 LSTM 的组合通过 2D CNN 提取空间特征,再结合 LSTM 捕捉时序信息,准确率明显提升,这表明引入时序建模能有效增强模型对动态变化的感知能力. 使用 ConvLSTM 替换 LSTM 则在建模时间依赖的同时保留了空间局部相关性,准确率进一步提升. SepConvLSTM 在 ConvLSTM 的基础上,首先利用深度卷积独立提取各通道的空间局部特征,再通过逐点卷积实现跨通道融合. 这种分离建模策略提升了模型的特征表达能力,因此取得最优的识别准确率.

表 3　模型使用不同权重进行融合时在 RWF-2000 数据集上的准确率

Table 3　Accuracy of the model with different fusion weights on the RWF-2000 dataset

融合权重	准确率
1∶1∶1	87.75%
1∶2∶1	88.00%
注意力权重	88.75%

表 4　模型选用不同特征提取网络在 RWF-2000 数据集上的准确率

Table 4　Accuracy of the model with different feature extraction networks on the RWF-2000 dataset

特征提取网络	准确率
3D CNN	84%
2D CNN+LSTM	86.75%
2D CNN+ConvLSTM	88.25%
2D CNN+SepConvLSTM	88.75%

4　结论

本文提出了一种基于 RGB 和深度特征融合的视频暴力行为识别方法. 针对实际场景下单一模态特征表达能力有限以及遮挡现象频发等挑战,利用 RGB 和深度模态的互补性优势,应对遮挡问题,增强暴力行为的特征表示. 具体来说,本文提出一个三通道融合网络(TSFN),3 个通道分别学习人体姿态特征、人体运动特征以及人物之间的相对位置特征,并添加注意力融合模块,充分融合这 3 种特征. 所提方法在公开数据集上达到较高的识别准确率,能够有效地识别暴力行为. 未来,我们将进一步探索实时暴力行为识别方法.

[参考文献]

[1] 王晓龙. 基于轨迹分析的暴力行为识别算法研究[D]. 上海:上海交通大学,2015.

[2] Senst T,Eiselein V,Kuhn A,et al. Crowd violence detection using global motion-compensated lagrangian features and scale-sensitive video-level representation[J]. IEEE Transactions on Information Forensics and Security,2017,12(12):2945-2956.

[3] Chen D,Wactlar H,Gao C,et al. Recognition of aggressive human behavior using binary local motion descriptors[C]//30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Vancouver,BC,Canada:IEEE,2008:5238-5241.

[4] Deb T,Arman A,Firoze A. Machine cognition of violence in videos using novel outlier-resistant VLAD[C]//17th IEEE International Conference on Machine Learning and Applications. Orlando,Florida,USA:IEEE,2018:989-994.

[5] Sun Z,Ke Q,Rahmani H,et al. Human action recognition from various data modalities:A review[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2023,45(3):3200-3225.

[6] Cheng M,Cai K,Li M. RWF-2000:an open large scale video database for violence detection[C]//25th International Conference on Pattern Recognition. Milan,Italy:IEEE,2021:4183-4190.

[7] Islam Z,Rukonuzzaman M,Ahmed R. Efficient two stream network for violence detection using separable convolutional LSTM[C]//International Joint Conference on Neural Networks. Shenzhen,China:IEEE,2021:1-8.

[8] Ilg E,Mayer N,Saikia T,et al. FlowNet 2.0:evolution of optical flow estimation with deep networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu,Hawaii,USA:IEEE,2017:1647-1655.

[9] Sudhakaran S,Lanz O. Learning to detect violent videos using convolutional long short-term memory[C]//14th IEEE International Conference on Advanced Video and Signal Based Surveillance. Lecce,Italy:IEEE,2017:1-6.

[10] Su Y,Lin G,Zhu J,et al. Human interaction learning on 3D skeleton point clouds for video violence recognition[C]//16th European Conference on Computer Vision(ECCV). Glasgow,UK:Springer,2020:74-90.

[11] Hu J,Shen L,Albanie S,et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020,42(8):2011-2023.

[12] Ding C,Fan S,Zhu M,et al. Violence detection in video by using 3D convolutional neural networks[C]//10th International Symposium on Visual Computing(ISVC). Las Vegas,NV,USA:Springer,2014:551-558.

[13] Li J,Jiang X,Sun T,et al. Efficient violence detection using 3D convolutional neural networks[C]//16th IEEE International Conference on Advanced Video and Signal Based Surveillance. Taipei,China:IEEE,2019:1-8.

[14] Hanson A,Pnvr K,Krishnagopal S,et al. Bidirectional convolutional LSTM for the detection of violence in videos[C]// Computer Vision-ECCV 2018 Workshops. Munich,Germany:Springer,2019:280-295.

[15] Pfeuffer A,Dietmayer K. Separable convolutional LSTMs for faster video segmentation[C]//2019 IEEE Intelligent Transportation Systems Conference(ITSC). Auckland,New Zealand:IEEE,2019:1072-1078.

[16] Guillermo G,Sanmiguel J. Human skeletons and change detection for efficient violence detection in surveillance videos[J]. Computer Vision and Image Understanding,2023,233:103739.

[17] Eigen D,Puhrsch C,Fergus R. Depth map prediction from a single image using a multi-scale deep network[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal,Canada:Curan Associates,2014:2366-2374.

[18] Ramamonjisoa M,Du Y,Lepetit V. Predicting sharp and accurate occlusion boundaries in monocular depth estimation using displacement fields[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle,Washington,USA:IEEE,2020:14648-14657.

[19] Garg R,Bg V K,Carneiro G,et al. Unsupervised cnn for single view depth estimation:geometry to the rescue[C]//14th European Conference on Computer Vision(ECCV). Amsterdam,The Netherlands:Springer,2016:740-756.

[20] Godard C,Mac Aodha O,Firman M,et al. Digging into self-supervised monocular depth estimation[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul,South Korea:IEEE,2019:3828-3838.

[21] Yang L H,Kang B Y,Huang Z L,et al. Depth anything:unleashing the power of large-scale unlabeled data[C]//2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle,WA,USA:IEEE,2024:10371-10381.

[22] Krizhevsky A,Sutskever I,Hinton E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing System 25(NIPS 2012). NV,USA:Curran Associates,2012:1106-1114.

[23] Sandler M,Howard A,Zhu M,et al. MobileNetV2:inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City,UT,USA:IEEE,2018:4510-4520.

[24] Lin M,Chen Q,Yan S. Network in network[C]//2nd Learning Representations(ICLR). Banff,Canada:Springer,2014.

[25] Nievas B,Suarez D,Sukthankar R,et al. Violence detection in video using computer vision techniques[C]//14th International Conference on Computer Analysis of Images and Patterns. Seville,Spain:Springer,2011:332-339.

[26] Wang W,Dong S,Zou K,et al. A lightweight network for violence detection[C]//Proceedings of the 2022 5th International Conference on Image and Graphics Processing. New York,USA:Association for Computing Machinery,2022:15-21.

[27] Vijeikis R,Raudonis V,Dervinis G. Efficient violence detection in surveillance[J]. Sensors,2022,22(6):2216.

[28] Honarjoo N,Abdari A,Mansouri A. Violence detection in compressed video[J]. Multimedia Tools and Applications,2024,83(29):73703-73716.

[责任编辑:黄　敏]