一种时序边界注意力循环暴力行为检测神经网络

刘邦义[1]，周激流[1]，张卫华[2]
(1. 四川大学电子信息学院，成都 610065；2. 四川大学计算机学院，成都 610065)

摘 要：暴力行为检测是行为识别的一个重要研究方向，在网络信息审查和智能安全领域具有广阔的应用前景. 针对目前的时序模型在复杂背景下不能有效提取人体运动特征和常规循环神经网络无法联系输入上下文的问题，本文提出一种时序边界注意力循环神经网络 TEAR-Net. 首先，以本文提出的一种全新的运动特征提取模块 MOE 为基础，在保留输入视频段序列背景信息的前提下加强运动边界区域. 运动边界对于动作识别的作用要远大于图像其他区域，因此运动边界加强能够有效提高动作特征的提取效率，从而提升后续网络的识别精度. 其次，引入了一种全新的结合上下文语境和注意力机制的循环卷积门单元(CSA-ConvGRU)，提取连续帧之间的流特征以及不同帧的独立特征，并关注关键帧，能够极大提升动作识别的效率，以少量参数和较低计算量的代价掌握视频流的全局信息，从而有效提高识别准确率. 本文提出的模型在目前最新的公开数据集 RWF-2000 和 RLVS 上进行了多种实验. 实验结果表明，本文提出的网络在模型规模和检测精度上均优于目前主流的暴力行为识别算法.

关键词：暴力行为；时序信息；运动边界；注意力机制；上下文
中图分类号：TP391 **文献标识码**：A DOI：10.19907/j.0490-6756.2023.023003

Temporal edge attention recurrent neural networkfor violence detection

LIU Bang-Yi[1], *ZHOU Ji-Liu*[1], *ZHANG Wei-Hua*[2]
(1. College of Electronic Information Engineering, Sichuan University, Chengdu 610065, China;
2. College of Computer Science, Sichuan University, Chengdu 610065, China)

Abstract: Violence detection is one of the most important research topic in behavior recognition, which has great potential applications in network information review and intelligent security. The published works cannot keep their performance in the complexity environments, because they cannot effectively extract movement features and contact consecutive frames. Hence, a novel method is proposed in this paper, which is referred to as temporal edge attention recurrent neural network (TEAR-Net). First, we propose a novel motion object enhancement (MOE) module, which enhances the motion edge while keeping the background information of the video sequences. Because the motion edge has a much greater effect on motion recognition than other areas of the image, the enhancement of motion edge can effectively improve the extraction efficiency of action features, and thus the recognition accuracy is improved. Then we introduce a novel recurrent convolutional gate unit CSA-ConvGRU, which combines context and attention mechanism. It can extract the stream features among consecutive frames and the independent features of each frame. Attention mechanism can help to focus on key frames, which greatly improve

收稿日期：2022-01-17
基金项目：四川省科技计划(2022YFQ0047)
作者简介：刘邦义(1998—)，男，硕士研究生，主要研究方向为计算机视觉、模式识别等. E-mail：228980603@qq.com
通讯作者：张卫华. E-mail：zhangweihua@scu.edu.cn

the efficiency of action recognition, capture the global information of the video stream with a lower cost, and thus effectively improve recognition accuracy. The proposed model has been tested on the currently lastest public datasets RWF-2000 and RLVS. The experimental results show that the proposed model outperforms the state-of-the-art violence detection algorithms in terms of computational cost and detection accuracy.

Keywords: Violence; Temporal information; Motion edge; Attention mechanism; Context

1 引　言

伴随着城镇化的规模不断扩大和人口的聚集,群众对公共区域安全的监管需求日益提升,各种监控设备被广泛部署. 如今大部分的公共安全监管仍然采取基于人工观察的方式,因此尽管底层感知设备布置较为完善,但是仍然具有较高的漏检率和错检率. 这导致有时不能及时处理应急事件. 因此一种基于计算机图像技术的异常行为检测算法是急需的,由于图像技术的快速、准确、轻量化、高拓展性且易于维护等特点,这类方法具有较大的社会价值和应用前景[1].

视频作为安防监控的载体,由连续的图像构成,具有高度的时空相关性[2]. 相较于图片,视频往往具有更为丰富的信息,例如帧间的时间语义信息和空间语义信息. 然而,视频具有输入维度高,时空信息难以解耦提取等难题. 因此,如何提出一种针对视频的实时准确轻量化的异常行为检测算法是十分困难的,具有较高的研究价值.

近年来,行为识别的研究热度不断提高,许多算法相继被提出,根据特征提取方法的不同,可以被分为基于传统图像处理和基于深度学习的方法. Wang 等[3]提出密集轨迹提取的相关算法(Dense Trajectories, DT),通过将三种特征描述子方向梯度直方图融合编码后进行分类,最终取得了较好的效果. 传统方法具有不需要训练以及对硬件设备要求低等优点,但是它们在面对背景复杂等情况下往往难以维持较好的表现,因为其提取的特征泛化能力不足.

神经网络近年来得到飞速发展,卷积神经网络(Convolutional Neural Network, CNN)作为神经网络的代表算法在诸多重要图像任务如分类[4]和分割[5]上均取得了瞩目的成绩. 深度学习能够获得样本深层特征表示,故基于深度学习的方法往往具有较强的鲁棒性和较高的识别精度. 行为识别算法可被分为多流卷积神经网络(Multi Stream CNN, MSCNN)的模型、时空序列模型和时间序列模型[6]. Simonyan 等[7]提出双流 CNN 模型,该模型分别从单帧 RGB 图像和多帧稠密光流图中提取视频的时空信息,最后将特征融合进行分类,取得较好的效果. 尽管多流 CNN 模型在提取动作信息特征有一定优势,但是光流信息的提取需要大量算力,增大了计算开销,难以在边缘设备部署.

时空模型采用三维卷积提取时空特征. Tran 等[8]在三维卷积的基础提出 3D 卷积模型(Convolutional 3D,C3D),C3D 为三维卷积和池化的线性组合,故模型简单训练速度较快. 为了缓解 C3D 参数量大的问题,Qiu 等[9]提出了伪 3D(Pseudo-3D, P3D)模型,Tran 等[10]提出了 R(2+1)D 模型. 两种模型的都采用将时空域分离进行灵活组合的思路,从而在保证模型参数量小的同时,能够提升所提取特征的鉴别性和鲁棒性.

时序模型采用 CNN 与递归神经网络级联的方式提取时序特征. Hochreiter 和 Schmidhuber[11]提出长短期记忆单元(Long Short Term Memory, LSTM),被认为是时序模型最有代表性的方法之一. 此后,Donahue 等[12]提出了长期循环卷积网络(Long-term Recurrent Convolution Network, LRCN),LRCN 将 CNN 和由 LSTM 单元组成的递归神经网络级联,分别提取输入的空间信息和时间信息. Melis 等[13]在自然语言处理领域中发现随着模型的复杂程度的加深,LSTM 单元中的输入与隐藏层之间的相关性会逐渐消失,进而提出了形变 LSTM. Cho 等[14]提出循环门单元(Gated Recurrent Units, GRU),GRU 相较于 LSTM 具有参数量更少,训练速度更快等优点. Shi 等[15]提出了 ConvLSTM,ConvLSTM 通过将 LSTM 中的全连接层替换为卷积层,使该结构不仅可以建立时序关系还可以像 CNN 一样提取局部空间特征. Lin 等[16]在 ConvLSTM 中增加基于记忆单元的自注意力模块(Memory-based Self-Attention Module, SAM)预测记录全局时空特征,以此来提取输入中具有代表性的时空特征.

尽管时序模型取得了较好的识别效果,但该类

方法在复杂背景下不能有效提取人体运动特征. 相较于其他任务的视频数据而言，安防监控视频需对大范围场景进行监控，因此安防监控视频中的运动主体并不突出，目标往往较小，同时背景非常复杂. 而时序模型针对具有突出主体的视频数据具有较好的识别效果，而当背景复杂并且运动目标较小时，该类方法难以保持其良好的识别性能. 其主要原因是冗杂信息占据视频的主导地位，直接使用原始视频训练如 LRCN 为代表的时序模型时难以取得良好的识别效果. 而这些算法为缓解这个问题往往使用帧差法消除背景信息的干扰，但与此同时运动区域的动作信息也会被一并删除.

时序模型的根基是以 ConvLSTM 为代表的递归神经网络. 该网络深度和视频长度相关，视频越长网络越深. 而递归神经网络中层级之间缺乏联系，在深度增加的同时网络会倾向关注近期输入的信息，而忽视早期信息随着深度的增加逐渐被忽视. 因此，这类模型难以对上下文信息建立有效联系.

为了缓解上述的问题，有效提取出视频主体运动的时空特征，减少复杂背景对模型的影响以提高识别准确率同时克服以往模型中上下文相关性不足的问题，本文提出了一种时序边界注意力循环神经网络(Temporal Edge Attention Recurrent Neural Network，TEAR-Net). 本文主要贡献可以总结如下：

(1) 本文提出一种全新的视频帧处理模块(Motion Object Enhancement，MOE)，MOE 可以基于帧间图像差异提取出运动边界，减少背景等因素的干扰，从而有效提高识别精度.

(2) 本文提出一种基于语境化的自注意力机制循环单元(Contextualization and Self-Attention ConvGRU，CSA-ConvGRU)对时序信息进行提取. 相较于其他结构，CSA-ConvGRU 在参数大小、显存占用和训练速度上更占优势. 此外相比于 ConvGRU，CSA-ConvGRU 对长时中的短时时序特征拥有更强的敏感性，相较于其他方法，本文提出方法在公共数据集上取得了令人满意的效果.

2　网络结构

为缓解识别率低，参数量大等问题，本文提出 TEAR-Net，整体框架图如图 1 所示，由三个主要部分组成分别为视频帧预处理模块，卷积特征提取网络和注意力-长短时特征提取网络. MOE 为视频帧预处理模块的主要成分，MOE 对聚合帧块做一系列的操作提取运动目标，在有效剔除可能造成干扰背景信息的同时保留主要运动目标. 卷积特征提取网络使用残差网络 (Residual Network，ResNet)[17] 来提取帧图像中的运动特征. 在注意力-长短时特征提取网络中，本文提出 CSA-ConvGRU，能够有效获取时间序列中的关键特征，使之能够提取更具有鉴别性的时空信息.

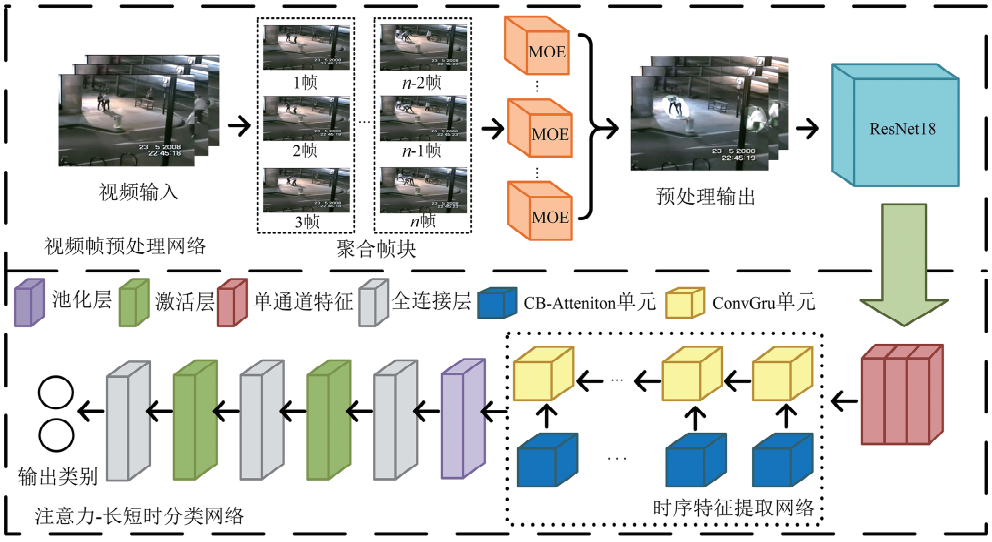


图 1　TEAR-Net 网络结构图

Fig. 1　The whole framework of the proposed TEAR-Net

为了避免歧义，本文将对一些基本内容进行数学定义. 其中，输入为视频 V，其中 V 的维度是 $V\in R^{N\times T\times C\times H\times W}$，$N$，$T$，$C$，$H$ 和 W 分别代表批尺寸，帧序列长度，图像通道数以及帧图像的长宽. V 由

帧图像组成,其中帧图像维度可以被表示为$V_t \in R^{N\times C\times H\times W}$, t 表示视频中的第 t 帧. δ 代表 sigmod 激活函数,pool 代表平均池化层.

2.1　运动目标检测模块

在行为识别中,背景信息往往不是重要特征并且通常会对结果造成极大的干扰. 因此,为了去除背景信息加强运动区域,本文提出一种全新的视频帧处理模块(Motion Object Enhancement, MOE)对运动区域进行加强. 不同于多流的方法,该方法不需计算额外的光流信息,即可对运动目标进行强化,具有快速,运算算力和硬件设施要求低等优点.

MOE 的整体框架图如图 2 所示,共分为 4 步. 首先,计算聚合帧块中相邻图像间不同通道维度的一阶导数的均方根(Root Mean Square, RMS),第 t 帧图像的 RMS 可以被表示为 v_t,其中 $v_t \in R^{N\times H\times W}$,具体计算过程由式(1)表示.

$$v_t = \sqrt{\frac{\sum_{i=t-1}^{t+1}\sum_{j=1}^{C}(V_{t+1}^{j}-V_{i}^{j})^2}{2}} \tag{1}$$

由式(1)求得图像的 RMS 后,利用全局平均池化层和激活层对空间信息 b_t 进行进一步提取,其中 $b_t \in R^{N\times 1\times H\times W}$,计算过程如式(2)所示. 通过这种方式,使 MOE 不仅可以有效提取运动边界,同时还能对提取的运动边界进行平滑与膨胀.

$$b_t = \mathrm{ReLU}(\mathrm{pool}(v_t)) \tag{2}$$

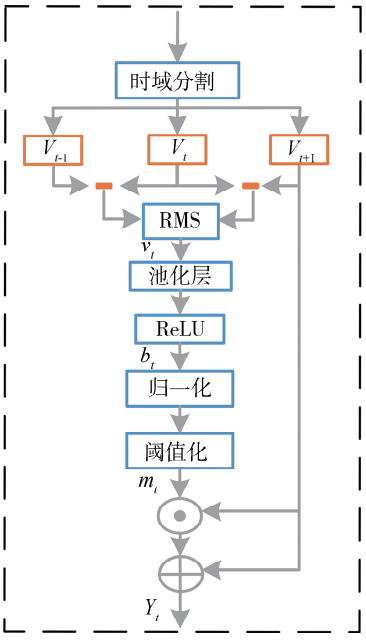


图 2　MOE 结构图

Fig. 2　The framework of MOE

其次,对得到的 b_t 进行归一化处理,并将处理的归一化数据进行掩码提取,从而得到边界清晰的运动区域 m_t. 本文采用的归一化方式是实例归一化(Instance Normalization, IN)[18],IN 可以进一步剔除图像的全局信息进而凸显个体差异. 具体计算方法如式(3)~(5)所示.

$$\mu_{ti} = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}x_{ti,lm} \tag{3}$$

$$\sigma_{ti}^2 = \frac{1}{HW}\sum_{l=1}^{W}\sum_{m=1}^{H}(x_{ti,lm}-\mu_{ti})^2 \tag{4}$$

$$y_{ti,jk} = \frac{x_{ti,jk}-\mu_{ti}}{\sqrt{\sigma_{ti}^2+\varepsilon}} \tag{5}$$

然而在式(2)中提取的运动边界仍然存在些许瑕疵,即图像中包含了非主要运动目标的噪声区域. 除此之外,运动目标的特征值往往较大,在之后进行处理的过程中会出现亮度失常的情况. 因此需要对归一化后的 b_t 的取值范围进行约束,去除不必要噪声并且避免亮度爆炸的问题,这些问题使得在之后对图像特征进行提取的时候,特征质量急剧下降从而导致较低的识别准确率. 因此,为了解决上述问题,本文在归一化操作之后引入阈值化操作,得到运动目标掩码图像 $m_t \in R^{N\times 1\times H\times W}$, $i \in R^{H\times W}$,具体的阈值化过程如式(6)所示.

$$m_{ti} = \begin{cases} 0, & \lfloor b_{ti} \rfloor < 0 \\ 0.8, & \lfloor b_{ti} \rfloor \geqslant 0 \end{cases} \tag{6}$$

得到掩码 m_t 后,将 m_t 与 V_t 点乘并与 V_t 求和组成残差结构得到最后的运动目标强化输出 $Y(t)$. MOE 中之所以采用残差方式来进行连接,是因为该方式不仅增强动态变化区域,还保留了视频中的场景信息,这使得网络被迫发现和获取差异化信息的时间特征. 具体过程如式(7)所示.

$$Y_t = V_{t+1} + m_t \odot V_{t+1} \tag{7}$$

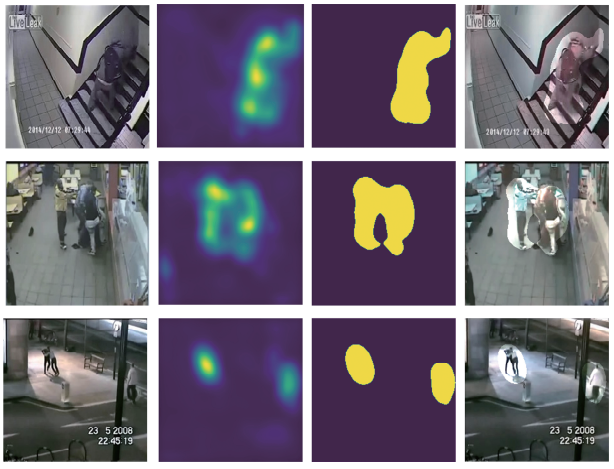
其中 $\odot$ 代表了张量之间的哈玛达积.

如图 3 所示,模块输出 $Y(t) \in R^{N\times 3\times H\times W}$ 将应该被予以关注的运动物体对应的图像区域进行凸显.

2.2　时序特征提取网络

ConvGRU 在保持参数量较少的同时,可以对时序信息进行有效提取,被广泛用于动作识别任务. 然而,随着网络深度的增加,浅层 ConvGRU 模块通过帧图像提取的隐藏层特征会随着级联层数的增加而逐渐被更新,在此过程中大部分浅层特征信息会丢失,但在动作识别领域中,浅层语义信息往往包含重要的信息,因此仅使用 ConvGRU 的识别网络无法达到令人满意的效果[16]. 为了缓解这个问题,本文将自注意力机制与 ConvGRU 相结

合，提出了 SA-ConvGRU 模块. 自注意力机制的作用是从信号中学习特征的权重，使得网络关注更重要的特征，从而保障后续分类的正确率，其早先被应用于自然语言处理领域，随后被广泛应用于图像识别、图像合成和视频预测等领域，并取得良好效果. 在 SAM 中，如图 4 所示，记忆单元M_t能够保存每一层中的部分特征，通过上一时刻的M_{t-1}与当前时刻的隐藏层输出H_t进行自相关操作并配合更新门更新输出$\hat{H}_t$与M_t，本文引入自注意力机制，使得 SA-ConvGRU 模块的输出$\hat{H}_t$能够从M_t中获取到浅层语义特征信息，从而缓解 ConvGRU 中浅层信息丢失的问题，并进而提升分类精度.



(a)　　(b)　　(c)　　(d)

图 3　各步骤效果示意图：(a)～(d)分别代表原图，b_t，m_t和处理后的示意图

Fig. 3　The samples of different steps in MOE: (a)～(d) Represent the original image, results of b_t, m_t and the processed image, respectively

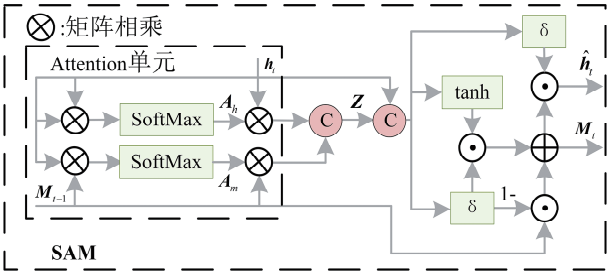


图 4　SAM 模块结构图

Fig. 4　The illustration of SAM

虽然 SA-ConvGRU 相较于 ConvGRU 可以对不同层级的特征进行学习，选择对识别任务影响较大的特征赋予更多的关注，有效缓解了 ConvGRU 难以保留浅层语义信息的问题. 但是该模块并没有建立不同时刻输入的联系，当前时刻的输入和上一时刻的隐藏层状态输出是完全独立的，忽略了上下文的关系，而这种关系对于处理时间步较长的数据是大有帮助的[13]. 为了更好地提取输入与语境之间的相关信息，增强模块对上下文的建模能力，本文将语境化模块(Contextualization Block，CB)和自注意力机制模块进行结合，提出 CSA-ConvGRU，其整体结构如图 5 所示，包含了三个单元模块：语境化模块、ConvGRU 和 SAM. CSA-ConvGRU 的整体更新递推公式如下所示：

$$\hat{x}_t, \hat{h}_{t-1} = \mathrm{CB}(x_t, h_{t-1}) \tag{8}$$

$$h_t = \mathrm{ConvGRU}(\hat{x}_t, \hat{h}_{t-1}) \tag{9}$$

$$\hat{h}_t, M_t = \mathrm{SAM}(h_t, M_{t-1}) \tag{10}$$

在 CB 中，上下文关联性是通过交互模式来提取的，具体公式如式(11)(12)所示. CB 通过引入额外的门控运算，使得两个输入之间能进行计算交互，最终使得输入x_t与h_{t-1}之间具有更加丰富的交互表示，从而提高 ConvGRU 模型的泛化能力与时序建模能力.

$$\hat{x}_t = 2 \times \delta(W_h * h_{t-1}) \odot x_t \tag{11}$$

$$\hat{h}_{t-1} = 2 \times \delta(W_x * \hat{x}_t) \odot h_{t-1} \tag{12}$$

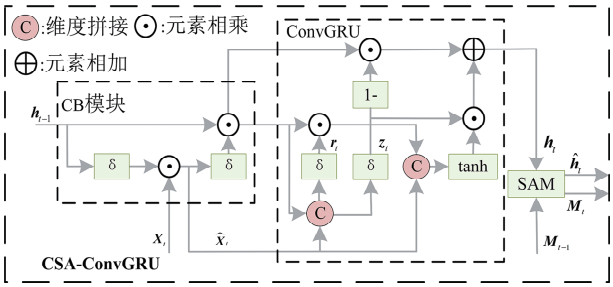


图 5　CSA-ConvGRU 模块结构图

Fig. 5　The framework of CSA-ConvGRU

ConvGRU 根据当前时刻输入、上一层的输出与自身记忆单元中的信息来获取时序特征，其框架图如图 5 所示，整体更新迭代计算过程如式(13)～(16)所示. 张量r_t和z_t分别为重置门和更新门的输出，r_t控制 CB 模块输出$\hat{h}_{t-1}$进入更新门后的记忆单元$\tilde{h}_t$中信息的数量. 更新门的输出z_t决定当前的输出状态中需要分别包含多少上一时刻的隐藏层收集的信息和当前时刻记忆的内容.

$$\hat{z}_t = \delta(W_z * [\hat{h}_{t-1}, \hat{x}_t]) \tag{13}$$

$$\hat{r}_t = \delta(W_f * [\hat{h}_{t-1}, \hat{x}_t]) \tag{14}$$

$$\tilde{h}_t = \tanh(W_h * [r_t \odot \hat{h}_{t-1}, \hat{x}_t]) \tag{15}$$

$$h_t=(1-z_t)\odot\hat{h}_{t-1}+z_t\odot\tilde{h}_t \tag{16}$$

SAM 通过引入 Attention 单元,使得输出的当前时刻的隐藏状态$\hat{h}_t$包含了当前的全局时空信息.整个单元的结构图如图 4 所示,相关公式如式(17)~(23)所示.使用两个注意力模块将两个输入分别提取出其中的重要特征后进行拼接得到聚合特征 Z.使用门控机制求得更新门门值i_t与更新值g_t,通过这两个值即能自适应的更新记忆单元M_t与最后的输出$\hat{h}_t$.

$$A_h=\mathrm{SoftMax}[(h_t^TW_{hq}^T)(W_{hk}^Th_t)] \tag{17}$$

$$A_m=\mathrm{SoftMax}[(M_{t-1}^TW_{hq}^T)(W_{mk}^TM_{t-1})] \tag{18}$$

$$Z=W_z*[(W_{hv}*h_t)A_h,(W_{mv}*M_{t-1})A_m] \tag{19}$$

$$i_t=\delta(W_{mi}*[Z,h_t]) \tag{20}$$

$$g_t=\tanh(W_{mg}*[Z,h_t]) \tag{21}$$

$$M_t=(1-i_t)\odot M_{t-1}+i_t\odot g_t \tag{22}$$

$$\hat{h}_t=\delta(W_{mo}*[Z,h_t])\odot M_t \tag{23}$$

3　实验结果与分析

为了验证本文提出的暴力检测框架的有效性,本文使用两个最新的数据集 RWF-2000[19] 和 Real Life Violence Situations(RLVS)[20] 进行多次实验.实验环境如下所示:Intel Xeon E3-1231 v3 CPU,Nvidia Geforce GTX1080 GPU,Ubuntu 20.04,PyTorch 1.9.0 架构.

3.1　实验数据集

RWF-2000 数据集包含从 YouTube 收集的 2000 个经过剪辑的监控视频.训练集包括 1600 个视频,测试集包括 400 个视频.该数据集中的视频片段全部由实际场景中的安防摄像头获取,所以视频具有场景丰富、人数众多和光照复杂等特点,是目前最具挑战性的暴力行为识别的数据集之一.

RLVS 数据集由 YouTube 收集的监控视频和人为拍摄的共计 2000 个视频经过剪辑组成.训练集包括 1600 个视频,测试集包括 400 个视频.其中的暴力行为视频是在监狱、街道和学校等环境中拍摄的,而非暴力行为视频是在运动场和竞技场等环境中拍摄的,包括游泳、射箭及打篮球等项目.

3.2　相关参数设置

预处理:首先将数据集中的视频样本进行逐帧切割,每一个视频段中包含 38 帧寸为 224×224 的帧图像.对于训练数据而言,首先将输入的视频帧的尺寸扩展到 256×256 像素大小,随后按照 50% 的概率对图像进行随机翻转,然后在图像 4 个顶角和中心的 5 个位置上随机剪裁 224×224 的像素格.最后使用 ImageNet 数据集中的均值与归一化系数进行标准化操作.

相关参数设置:从预处理完后的视频段以等差间隔的方式抽取 30 个视频帧作为网络输入.在训练过程中使用在 ImageNet 数据集上预训练的 ResNet18 模型提取视频帧图像特征,使用 Adam 优化器和交叉熵损失函数对网络进行优化,优化器中参数设置为:第一次估计的指数衰减率$\beta_1=0.5$;第二次估计的指数衰次减率$\beta_2=0.99$、权重衰减$L_2=5\times e^{-4}$.初始学习率被设置为 1×10^{-4},学习率衰减使用余弦退火衰减方法,其中余弦函数周期T_{max}设置为 64,总共迭代轮次被设置为 150.

3.3　CNN 主干网络探索实验

在 RWF-2000 数据集上,保证网络另外两个模块不变的情况下使用三种主流神经网络模型进行测试,分别为 AlexNet[21],VGG[22] 和 ResNet.实验结果如表 1 所示.

表 1　不同 CNN 网络作为主干网络的实验结果

Tab.1　Experiment results on different backbone selection

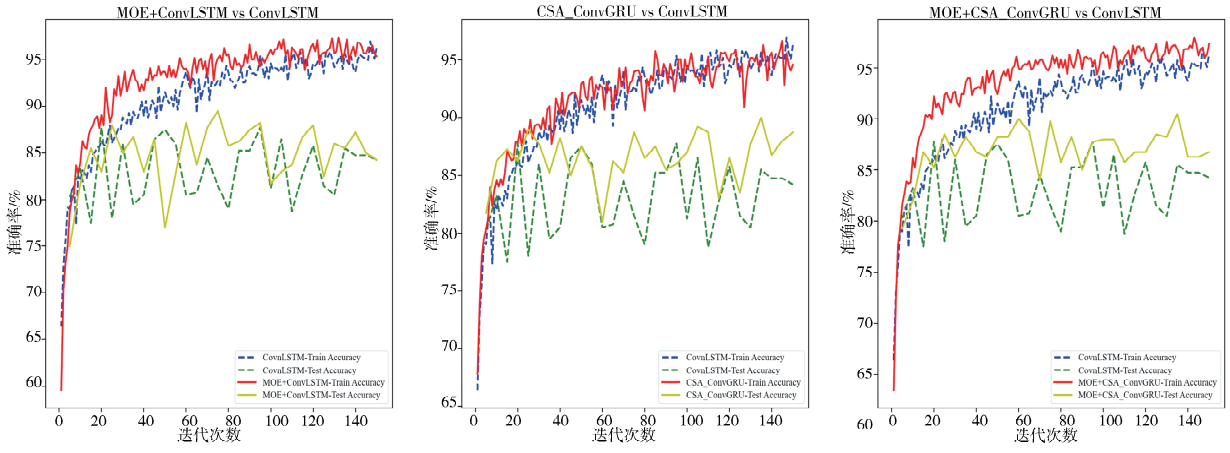
网络名称	显存占用/m	模型大小/m	准确度/%
AlexNet	947	11.44	87.00
Vgg16	1041	25.45	88.25
ResNet18	1045	21.91	90.50

通过实验结果不难发现,尽管在视频帧预处理网络中采用了 MOE 对运动边界进行了增强,但是通过分析数据集视频不难得知真实场景下的暴力冲突图像场景往往十分复杂.由于 AlexNet 深度过浅并不能有效地提取出其中的特征信息,导致识别准确度不高.而 Vgg 网络虽然有着足够的深度能够提取出深层抽象特征,但是 Vgg16 的模型大小已经增加到 25.45 M,并且模型训练时间较长收敛缓慢.考虑到模型大小和识别精度,本文采用 ResNet18 来提取帧图像中的运动特征信息.残差结构可以有效避免梯度消失,同时在保证模型参数轻量的情况下拥有足够的网络深度提取出单帧图像运动特征

3.4　消融实验

为了验证本文提出的 MOE 和 CSA-ConvGRU 结构在暴力行为检测中的有效性和泛化性,在相同的实验条件下,我们在 RWF-2000 数据集上对这两个模块的所有组合进行了消融实验.从显存

占用情况、模型大小和识别准确率三个方面进行对比，其中显存占用大小是在 batch size 为 1 的条件下检测，实验结果如表 2 所示，组合模型的训练和测试准确度的变化情况如图 6a～6c 所示.



(a)　(b)　(c)

图 6　两个模块在 RWF−2000 上的消融实验结果：(a)～(c)分别代表不同模块组合对比

Fig. 6　Results of ablation experiments on RWF-2000: (a)～(c) Represent the results of the combination of modules

表 2　两个模块在 RWF-2000 上的消融实验结果

Tab. 2　Results of ablation experiments on RWF-2000

MOE	CSA-ConvGRU	显存占用/M	模型大小/M	准确度/%
		1035	21.31	87.25
√		1037	21.31	89.50
	√	1037	21.91	90.00
√	√	1045	21.91	90.50

一方面，由于 MOE 使用了连续三帧的信息并进行了一系列的归一化和阈值化操作，所以相较于将两帧之差作为运动特征的帧差法能够包含更精确和丰富的人体运动信息，从而在不引入额外参数的情况下有效地提高识别精度. 从表 2 中可以发现 MOE 相比于简单的帧差法在模型大小和显存占用上并没有明显的增加，但是识别准确上升了 2.25%. 这也说明本文所提出的 MOE 能在不增加模型大小以及运算复杂度的基础上显著的提升识别精度. 从图 6a 和图 6c 中可以看出添加了 MOE 的网络在训练过程中收敛速度更快. 这也从侧面说明了 MOE 能够有效地提取出人体运动边界从而加快训练速度.

另一方面，添加了 CB 模块和 SAM 的 CSA-ConvGRU 结构相比于 ConvLSTM 结构能够获得更多的上下文特征信息，并且更加关注视频段中发生暴力冲突的片段，从而拥有了更多的关键且丰富的特征信息. 如表 2 所示，CSA-ConvGRU 模型大小与占用显存几乎与 ConvLSTM 持平，但是识别准确率大约提升了 2.75%. 图 6b 和图 6c 显示出新的结构在每次测试时均拥有更高的准确度. 二者都说明了 CSA-ConvGRU 结构的优越性与泛化性.

3.5　与其他主流检测方法对比

为了进一步验证本文提出的模型的性能，我们分别在公开的暴力行为数据集 RWF-2000 和 RLVS 上进行训练和测试. 数据集的稀缺极大程度阻碍了暴力检测的相关研究. 为了全面评估本文提出模型的识别能力，本节将所对比的方法分为两类，分别是最新提出的暴力行为检测方法和近年来提出的先进的通用动作识别方法. 其中，暴力行为检测方法包括 Inception-Resnet-V2＋DI[23]、Flow Gated[19]、SPIL[24]、X3D XS[25]、ECA-two cascade TSM[26]、SepConvLSTM-M[27]、SAM-GhostNet-ConvLSTM[6] 和 ViolenceNet OF[28]. 而通用动作识别方法包含 C3D[8]、P3D[9]、TEA[29]、R(2＋1)D[10] 以及 LRCN[12]，所有方法的识别准确率在表 3 进行展示.

SAM-GhostNet-ConvLSTM 是基于 LRCN 的改进模型，二者均使用了与本文模型类似的循环卷积神经网络. LRCN 方法在 RWF-2000 上的效果并不理想，主要原因是该网络没有对各帧图像进行处理而是直接将其送入网络中并仅采用简单的 LSTM 提取时序特征，故时空特征中混入了大量的干扰信息，从而导致识别的准确率较低. SAM-GhostNet-ConvLSTM 模型使用普通的帧差法对运动特征进行提取从而去除了无用的背景干扰，缓解 LRCN 中引入过多背景信息的问题. 但是由于帧差法在复杂背景下的提取效果并不令人满意，同

时 ConvLSTM 模块存在浅层语义信息丢失和无法利用上下文信息的问题,导致最后的识别准确率仍然无法令人满意. 相较于上述两种方法,本文提出的 MOE 能够加强视频帧的运动特征,使 CNN 能够更加有效地对运动主体的特征进行提取,并且 CSA-ConvGRU 单元可以更加有效地提取全局特征,可以缓解上述两种方法中存在的问题. 从结果来看,得益于 MOE 和 CSA-ConvGRU 单元有效建立不同帧间联系并自动化地对不同层级特征进行关注,所提出的模型得到了显著的效果提升,在两个公开的暴力行为检测数据集上均取得了最好的识别效果.

表 3　在两种暴力行为数据集上的识别准确率

Tab. 3　Recognition accuracies of different methods on two public violence datasets

方法	RWF-2000/%	RLVS/%
Inception-Resnet-V2+DI[23]	—	86.78
Flow Gated[19]	87.25	—
SPIL[24]	89.30	—
C3D[8]	82.75	90.50
P3D[9]	84.50	90.75
X3D XS[25]	81.75	97.50
TEA[29]	88.50	—
R(2+1)D[10]	83.25	88.50
ECA-two cascade TSM[26]	89.27	—
LRCN[12]	77.00	—
SepConvLSTM-M[27]	89.75	—
SAM-GhostNet-ConvLSTM[6]	87.50	—
ViolenceNet OF[28]	—	95.60
Ours	90.50	97.75

4　结　论

本文提出了一种时序边界注意力循环神经网络(TEAR-Net). 该方法可以有效解决运动人体特征提取不足的难题,同时还使用具有语境化和注意力机制的 CSA-ConvGRU 结构进一步增加模型对于输入信息的上下文和全局特征的提取. 所提出的 TEAR-Net 在 RWF-2000 和 RLVS 两个基于真实场景下所收集的公开数据集上的识别精度可以达到 90.50%和 97.75%. 实验结果表明,相比于目前的暴力行为识别方法,TEAR-Net 具有更高的识别率,能够适应监控场景下的暴力行为检测任务.

参考文献:

[1] Thys S, Van Ranst W, Goedemé T. Fooling automated surveillance cameras: adversarial patches to attack person detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, CA, USA: IEEE, 2019.

[2] Vuran M C, Akan Ö B, Akyildiz I F. Spatio-temporal correlation: theory and applications for wireless sensor networks [J]. Comput Netw, 2004, 45: 245.

[3] Wang H, Kläser A, Schmid C, *et al*. Dense trajectories and motion boundary descriptors for action recognition [J]. Int J Comput Vis, 2013, 103: 60.

[4] 池涛, 王洋, 陈明. 多层局部感知卷积神经网络的高光谱图像分类[J]. 四川大学学报: 自然科学版, 2020, 57: 103.

[5] 李頔, 王艳, 马宗庆, 等. 基于 DenseASPP 模型的超声图像分割[J]. 四川大学学报: 自然科学版, 2020, 57: 741.

[6] Liang Q, Li Y, Yang K, *et al*. Long-term recurrent convolutional network violent behaviour recognition with attention mechanism [C]//MATEC Web of Conferences. Sanya, China: EDP Sciences, 2021, 336: 05013.

[7] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[J]. Adv Condens Matter Phys, 2014, 27: 1.

[8] Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3d convolutional networks [C]//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015.

[9] Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3d residual networks [C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017.

[10] Tran D, Wang H, Torresani L, *et al*. A closer look at spatiotemporal convolutions for action recognition [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018.

[11] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Comput, 1997, 9: 1735.

[12] Donahue J, Anne Hendricks L, Guadarrama S, *et al*. Long-term recurrent convolutional networks for visual recognition and description [C]//Proceedings

of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015.

[13] Melis G, Koćiský T, Blunsom P. Mogrifier LSTM [EB/OL]. [2022-01-03]. https://arxiv. org/abs/1909. 01792.

[14] Cho K, van Merriënboer B, Gu¨lçehre Ç, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar: ACL, 2014.

[15] Shi X, Chen Z, Wang H, *et al*. Convolutional LSTM network: A machine learning approach for precipitation nowcasting[C]//Advances In Neural Information Processing Systems. Montréal, CANADA: MIT Press, 2015.

[16] Lin Z, Li M, Zheng Z, *et al*. Self-attention convlstm for spatiotemporal prediction[C]//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA: AAAI, 2020.

[17] He K, Zhang X, Ren S, *et al*. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, Nevada, USA: IEEE, 2016.

[18] Ulyanov D, Vedaldi A, Lempitsky V. Instance normalization: the missing ingredient for fast stylization [EB/OL]. [2022-01-03]. https://arxiv. org/abs/1607. 08022.

[19] Cheng M, Cai K, Li M. Rwf-2000: An open large scale video database for violence detection[C]//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE, 2021.

[20] Soliman M M, Kamal M H, Nashed MAEM, *et al*. Violence recognition from videos using deep learning techniques[C]//Proceedings of the 2019 Ninth International Conference on Intelligent Computing and Information Systems. New York, USA: IEEE, 2019.

[21] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Commun ACM, 2017, 60: 84.

[22] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition [EB/OL]. [2022-01-23]. https://arxiv. org/abs/1409. 1556.

[23] Jain A, Vishwakarma D K. Deep NeuralNet for violence detection using motion features from dynamic images[C]//Proceedings of the Third International Conference on Smart Systems and Inventive Technology. Tirunelveli, India: IEEE, 2020.

[24] Su Y, Lin G, Zhu J, *et al*. Human interaction learning on 3D skeleton point clouds for video violence recognition[C]//Proceedings of the European Conference on Computer Vision. Glasgow, UK: Springer, 2020.

[25] Santos F, Durães D, Marcondes F S, *et al*. Efficient violence detection using transfer learning[C]//Proceedings of the Practical Applications of Agents and Multi-Agent Systems. Salamanca, Spain: Spinger, 2021.

[26] Liang Q, Li Y, Chen B, *et al*. Violence behavior recognition of two-cascade temporal shift module with attention mechanism[J]. J Electron Imag, 2021, 30: 043009.

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

[28] Rendón-Segador F J, Álvarez-García J A, Enríquez F, *et al*. ViolenceNet: dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence[J]. Electronics, 2021, 10: 1601.

[29] Li Y, Ji B, Shi X, *et al*. Tea: Temporal excitation and aggregation for action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, Washington, USA: IEEE, 2020.

引用本文格式：

中　文：刘邦义，周激流，张卫华. 一种时序边界注意力循环暴力行为检测神经网络[J]. 四川大学学报：自然科学版，2023，60：023003.

英　文：Liu B Y, Zhou J L, Zhang W H. Temporal edge attention recurrent neural network for violence detection[J]. J Sichuan Univ: Nat Sci Ed, 2023, 60: 023003.