计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(9):211−220 [doi: 10.15888/j.cnki.csa.009229]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于视角置信度和注意力的暴力行为识别①

夏良伟, 朱　明

(中国科学技术大学 信息科学技术学院, 合肥 230026)
通信作者: 夏良伟, E-mail: xlw1998@mail.ustc.edu.cn

摘　要: 暴力行为容易出现遮挡情况, 识别准确率较低. 目前, 一些算法加入多视角视频输入来解决遮挡问题, 以等量权重将所有视角数据融合, 但是不同视角的视频因拍摄距离和遮挡情况本身就对识别存在差异性. 针对该问题, 本文提出一种基于视角置信度和注意力的暴力行为识别方法, 提高暴力识别的准确率. 本文将时序差分模块 TDM 的输入扩展成多视角, 将通道注意力机制运用在片段维度来增强 TDM 中跨段特征提取能力, 通过背景抑制方法突显移动目标的纹理特征并计算出每个视角图像的置信度, 引入双线性池化方法融合多视角视频特征, 根据视角置信度分配每个视角局部特征的权重. 本文在公开数据集 CASIA-Action 和自制数据集上进行了验证. 实验表明, 本文提出的视角置信度方法优于改进前的双线性池化方法, 暴力行为准确率相较于现有的行为识别方法取得了更好的效果.

关键词: 暴力行为识别; 注意力; 双线性池化; 视角置信度

引用格式: 夏良伟,朱明.基于视角置信度和注意力的暴力行为识别.计算机系统应用,2023,32(9):211–220. http://www.c-s-a.org.cn/1003-3254/9229.html

Violence Recognition Based on View Confidence and Attention

XIA Liang-Wei, ZHU Ming

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: Violence can be easily occluded, and the recognition accuracy is low. At present, some algorithms add multi-view video input to solve the occlusion problem and fuse all view data with equal weight. However, video from different views differs in recognition due to shooting distance and occlusion itself. To solve this problem, this study proposes a violence recognition method based on view confidence and attention to improve the accuracy of violence recognition. The input of the temporal difference module (TDM) is expanded to a multi-view angle. The channel attention mechanism is applied to the segment dimension to enhance the ability of cross-segment feature extraction in TDM. The background suppression method is used to highlight the texture features of moving objects and calculate the image confidence of each view. The bilinear pooling method is introduced to fuse multi-view video features, and the weight of local features of each view is assigned according to the view confidence. In this study, validation is performed on both the public dataset CASIA-Action and the self-made dataset. Experiments show that the view confidence method proposed in this study is better than the bilinear pooling method before improvement, and the accuracy of violence recognition is better than that of the existing behavior recognition methods.

Key words: violence recognition; attention; bilinear pooling; view confidence

1　引言

随着人工智能和深度学习的快速发展, 人们可以从监控视频中以智能化的方式获取各种数据信息, 对视频里出现的行为自动识别、理解和分析, 以可视化

① 基金项目: 科技创新特区计划 (20-163-14-LZ-001-004-01)
收稿时间: 2023-02-20; 修改时间: 2023-03-20; 采用时间: 2023-04-07; csa 在线出版时间: 2023-07-17
CNKI 网络首发时间: 2023-07-18

的结果呈现出来, 并在设置的条件内有效报警, 实现准确高效的监管. 因此基于深度学习的行为识别技术在智能视频监控领域中至关重要. 在各类异常行为中, 暴力行为作为一种恶劣且复杂的群体行为, 涉及个体之间的肢体接触, 人员之间距离更近, 遮挡情况也会更严重, 识别难度相比其他行为更高. 目前根据算法输入的不同, 可以分为以下 3 种方法.

(1) 单视角方法. Tran 等[1] 提出一种新的 3D 卷积神经网络 (convolutional 3D, C3D), 将二维卷积核扩展到时间轴, 不但可以提取单帧视频的空间特征还能够提取连续帧之间的时序特征. Ding 等[2] 最早使用 3D 卷积网络直接从原始输入中识别暴力. 为减小无人帧带来的计算代价, Ullah 等[3] 先通过 MobileNet 去除无人帧, 再通过 3D 卷积只对含有人员的帧提取特征. 但是 3D 卷积计算量和参数量过大, 且在时间维度上的卷积感受野有限, 只适合提取短暂时间的暴力特征. Simonyan 等[4] 设计了双流网络, 使用二维卷积提取视频帧与堆叠光流图中的空间和时间特征, 但是密集采样出现了较多的冗余帧. Wang 等[5] 改进了二维的双流网络, 提出了时序分段网络 (temporal segment networks, TSN), 使用稀疏采样代替密集采样, 降低无用信息冗余. Lin 等[6] 提出的将部分通道沿着时间轴移位来提取帧间信息的时序位移模块 (temporal shift module, TSM), 能即插即用到 CNN 中的时态建模模块来降低计算量. Cheng 等[7] 基于伪 3D 卷积思想分别从暴力片段中提取外观特征和运动特征, 减小计算量. Dong 等[8] 在双流网络的基础上添加了具有空间和时间信息的加速流, 并用 LSTM 建模长时间信息, 用于检测人与人之间的暴力, 但是多流网络的计算代价过大. FAIR[9] 提出了 SlowFast 网络, 设计了慢快路径结合, 并在 Kinetics-400 和 ava 数据集上得到很高的精度. 基于双流网络中光流提取成本较大的问题, Islam 等[10] 提出背景抑制后的 rgb 流与帧差流相结合的双流暴力识别网络, 每个流的分支使用二维 CNN 与深度可分离卷积 LSTM 相结合提取特征. Wang 等[11] 在 TSN 的基础上提出了时序差分模块 (temporal difference module, TDM), 同样使用帧差流代替光流减少计算量, 提取短期时间和长期时间两种特征来增强段间和跨段运动变化信息. 这类方法特点是输入比较单一, 当遮挡非常严重时, 信息的缺失导致识别效果有限.

(2) 多模态方法. 多模态就是在 RGB 视频输入的基础上添加一些扩充信息, 如深度信息、骨骼信息等等来解决遮挡问题, 提升识别准确率. Singh 等[12] 通过姿态估计网络提取每个人的骨骼关键点数据后用于暴力行为识别, 但是提取每个人的骨骼数据计算量较大, 且 2D 数据无法解决遮挡问题. Yan 等[13] 将 3D 骨骼关键点数据与 RGB 视频同时作为网络的输入, 以弥补空间尺度信息和遮挡人员特征信息的缺失. 这类研究目前已经取得了非常好的效果, 然而想要获取完整的 3D 人体骨骼关键点的三维坐标并非易事. 一般通过价格高昂的传感器捕获或通过深度相机或激光雷达的深度数据进行伪 3D 骨骼点估计, 两种方法因为高昂的成本很难在实际场景中应用. Peixoto 等[14] 在暴力识别算法中引入音频信息来提升精度, 但是音频数据本身干扰很大, 也不能解决视频中出现的遮挡问题.

(3) 多视角方法. 多视角方法只需要多个 RGB 摄像头的输入数据, 相较于上述多模态数据, 多视角数据不需要更多的预处理, 成本较低. 并且随着智能视频监控系统硬件体系的整体提升, 对公共区域的多视角监控成为可能, 能实现对监控区域的全方位无死角覆盖. Su 等[15] 首次使用多相机系统从若干不同的角度拍摄三维物体, 得到多视角的二维图像, 提出多视角卷积神经网络 (multi-view CNN, MVCNN), 使用全局最大池化算法将各个视角提取的特征融合起来, 该算法没有考虑到不同视图之间的相似度关系. Feng 等[16] 在文献 [15] 的基础上提出分组视角卷积网络 (group-view CNN, GVCNN), 先将不同视角下 CNN 提取的特征根据其相似程度进行分组, 再进行组内最大池化和组间特征融合, 但该算法依旧使用了最大池化, 忽略了非最大元素, 导致信息的丢失. Hou 等[17] 提出多视角检测网络, 利用外参将 CNN 中的特征图映射到鸟瞰图上进行融合, 但是该方法需要多视角视频之间外部参数的先验信息, 在实际场景中较难获取. Yu 等[18] 提出使用双线性池化方法融合多视角的局部卷积特征, 生成一个紧凑的全局描述符. Gao 等[19] 提出基于群稀疏性和图模型的多视角判别结构化字典模型用于行为识别. Xia 等[20] 提出一种基于局部自相似描述符和图共享多任务学习的多视角交互行为识别方法. 但是上述算法将所有视角特征等量融合, 忽略了不同视角的视频因拍摄角度和拍摄距离本身就对识别精度存在差异性的问题.

基于此, 本文的贡献点如下: (1) 提出了一种基于注意力和视角置信度的暴力行为识别方法, 将通道注

意力机制运用在片段维度来增强 TDM 中跨段特征提取能力, 使用改进后的跨段注意力 TDM 对多视角视频进行特征提取. (2) 通过背景抑制方法来突显移动目标的纹理特征并根据移动目标的区域大小计算出该视角图像的置信度. (3) 引入双线性池化方法融合多视角视频特征, 并根据视角置信度得分分配每个视角局部特征的权重. 实验表明, 本文提出的置信度得分方法优于改进前的双线性池化方法, 暴力行为准确率相较于现有的行为识别方法取得了更好的效果.

2 相关工作

2.1 时序差分模块 TDM

时序差分模块 TDM[11] 分为短期 TDM 和长期 TDM 两个步骤. 先采用分段采样的方法将视频分为k个片段, 短期 TDM 用于提取片段内的局部运动信息, 长期 TDM 用于提取跨段的长期运动信息.

短期 TDM 是 RGB 流与帧差流相结合的双流网络, 结构如图 1 所示, 其中一条流提取片段内关键帧的 RGB 图像的空间特征, 另一条流提取关键帧前后共 $2m$ 个帧差图的时间特征, 最后融合两条分支的信息得到片段内 $2m$+1 帧连续图像的段内局部运动信息. 数学模型为:

$$M_k = CNN(F_k) + S(D_k) \tag{1}$$

其中, M_k 为第k个片段通过短期 TDM 的输出, F_k和 D_k分别是第k个片段的关键帧和关键帧附近的$2m$帧帧差图, CNN 和 S 分别是 RGB 流和帧差流的模型.

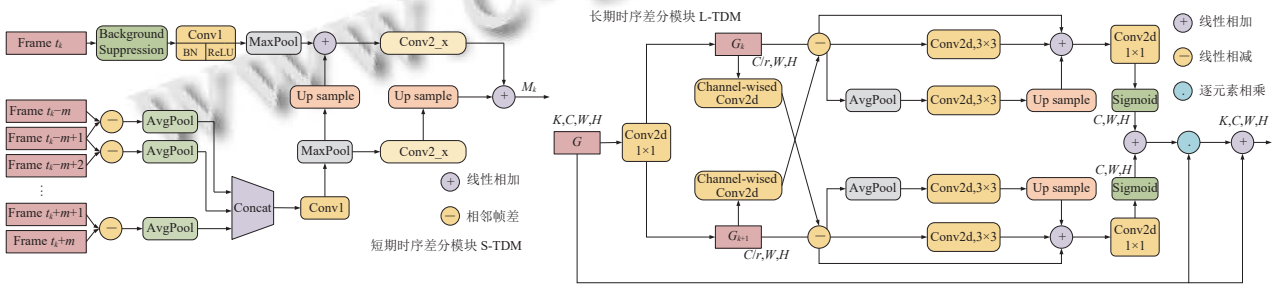


图 1 时序差分模块 TDM

长期 TDM 融合相邻段的特征信息, 实现跨段的时序特征提取, 结构如图 1 所示. 首先将每个片段的通道以r的比率压缩, 接着将每个片段的信息经过通道卷积跳跃连接到相邻的片段中, 实现跨片段之间的全局信息建模. 最后经过三分支结构学习多尺度特征, 上采样到原来的通道数与原特征进行残差连接. 数学模型为:

$$R_k = G_k + G_k \odot L(G_k, G_{k+1}) \tag{2}$$

其中, R_k为第k个片段通过长期 TDM 的输出, G_k和 G_{k+1}分别是第k和第$k+1$个片段的输入特征, L 是长期 TDM 残差模块主分支的模型, $\odot$是逐元素相乘.

2.2 双线性池化方法

设视频 A 和视频 B 有V个视角, 每个视角特征的尺寸是d. 视频 A 和视频 B 的所有视角特征集合分别为 $X_A = [X_A^1, X_A^2, \cdots, X_A^V]$和$X_B = [X_B^1, X_B^2, \cdots, X_B^V]$. 通常通过相关性来评估融合特征的鉴别能力, 定义视频 A 和视频 B 的相关性为:

$$sim(A,B) = \sum_{x\in X_A}\sum_{y\in X_B} < x,y>^2 \tag{3}$$

A 和 B 的相关性为 A 和 B 中所有视角特征两两匹配对的内积平方和. 两个向量之间内积越大, 说明两个视角越匹配, 其中一个$< x,y >$能看成该匹配对的权重, 也会比较大. 因此该相似度可以累加匹配的视角特征对, 抑制非匹配视角的影响. 式 (3) 又可以写成:

$$\begin{aligned} sim(A,B) &= \sum_{x\in X_A}\sum_{y\in X_B} < vec(xx^{\mathrm{T}}), vec(yy^{\mathrm{T}}) > \\ &=< vec\left(\sum_{x\in X_A} xx^{\mathrm{T}}\right), vec\left(\sum_{y\in X_B} yy^{\mathrm{T}}\right) > \end{aligned} \tag{4}$$

其中, $vec\left(\sum_x xx^{\mathrm{T}}\right)$为视频的双线性池化特征[18].

也就是说, 两组视频视角特征集合的相关性等于他们的双线性池化特征的内积. 双线性池化特征可以反映不同视角特征之间两两匹配的结果. 双线性池化特征是先将相同位置上的每一组特征与其自身线性相乘后, 再对所有位置上的线性相乘结果进行求和池化, 能够提取原有特征的二阶协方差统计信息, 比一阶均值信息具有更丰富的特征表达能力, 在多模态特征融合等任务中表现较好.

3 算法设计

本文基于上述研究提出了一种端到端的基于视角置信度和注意力的暴力行为识别方法. 结构如图 2 所示. 图中设定输入为 4 个视角的视频流, 输出为判断是否出现了暴力行为. 先对多视角视频进行稀疏采样, 并且均匀分成 K 个片段, 再从每个片段中随机采样. 采样后的多视角视频帧通过短期 TDM 提取片段内局部特征. 短期 TDM 输入使用全局背景抑制后的帧代替原始 RGB 视频帧来突出帧中移动目标, 并基于分段背景抑制方法得到每个片段内关键帧移动目标的区域大小, 从而计算出每个视角的置信度. 然后通过加入跨段注意力机制的长期 TDM 提取多尺度的跨段全局特征, 长期 TDM 提取特征时, 相邻片段段的特征信息会跳跃连接进来. 每个视角共享特征提取网络的参数, 然后通过基于视角置信度的双线性池化模块融合多个视角的视频特征, 最后通过全连接层分类后对所有片段的分类结果取均值, 就得到整个视频流的分类结果. 分类结束后, 通过 Grad-CAM[21] 对结果进行热力图可视化.

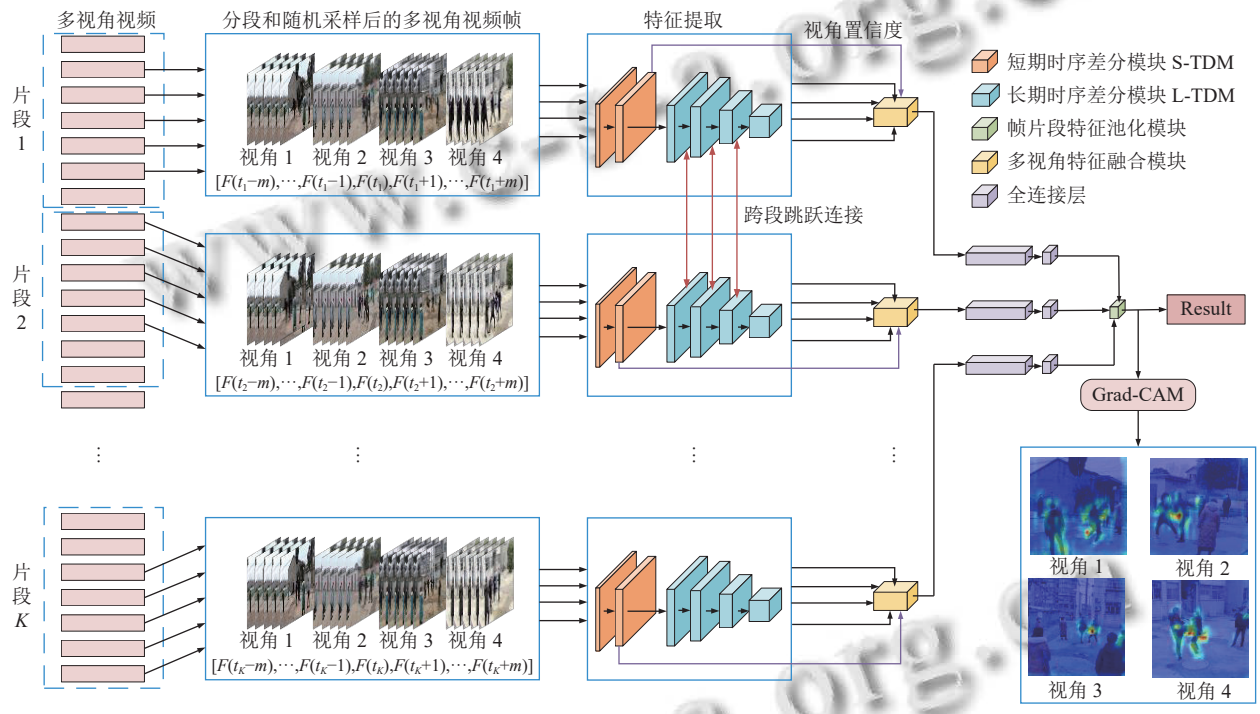


图 2 网络结构

3.1 稀疏分段随机采样

多视角视频经过稀疏采样得到多视角视频帧. 稀疏采样相比密集采样, 能得到更多有效帧. 每一帧由 V 个视角的图像组合而来. 再将多视角视频流分成等长且互不重叠的 K 个片段, 从每个片段中随机选取 $2m+1$ 张连续视频帧作为该视频片段的输入, 连续帧最中间的帧作为该片段的关键帧. 将 K 个视频片段的输入组合起来是整个网络的输入 I, 尺寸为 $K\times(2m+1)\times V\times C\times W\times H$, 有 6 个维度, 分别是视频片段数、段内采样帧数、视角数、通道、宽度和高度.

3.2 背景抑制方法

本文短期 TDM 中的关键帧输入不直接使用原始 RGB 图像帧, 而是选择背景抑制之后的帧. 由于暴力行为的特征大部分都是变化的肢体运动特征, 而不是固定不变的背景特征, 这促使模型更加关注变化的运动信息. 因此可以通过背景抑制方法来突显帧中的移动目标的纹理特征.

首先计算整个视频里所有帧的平均值, 平均帧主要包含背景信息, 它们在所有帧中固定不变. 再将每一个关键帧都减去这个平均值就得到了背景抑制之后的帧. 因为背景部分的像素值保持不变, 相减之后为 0, 剩下的就是帧中变化的像素点, 对应运动的前景目标.

$$avg=\frac{1}{N}\sum_{i=1}^{N}F_i \tag{5}$$

$$B_i = |F_i - avg| \tag{6}$$

其中, avg为视频里所有帧的平均值, B_i为关键帧F_i背景抑制之后的结果, N为整个视频采样的总帧数.

3.3 基于分段背景抑制方法的视角置信度计算

不同视角的视频因拍摄距离以及遮挡情况本身就对识别存在差异性. 拍摄距离较远时, 运动目标在图像中较小. 遮挡情况导致运动目标区域被静止目标掩盖. 上述情况能通过运动目标区域大小占整张图像的比值来衡量. 本文定义视频片段k内视角v的置信度为:

$$\lambda_k^v = \frac{S_{\text{motion}}}{S_{\text{image}}} \tag{7}$$

其中, S_{motion}为目标运动区域的面积, S_{image}为图像的总面积.

视角置信度计算过程如图 3 所示, 与短期 TDM 中的关键帧输入在背景抑制时使用整个视频的平均帧不同, 这里首先计算片段内的平均帧与关键帧的差值图. 因为片段内帧相比整个视频的帧, 间隔时间短, 背景抑制后的干扰信息较少, 方便后续的计算. 背景抑制后的帧依然有 3 个通道, 接着对差值图进行灰度化降成 1 个通道, 再进行二值化区分背景信息和运动信息, 然后经过腐蚀的形态学运算对背景中的小颗粒噪声进行消除, 最后经过膨胀对运动边界进行平滑, 腐蚀和膨胀是分别在滤波核中求局部最小值和最大值的操作, 滤波核尺寸是 3, 得到的二值图中白色像素点的个数就是S_{motion}. 白色像素点描述的是运动的轮廓, 轮廓的厚度表示运动的剧烈程度.

图 3 视角置信度计算过程

对片段内V个视角的关键帧分别计算置信度, 接下来需要对其进行归一化:

$$\omega_k^v = \frac{\lambda_k^v V}{\sum_{i=1}^{V} \lambda_k^i} \tag{8}$$

其中, ω_k^v为视角v归一化后的置信度, 这样保证V个视角归一化后的置信度相加等于V. 在后续特征融合作为每个视角的权重时, 不会对数据的尺度产生影响.

一组多视角关键帧计算的置信度如图 4 所示, 红色矩形框表示白色像素点的密集区域, 即运动目标的区域. 视角 1 中的暴力行为存在遮挡问题, 置信度最低. 视角 3 中的拍摄距离较远, 置信度也相对较低. 视角 2 和视角 4 的特征较为明显, 因此置信度较高. 本文对置信度设置了一个阈值threshold, 若某个视角归一化后的置信度$\omega_k^v < \text{threshold}$, 则不对该视角视频进行特征提取和融合, 以减小计算量.

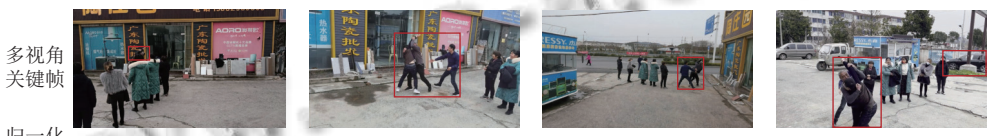


图 4 多视角视频帧的置信度

3.4 基于跨段注意力的时序差分特征提取网络

TDM 将 ResNet 后 3 个阶段的每个 Bottleneck 残差模块中添加了长期 TDM 来提取多尺度的跨段全局信息. 本文将通道注意力机制运用在片段维度, 在每个长期 TDM 的后面加入了跨段注意力模块, 进一步增强段与段之间的联系和全局特征提取能力. 网络结构图如图 5 所示.

首先对空间特征进行平均池化以减小计算量, 接着通过 卷积将通道压缩到原来的 1/16, 实现相同片段不同通道的信息交互和融合. 然后将片段维度调整到最后一维, 并对其进行一维卷积, 增强相同通道内跨片段信息的联系. 最后调整维度并上采样到原来的通道数形成跨段的注意力图, 并融合原特征和增强后的特征, 以逐元素相乘和相加的方式进行残差连接.

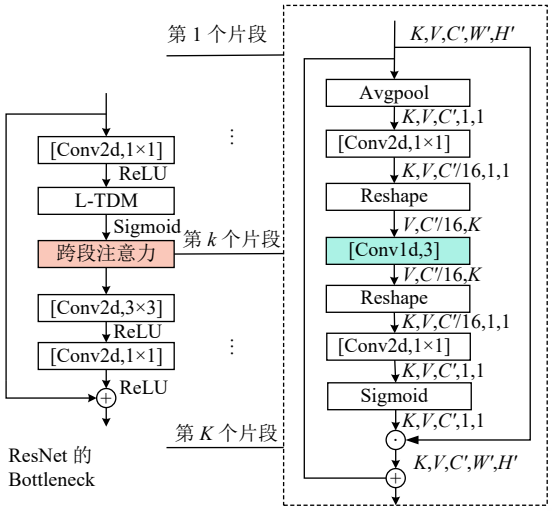


图 5 跨段注意力模块

3.5 基于视角置信度的双线性池化方法

本文基于第 3.3 节中计算的视角置信度, 提出了基于视角置信度的双线性池化方法. 差异较大的视图中可能存在某块局部区域的特征是匹配的, 式 (3) 里基于全局视角图像的匹配方式会忽略该相关信息, 如图 6 所示. 基于此, 本文将多视角特征的视角个数V、宽度W和高度H合并成一个维度N, 并通过1×1卷积将通道数C削减为d减小计算量. 双线性池化让合并维度的所有局部特征块都可以进行匹配并累加, 相比全局视角的匹配能提取更精细的局部区域匹配信息, 从而根据式 (4) 得到两组视频之间的相关性.

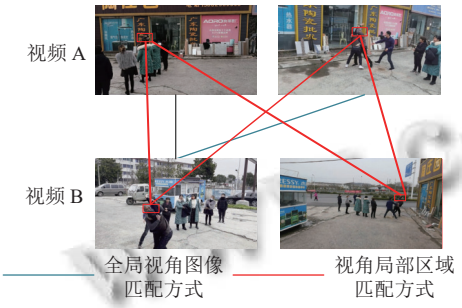


图 6 全局图像和局部区域匹配方式的区别

设第k个视频片段内维度合并和 1×1 卷积后的特征为X_k, 尺寸为$[d,N]$, 其中$N=V\times W\times H$. X_k包含所有视角的N组局部特征块, 其双线性池化特征为:

$$Y_k=\sum_l x_k^l(x_k^l)^{\mathrm{T}}=X_kX_k^{\mathrm{T}} \tag{9}$$

其中, x_k^l是X_k的一组局部特征, 尺寸为$d\times1$. Y_k的尺寸为$d\times d$.

本文在双线性池化中加入了归一化后的视角置信度作为权重, 改进后的双线性池化特征为:

$$Z_k=\sum_l \omega_k^l x_k^l(x_k^l)^{\mathrm{T}} \tag{10}$$

其中, ω_k^l是局部特征x_k^l所属视角的置信度, 作为权重来突出高价值的视角特征. 式 (10) 可以简化为使用X_k'代替X_k进行双线性池化, 如式 (11) 和式 (12) 所示:

$$X_k'=W_k\odot X_k \tag{11}$$

$$W_k=Concat\left(\sqrt{\omega_k^1}A,\sqrt{\omega_k^2}A,\cdots,\sqrt{\omega_k^V}A\right) \tag{12}$$

其中, $\odot$是逐元素相乘, $Concat$ 是拼接操作, A是尺寸为$[d,W\times H]$, 所有元素值为 1 的矩阵, ω_k^v为视角v归一化后的置信度, 拼接后W_k尺寸为$[d,V\times W\times H]$.

直接对X_k'进行双线性池化与式 (10) 等价:

$$Z_k=X_k'X_k'^{\mathrm{T}} \tag{13}$$

为了简化双线性池化的计算量, 可以先将X_k'进行 SVD 分解:

$$X_k'=U_k\Sigma_kV_k^{\mathrm{T}}=\sum_{i=1}^{d}\sigma_k^iu_k^i(v_k^i)^{\mathrm{T}} \tag{14}$$

其中, U_k和V_k^{T}是正交矩阵, 尺寸分别为$d\times d$和$N\times N$, Σ_k是对角矩阵, 尺寸为$d\times N$. $\{\sigma_k^i\}_{i=1}^d$是X_k'的奇异值, 而$\{u_k^i\}_{i=1}^d$和$\{(v_k^i)^{\mathrm{T}}\}_{i=1}^d$分别是对应的左右奇异向量.

将式 (14) 代入式 (13) 可得:

$$Z_k=\left(\sum_{i=1}^{d}\sigma_k^iu_k^i(v_k^i)^{\mathrm{T}}\right)\left(\sum_{j=1}^{d}\sigma_k^ju_k^j(v_k^j)^{\mathrm{T}}\right)^{\mathrm{T}}=\sum_{i=1}^{d}(\sigma_k^i)^2u_k^i(u_k^i)^{\mathrm{T}} \tag{15}$$

其中, $d<<N$的关系让计算量得到简化.

基于视角置信度的双线性池化方法结构如图 7 所示. 在双线性池化前后进行矩归一化来进一步减小特征的方差, 接着通过 L2 范数归一化消除奇异特征数据导致的不良影响, 加快梯度下降的收敛速度. 最后得到长度为d^2的特征向量送入全连接层分类.

3.6 损失函数

本文使用的损失函数参考 TSN[5] 中的分段交叉熵损失函数:

$$L(y,G)=-\sum_{i=1}^{C}y_i\left(G_i-\log\sum_{j=1}^{C}\mathrm{e}^{G_j}\right) \tag{16}$$

其中, C是类别数, 暴力识别是二分类网络, 因此C的值为 2. y_i是暴力和非暴力的真实标签, G_i为K个片段分类得分的平均值, 即:

$$G_i=\frac{\sum_{k=1}^{K}F_i(I_k)}{K} \tag{17}$$

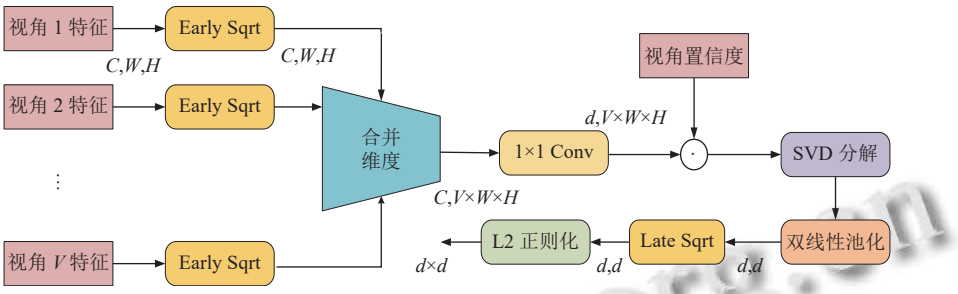


图 7 基于视角置信度的双线性池化方法结构图

4 实验分析

4.1 数据集

本文分别在公开数据集 CASIA-Action[22] 和自制数据集上进行了算法的验证实验. 两个数据集都是多视角的行为识别数据集, 其中 CASIA-Action 包含 3 个视角, 自制数据集包含 4 个视角.

CASIA-Action 数据集包含了 422 组视频数据, 每组视频由室外环境下 3 个不同视角的摄像机拍摄而成. 数据集包含了走路、跳跃等单人行为和抢劫、打斗、会合等交互行为. 本文只区分其中的暴力行为和非暴力行为作为暴力行为数据集. 视频的帧率是 25 FPS, 分辨率是320×240.

考虑到 CASIA-Action 数据集中的暴力样本和视角个数较少以及视频中人员数量较少, 并且目前没有复杂场景下的多视角暴力行为数据集. 本文使用 4 个1920×1080分辨率的 RGB 摄像头在不同的角度对同一块区域同时进行拍摄, 自制了一个多视角暴力数据集. 整个数据集拍摄了 10 人左右模拟的一些行为动作. 为了验证算法的有效性, 视频中除了包含拳打、脚踢、扭打、摔跤等暴力行为以外, 还有一些容易对暴力行为产生干扰的奔跑、拥抱、跳绳等非暴力行为. 4 个摄像头在 5 个不同户外场景中采集了时长 90 min 的视频, 经过剪辑选取了 600 组有明显暴力和非暴力行为区分的视频数据, 每组视频数据都包含 4 个摄像头在同一时刻以不同角度拍摄的 4 段视频, 合计 2 400 段视频. 将其中 400 组视频作为训练集, 40 组视频作为验证集, 160 组视频作为测试集.

4.2 实验过程

实验所用的机器配置如表 1 所示.

表 1 训练所用机器配置表

类型	型号	参数
系统	Ubuntu	16.04.2 LTS
CPU	Intel(R) Core(TM) i9-9900	8核
GPU	Nvidia GeForce GTX 1080Ti	11 GB
内存	DDR4	32 GB

实验中特征提取网络选择 ResNet50, 特征提取网络的参数使用 ImageNet 数据集上的预训练模型初始化. 关键帧的前后帧数量m设置为 2, 一组视频拆分的视频片段数量K设置为 8, 训练的 batch size 设置为 8, 优化器是 SGD, 初始学习率设置为 0.001, 置信度阈值 threshold 设置为 0.2.

为了提高暴力行为识别的鲁棒性, 本文训练时, 首先将视频帧的较短边长度随机调整到[256,320]区间内, 并且保持较长边与较短边的比例不变, 再随机裁剪, 得到224×224的输入数据.

本文评价指标是准确率 *Accuracy* (如式 (18) 所示)、参数量和 GFLOPS.

$$Accuracy=\frac{TP+TN}{TP+FP+TM+FN} \tag{18}$$

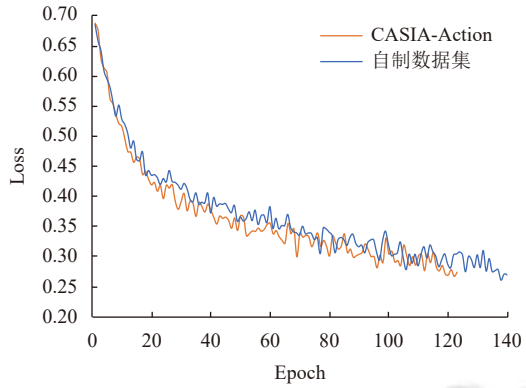
其中, *TP* 是真阳性, *TN* 是真阴性, *FN* 是假阴性, *FP* 是假阳性.

参数量是指模型要训练的参数总数. GFLOPS 是每秒 1G (10 亿) 次的浮点运算次数, 衡量模型的计算量.

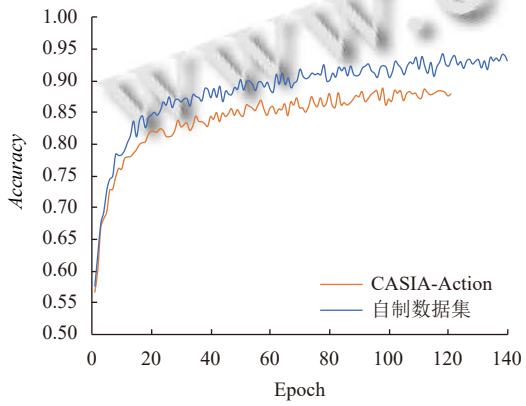
4.3 与现有方法的对比实验

为了验证本文基于视角置信度的双线性池化方法的有效性, 本文在 CASIA-Action 和自制数据集上分别融合三视角和四视角的视频特征, loss 和 *Accuracy* 曲线如图 8 所示, 并与其他的多视角特征融合方法进行

了对比, 实验结果如表 2 所示, 可以发现, 本文提出的基于视角置信度的双线性池化方法优于改进前的双线性池化方法, 也优于其他多视角特征融合方法.



(a) Loss 变化曲线



(b) *Accuracy* 变化曲线

图 8 Loss 和 *Accuracy* 变化曲线

表 2 与其他多视角特征融合方法的对比 (%)

多视角特征融合方法	*Accuracy*	
	CASIA-Action	自制数据集
拼接	83.53	88.13
MVCNN[15]	84.21	90.63
GVCNN[16]	84.21	91.25
双线性池化[18]	86.18	93.75
本文方法	**88.23**	**95.00**

为了验证本文暴力行为识别方法的有效性, 本文在 CASIA-Action 和自制数据集上与单视角和多视角行为识别算法进行了对比实验. 对于单视角的行为识别算法, 本文将 CASIA-Action 和自制数据集中的每一组多视角视频数据分别拆分成 3 组和 4 组单视角数据集作为网络的输入, 再计算所有视频片段的准确率. 对于多视角的行为识别算法, CASIA-Action 和自制数据集上的网络输入分别使用三视角和四视角视频帧. 实验结果如表 3 所示. 本文方法对比其他行为识别算法, 准确率更高, 而多视角视频特征提取共用 TDM 的参数, 因此参数量依旧较小, 计算量取决于视角的个数.

表 3 与其他行为识别网络的对比

方法	GFLOPS	参数量 (M)	*Accuracy* (%)	
			CASIA-Action	自制数据集
TSN[5]	33	24.3	71.04	74.06
SlowFast[9]	65.7	34.4	81.96	82.50
ConvLstm[10]	14.4	9.6	80.39	84.84
TSM[6]	33	24.3	77.64	83.69
TDM[11]	36	26.2	82.35	84.69
Xia等*[20]	24.4×V	—	86.4	—
GAO等*[19]	37.2×V	—	82.76	89.38
本文方法	**36.6×V**	**26.5**	**88.23**	**95.00**

注: *为多视角网络, V为视角的个数

4.4 消融实验

为了验证本文改进方法的有效性, 本文做了消融实验, 在基于双线性池化融合方法的多视角 TDM 网络的基础上, 将加入跨段注意力模块设为改进 1, 将视角置信度方法与双线性池化方法结合设为改进 2. 实验结果如表 4 所示. 可以发现, 两种改进均提高了暴力识别的准确率, 证明了本文方法的有效性.

表 4 注意力和视角置信度的影响 (%)

方法	*Accuracy*	
	CASIA-Action	自制数据集
TDM	82.35	84.69
TDM+双线性池化	86.18	93.75
添加改进1	86.84	94.38
添加改进1+改进2	88.23	95.00

本文将特征提取网络 ResNet50 替换 MobileNetV2 实现轻量化, 实验结果如表 5 所示, 在略微降低精度的情况下, 参数量和计算量大量较少. 在实际运用中, 可以考虑该轻量化模型, 来实现实时的监控.

表 5 特征提取网络的对比

特征提取网络	GFLOPS	参数量 (M)	*Accuracy* (%)	
			CASIA-Action	自制数据集
ResNet50	36.6×V	26.5	88.23	95.00
MobileNetV2	26.4×V	5.8	85.05	93.13

4.5 可视化

本文使用 Grad-CAM 技术[21] 可视化多视角 TDM

的特征表示. Grad-CAM 基于梯度定位创建类激活映射, 生成卷积层特征的热力图来说明深度学习模型的可解释性和透明性. 本文对 S-TDM 模块的 conv2_x 层的特征进行可视化, 计算卷积层中每个通道的特征图对识别类别的权重, 并求每个特征图的加权和, 最后将加权特征图映射到原图中. 以片段内所有帧作为输入, 并只在关键帧中绘制激活映射, 一组四视角关键帧的热力图如图 9 所示. 可以看到, 图像中的两个人员出现了暴力行为, 热力图在 4 个视角图像中都能清晰的定位到暴力特征的区域, 并且精细到人员暴力接触的肢体, 验证了本文方法的有效性.

(a) 原图 1 (b) 热力图 1 (c) 原图 2 (d) 热力图 2

图 9 Grad-CAM 热力图

5 结论与展望

本文提出了一种视角置信度和注意力的暴力行为识别方法. 通过背景抑制方法来突显移动目标的纹理特征并计算出每个视角的置信度, 将通道注意力机制运用在片段维度来增强 TDM 中跨段特征提取能力, 引入双线性池化网络融合多视角视频特征, 并根据视角置信度分配每个视角局部特征的权重. 相较于现有的多视角特征融合方法, 取得了更好的效果. 下一步研究工作将着重于将多视角暴力行为识别算法应用到智能视频监控系统中, 实现准确高效的监管.

参考文献

1 Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 4489–4497. [doi: 10.1109/ICCV.2015.510]

2 Ding CH, Fan SK, Zhu M, *et al*. Violence detection in video by using 3D convolutional neural networks. Proceedings of the 10th International Symposium on Advances in Visual Computing. Las Vega: Springer, 2014. 551–558.

3 Ullah FUM, Ullah A, Muhammad K, *et al*. Violence detection using spatiotemporal features with 3D convolutional neural network. Sensors, 2019, 19(11): 2472. [doi: 10.3390/s19112472]

4 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2014. 568–576.

5 Wang LM, Xiong YJ, Wang Z, *et al*. Temporal segment networks: Towards good practices for deep action recognition. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 20–36.

6 Lin J, Gan C, Han S. TSM: Temporal shift module for efficient video understanding. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 7082–7092. [doi: 10.1109/ICCV.2019.00718]

7 Cheng M, Cai KJ, Li M. RWF-2000: An open large scale video database for violence detection. Proceedings of the 25th International Conference on Pattern Recognition (ICPR). Milan: IEEE, 2021. 4183–4190.

8 Dong ZH, Qin J, Wang YH. Multi-stream deep networks for person to person violence detection in videos. Proceedings of the 7th Chinese Conference on Pattern Recognition. Chengdu: Springer, 2016. 517–531.

9 Feichtenhofer C, Fan HQ, Malik J, *et al*. SlowFast networks for video recognition. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 6201–6210. [doi: 10.1109/ICCV.2019.00630]

10 Islam Z, Rukonuzzaman M, Ahmed R, *et al*. Efficient two-stream network for violence detection using separable convolutional LSTM. Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN). Shenzhen: IEEE, 2021. 1–8. [doi: 10.1109/IJCNN52387.2021.9534280]

11 Wang LM, Tong Z, Ji B, *et al*. TDN: Temporal difference networks for efficient action recognition. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021. 1895–1904. [doi: 10.1109/CVPR46437.2021.00193]

12 Singh A, Patil D, Omkar SN. Eye in the sky: Real-time drone surveillance system (DSS) for violent individuals identification using scatternet hybrid deep learning network. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City: IEEE, 2018. 1710–1718. [doi: 10.1109/CVPRW.2018.00214]

13 Yan SJ, Xiong YJ, Lin DH. Spatial temporal graph convolutional networks for skeleton-based action recognition. Proceedings of the 32nd AAAI Conference on Artificial Intelligence, the 30th Innovative Applications of Artificial Intelligence Conference and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence. New Orleans: AAAI, 2018. 912.

14 Peixoto B, Lavi B, Bestagini P, *et al*. Multimodal violence detection in videos. Proceedings of the 2020 ICASSP IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 2957–2961. [doi: 10.1109/ICASSP40776.2020.9054018]

15 Su H, Maji S, Kalogerakis E, *et al*. Multi-view convolutional neural networks for 3D shape recognition. Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015. 945–953. [doi: 10.1109/ICCV.2015.114]

16 Feng YF, Zhang ZZ, Zhao XB, *et al*. GVCNN: Group-view convolutional neural networks for 3D shape recognition. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 264–272. [doi: 10.1109/CVPR.2018.00035]

17 Hou YZ, Zheng L, Gould S. Multiview detection with feature perspective transformation. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 1–18.

18 Yu T, Meng JJ, Yuan JS. Multi-view harmonized bilinear network for 3D object recognition. Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 186–194. [doi: 10.1109/CVPR.2018.00027]

19 Gao Z, Zhang H, Xu GP, *et al*. Multi-view discriminative and structured dictionary learning with group sparsity for human action recognition. Signal Processing, 2015, 112: 83–97. [doi: 10.1016/j.sigpro.2014.08.034]

20 Xia LM, Guo WT, Wang H. Interaction behavior recognition from multiple views. Journal of Central South University, 2020, 27(1): 101–113. [doi: 10.1007/s11771-020-4281-6]

21 Selvaraju RR, Cogswell M, Das A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of the 2017 IEEE International Conference on Computer Vision. 2017. 618–626. [doi: 10.1109/ICCV.2017.74]

22 Zhang Z, Huang KQ, Tan TN. Multi-thread parsing for recognizing complex events in videos. Proceedings of the 10th European Conference on Computer Vision. Marseille: Springer, 2008. 738–751. [doi: 10.1007/978-3-540-88690-7_55]

(校对责编: 牛欣悦)