计算机科学
COMPUTER SCIENCE

基于时空图卷积和注意力模型的航拍暴力行为识别

邵延华，李文峰，张晓强，楚红雨，饶云波，陈璐

计算机科学, 2022, 49(6): 313-318. https://doi.org/10.11896/jsjkx.210400101

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210400272

基于时空图卷积和注意力模型的航拍暴力行为识别

邵延华[1] 李文峰[1] 张晓强[1] 楚红雨[1] 饶云波[2] 陈 璐[1]
1 西南科技大学信息工程学院 四川 绵阳 621000
2 电子科技大学信息与软件工程学院 成都 610054
(syh@cqu.edu.cn)

摘 要 公共区域暴力行为频繁发生,视频监控对维护公共安全具有重要意义。相比固定摄像头,无人机具有监控灵活性,然而航拍成像中无人机快速运动以及姿态、高度的变化,使得目标出现运动模糊、尺度变化大的问题,针对该问题,设计了一种融合注意力机制的时空图卷积网络 AST-GCN(Attention Spatial-Temporal Graph Convolutional Networks),用于实现航拍视频暴力行为识别。该方法主要分为两步:利用关键帧检测网络完成初定位以及 AST-GCN 网络通过序列特征完成行为识别确认。首先,针对视频暴力行为定位,设计关键帧级联检测网络,实现基于人体姿态估计的暴力行为关键帧检测,初步判断暴力行为的发生时间。其次,在视频序列中提取关键帧前后的多帧人体骨架信息,对骨架数据进行归一化、筛选和补全,以提高对不同场景及部分关节点缺失的鲁棒性,并根据提取的骨架信息构建骨架时序-空间信息表达矩阵。最后,时空图卷积对多帧人体骨架信息进行分析识别,融合注意力模块,提升特征表达能力,完成暴力行为识别。在自建航拍暴力行为数据集上进行验证,实验结果表明,融合注意力机制的时空图卷积 AST-GCN 能实现航拍场景暴力行为识别,识别准确率达 86.6%。提出的航拍暴力行为识别方法对于航拍视频监控和行为理解等应用具有重要的工程价值和科学意义。

关键词:暴力行为识别;人体姿态估计;航拍;时空图卷积;级联网络;注意力机制
中图法分类号 TP391

Aerial Violence Recognition Based on Spatial-Temporal Graph Convolutional Networks and Attention Model

SHAO Yan-hua[1], LI Wen-feng[1], ZHANG Xiao-qiang[1], CHU Hong-yu[1], RAO Yun-bo[2] and CHEN Lu[1]
1 School of Information, Southwest University of Science and Technology, Mianyang, Sichuan 621000, China
2 School of Information and Software Engineering, University of Electronic Science & Technology, Chengdu 610054, China

Abstract The violence in public areas occurs frequently and video surveillance is of great significance for maintaining public safety. Compared with fixed cameras, unmanned aerial vehicles (UAVs) have surveillance mobility. However, in aerial images, the rapid movement of UAVs as well as the change of posture and height cause the problem of motion blur and large-scale change of target. To solve this problem, an attention spatial-temporal convolutional network (AST-GCN) combining attention mechanism is designed to realize the identification of violent behavior in aerial video. The proposed method is divided into two steps: the key frame detection network completes the initial positioning, and the AST-GCN network completes the behavior identification through the sequence features. Firstly, aiming at video violence localization, a key frame cascade detection network is designed to realize violence key frame detection based on human posture estimation, and preliminarily judge the occurrence time of violence. Secondly, the skeleton information of multiple frames around key frames is extracted from the video sequence, and the skeleton data is pre-processed, including normalization, screening and completion, so as to improve the robustness of different scenes and the partial missing of key nodes. And the skeleton temporal-spatial representation matrix is constructed according to the extracted skeleton information. Finally, AST-GCN network analyzes and identifies multiple frames of human skeleton information, to integrate attention module, improve feature expression ability, and complete the recognition of violent behavior. The method is validated on self-built aerial violence data set, and experimental results show that the AST-GCN can realize the recognition of aerial

到稿日期:2021-04-26 返修日期:2021-08-11
基金项目:国家自然科学基金(61601382);四川省教育厅项目(17ZB0454);西南科技大学博士基金(19zx7123);西南科技大学龙山人才(18LZX632)
This work was supported by the National Natural Science Foundation of China(61601382), Project of Sichuan Provincial Department of Education (17ZB0454), Doctoral Fund of Southwest University of Science and Technology(19zx7123) and Longshan Talent of Southwest University of Science and technology(18LZX632).
通信作者:李文峰(lwf.swust@qq.com)

scene violence, and the recognition accuracy is 86.6%. The proposed method has important engineering value and scientific significance for the realization of aerial video surveillance and human pose understanding applications.

Keywords Violence recognition, Human pose estimation, Aerial photography, Spatial-temporal graph convolutional, Cascade network, Attention mechanism

1 引言

近年来，公共区域个体与团体暴力事件频繁发生，严重危害人民群众的生命安全和心理健康，造成了巨大的人员伤亡和财产损失。视频监控作为目前最主要的安防监控手段，也成为了执法机构用于监控和遏制暴力行为的最常用方法[1]。然而，传统摄像头监视范围固定，难以实现灵活监控，因此，利用高灵活的性无人机进行监测，实现航拍暴力行为监测具有重要意义[2-3]。

对航拍视频进行人体行为识别，检测其中的行为是否为暴力行为并发出预警信号，成为了航拍监测的重点之一[4]。暴力视频检测一般指对视频中的暴力和暴力场景的检测，在一定程度上，暴力视频检测可以看作是视频动作识别的特例[5]。人体行为检测与识别可以从外形、动态光流、骨架等信息中提取特征[6-7]，这些信息中，人体骨架关节点因其对称性通常包含着重要的互补信息。然而，与外形和动态光流的特征建模相比，受限于传感器和计算量等原因，基于动态骨架的建模受到的关注较少[8]。

CNN 在图像领域的成功启发了基于 CNN 的动作视频识别研究[9-10]。近年来，人体姿态估计算法相继被提出[11-12]，其能够生成动态骨架，表示为一系列人体关节位置坐标的时间序列。Liu 等[13]试图利用关节间的自然连接来实现人体行为识别。根据人体的物理结构，将人体骨骼分解为 5 个部分，即两条胳膊、两条腿和一个躯干。为了有效地识别人类的各种行为，Li 等[14]将骨骼点提取与动作识别相结合，通过 Openpose 算法得到人体骨骼关键点数据，使用 SVM 作为分类器来完成多人行为识别，这表明了关节连接信息的重要性。然而，基于人体骨架关节点的行为识别，需要同时关注行为运动过程中空间与时间两个维度关节点的变化情况。Kim 等[15]提出了一种学习人体动作的可解释时空表示方法，直接使用时间卷积网络(temporal conv)分析动作的时间维度和空间维度特征，从而完成动作的识别。

随着图神经网络的发展，图卷积 GCN(Graph Convolutional Networks)因具有处理任意结构数据的特性，逐渐受到关注。人体骨架信息为典型非欧式结构数据，图卷积较传统卷积更适合用于处理该类问题。Yan 等[8]将图神经网络扩展为时空图卷积网络 ST-GCN(Spatial Temporal GCN)，利用时空图来形成骨架序列的层次化表示，实现基于骨架的人体行为识别。Chen 等[16]提出了改进的移位图进化网络(Shift-GCN)，弥补了传统 GCN 计算复杂度高、表达能力有限的缺点，提升了行为识别的精度。

目前已在多个任务中证明了注意力机制的有效性[17]，因此有必要对注意力机制在行为识别中的应用进行研究。Shi 等[18]设计了一种基于注意力的自适应图卷积网络，引入时空通道注意模块，更加关注关节、帧与通道特征，实现了基于骨架的行为识别。Markovitz 等[19]提出了一种图嵌入式位姿聚类算法，引入注意力子模块，用于异常行为检测。

机载计算平台受功耗等因素的影响，计算能力有限。因此，真实航拍场景在线行为理解需要先定位关键帧以减小识别对象空间，再进行视频识别以确认真实情况。本文设计了一种关键帧级联识别网络和融合注意力机制的时空图卷积 AST-GCN(Attention ST-GCN)，实现了航拍场景中的暴力行为识别。首先，设计级联识别网络，利用人体姿态信息实现基于人体姿态的暴力行为关键帧检测，初步判断暴力行为在视频中发生的时间；然后，在视频序列中提取关键帧前后的多帧人体骨架信息，对骨架数据进行归一化、筛选和补全，解决人体尺度变化及部分关节点缺失的问题，并将构建的暴力行为骨架时序-空间信息表示矩阵作为时空图卷积的输入；最后，利用时空图卷积对多帧人体骨架信息进行分析，在时空图卷积中引入注意力模块，以进一步提升特征描述能力和识别性能，完成航拍暴力行为识别验证。鉴于航拍暴力行为数据集的缺乏，本文自建了航拍暴力行为关键帧数据集和航拍暴力行为视频数据集，以进行相关实验。

2 航拍场景暴力行为识别

视频行为识别方法主要分为基于视频图像和基于人体骨架的方法。基于视频图像的行为识别方法将视频序列图像作为网络输入，识别效果受环境因素影响较大。基于人体骨架的行为识别方法依赖于准确的人体姿态估计，以提取到的人体的骨架信息作为输入，具有较强的鲁棒性，受环境因素的影响较小。因此，本文基于人体姿态估计设计了一种融合注意力机制的时空图卷积 AST-GCN，用于实现航拍场景中的暴力行为识别。

本文将航拍视频暴力行为识别分解为由粗到细的两个步骤。首先，基于姿态估计的暴力行为关键帧检测，实现对视频中暴力行为发生时间的快速定位；然后，提取关键帧前后的多帧图像，构建骨架时序-空间信息表达矩阵，实现序列暴力行为识别，进一步完成航拍暴力行为识别验证。

图 1 给出了本文航拍视频暴力行为识别的整体流程。首先由机载视觉系统获取感兴趣区域的视频序列，通过人体姿态估计和级联识别网络，实现视频暴力行为关键帧检测，并定位暴力行为的发生时间；然后提取关键帧前后的多帧人体骨架关节信息，构建骨架时序-空间信息表达矩阵，通过融合注意力机制的 AST-GCN，来实现暴力行为识别确认。时空图卷积能够充分提取人体骨架的空间特征和时间维度特征，对时序人体骨架数据进行深度特征分析，同时利用注意力机制学习通道间的特征相关性，实现通道特征重要性分配，提升行为识别精度。

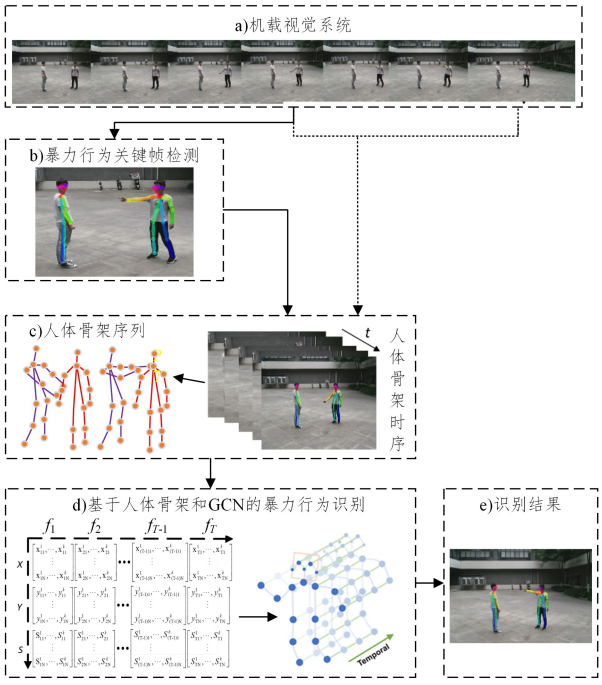


图 1 基于 AST-GCN 的航拍视频暴力行为识别

Fig. 1 Aerial violence recognition based on AST-GCN

3 基于姿态估计的关键帧检测

针对视频暴力行为关键帧定位，本文构建了一种端到端的暴力行为关键帧级联检测识别网络，如图 2 所示。

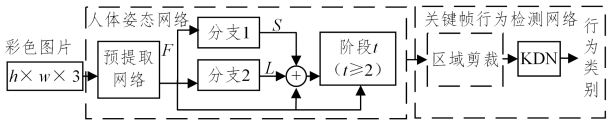


图 2 关键帧级联检测网络的结构

Fig. 2 Network structure of key frame cascade detection

关键帧级联检测网络主要包含人体姿态网络和关键帧行为检测网络 KDN(Keyframe Detection Network)两大模块。首先，通过人体姿态估计提取输入图片的特征，得到关节点位置和特征图，根据人体关节点位置进行人体区域裁剪，去除噪声，保留关键特征，然后将特征图输入到关键帧行为检测网络 KDN 中，实现暴力行为关键帧初定位。

3.1 人体关节点提取

人体姿态估计使用一种实时的多人关节点检测模型 CMU-Openpose 来提取特征[12]。图 2 所示的人体姿态网络采用大小为 $h\times w\times 3$ 的彩色图像作为输入，预提取网络生成一组特征 F，用于以后各个阶段的输入。网络每个阶段包含两个具有相同网络结构的分支，分别预测置信度图 S 和关节亲和域 L。在每个分支之后，将特征图 F、置信度图 S 和关节亲和域 L 连接起来，作为下一阶段的输入。多阶段提高预测精度，CMU-Openpose 有 6 个预测阶段，通过在 $t\geqslant 2$ 的整个阶段改进其自身的预测，并充分利用上下文信息精确地定位人体关节，提取感兴趣的关节点坐标信息。

3.2 关键帧行为识别

关键帧行为检测网络的 KDN 架构如图 3 所示。该网络具有的特点如下：1)通过双分支融合保留更多信息；2)采用全局均值池化层 GAP(Global Average Pooling)[20] 替代传统全连接层，GAP 具有降低参数量并可防止在该层出现过拟合的优势。

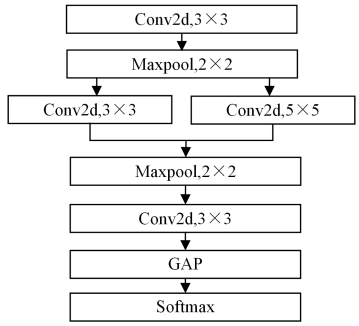


图 3 关键帧检测网络

Fig. 3 Key frame detection network

3.2.1 特征图

人体姿态骨架采用 COCO 关节点格式，关节点示意图如图 4 所示。通过人体姿态估计网络获取每个人的关节点二维坐标，使用 CMU-Openpose[12] 预测置信度图 S 和关节亲和域 L，整合后得到识别模块的输入特征图。同时，根据关节点对动作贡献的大小，本文级联识别网络的输入为序号为 1－13 的关节点对应的特征图，其尺寸为 184×216×39 张量。

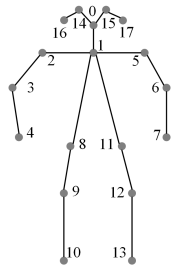


图 4 人体姿态骨架关节点示意图

Fig. 4 Schematic diagram of joint points of human skeleton

3.2.2 人体区域裁剪

不同的人员个体尺度差异较大，为提升暴力行为识别的准确率，需要根据关节点位置对人体区域进行裁剪，从而去除噪声，保留关键特征，本文定义裁剪区域 S_{crop} 如图 5 所示，矩形大小为 $(d_2+5d_1, 4d_1)$，最后将裁剪区域图像尺寸调整为 432×368。其中，d_1 为颈部与髋关节之间的较大距离，d_2 为双方颈部之间的水平距离。

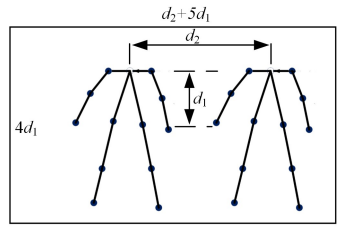


图 5 人体区域裁剪

Fig. 5 Body area cropping

3.2.3 损失函数

关键帧检测识别网络采用交叉熵(Crossing Entropy)

损失函数，如式(1)所示：

$$L(y,y^*)=\frac{1}{5}\sum_{p=1}^{5}y_p(i)\log^{y_p^*(i)} \tag{1}$$

其中，y 和 y^* 分别是网络输出的预测行为类别以及真实行为类别，行为类别使用独热(One-hot)编码格式。本文定义的典型暴力行为为 4 类，再加上正常的行为 1 类，因此关键帧识别网络的最后结果为 5 类。

4 基于 AST-GCN 的暴力行为识别

4.1 网络结构

基于 AST-GCN 的暴力行为识别的基本流程如图 6 所示。首先通过暴力视频片段提取骨架序列，构建骨架时序-空间信息表达矩阵，然后将该矩阵作为 AST-GCN 网络模型的输入，完成航拍视频暴力行为识别。

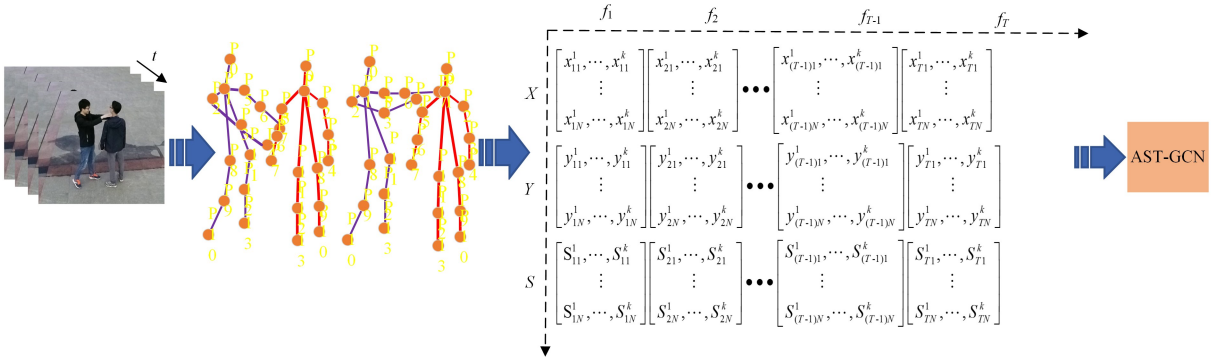


图 6 时序暴力行为识别

Fig. 6 Chronological violence recognition

4.2 骨架时序-空间信息表达矩阵

首先，计算骨架时序。利用人体姿态估计提取视频片段中多帧人体骨架数据信息，保存数据格式为 $(x_k^0,y_k^0,s_k^0,\cdots,x_k^t,y_k^t,s_k^t,\cdots,x_k^{17},y_k^{17},s_k^{17})$，其中，对于每帧的骨架信息，$x$，$y$ 分别表示关节点水平坐标和垂直坐标，s 表示关节点置信度，k 表示当前帧中第 k 个人员实例，t 表示相应关节点索引。

然后，对关节点数据进行归一化。不同行为动作以及人体在图像中的位置、距离摄像头的距离等几何尺度，都会对姿态估计提取的关节点坐标产生影响，因此需要对关节点进行合适的预处理。如图 7 所示，处于同样姿态的两人，相同关节坐标的差异较大。对关节点数据进行归一化，使得骨架图关节点特征只取决于行为动作本身，消除其他外在因素对特征提取的影响。对关节点进行归一化处理，如式(2)所示：

$$(x_n,y_n)=\left(\frac{x-x_c}{w},\frac{y-y_c}{h}\right) \tag{2}$$

其中，(x_n,y_n) 为经过归一化后的关节点坐标，(w,h) 代表图像尺寸，(x,y) 为关节点横纵坐标，(x_c,y_c) 定义为坐标中心点。本文以 1 号关节点为归一化坐标中心。

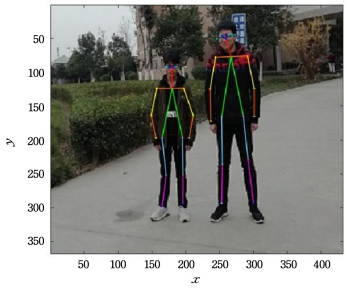


图 7 人体姿态估计尺度示意

Fig. 7 Schematic diagram of human body posture estimation scale

最后，对归一化骨架关节数据进行筛选、补全。由于行为识别的网络要求输入特定的骨架图关节点数量，因此，当提取的骨架图存在部分关节点丢失时，需对缺失的关节点进行筛选补全。若未提取到当前帧骨架图中上半身或下半身部分的重要关节点，则舍弃该帧，并将其替换为后续帧中较准确的结果。关节点补全指当未检测到某一关节点时，使用对称位置的关节点来替代，若两关节点均未检测到，则两者的坐标设为(0,0)。

经过以上步骤，按时序排列构建骨架时序-空间信息表达矩阵，如图 8 所示，纵轴(X,Y,S)三维坐标值如同彩色 RGB 图的 R,G,B 3 个通道矩阵，横轴 $f_1,f_2,\cdots,f_{T-1}$ 包含时序信息。骨架时序-空间信息表达矩阵的维度为[$T\times18\times3\times K$]，其中，T 代表骨架时序帧数，18 代表人体骨架中的 18 个关节点，如图 5 所示，3 代表各关节点的横纵坐标及置信度，K 代表该帧中检测到的人数。在将该矩阵输入网络模型前需进行矩阵变换，变为[$batchsize\times K,T,18,3$]的张量。

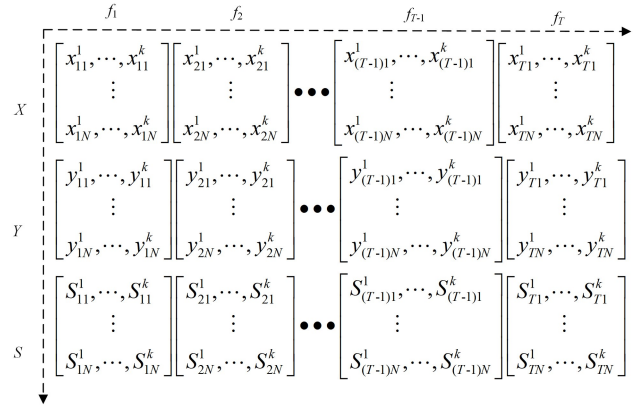


图 8 骨架时序-空间信息表达矩阵

Fig. 8 Temporal-spatial representation matrix of skeleton

4.3 图卷积

简单的图卷积层可以写成一个非线性函数，具体如下：

$$\boldsymbol{H}^{(l+1)}=f(\boldsymbol{X},\boldsymbol{A})=\sigma(\boldsymbol{A}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \tag{3}$$

其中，$\boldsymbol{A}$ 为邻接矩阵，$\boldsymbol{H}^{(l)}$ 是第 l 层输出的激活特征矩阵，$\boldsymbol{W}^{(l)}$

为第 l 层可训练权重参数矩阵，$\sigma(\cdot)$为非线性激活函数。针对输入特征 $\boldsymbol{X}$，考虑自身节点信息自传递的问题以及对邻接矩阵进行归一化，图卷积计算公式可进一步表示为：

$$\boldsymbol{H}^{(l+1)}=\sigma(\tilde{\boldsymbol{\Lambda}}^{-\frac{1}{2}}\tilde{\boldsymbol{A}}\tilde{\boldsymbol{\Lambda}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}) \tag{4}$$

其中，$\tilde{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}_N$ 是邻接矩阵和自连接单位矩阵 $\boldsymbol{I}_N$ 的加权，$\tilde{\boldsymbol{\Lambda}}_{ij}=\sum_j\tilde{\boldsymbol{A}}_{ij}$ 为度矩阵，$\boldsymbol{H}^{(0)}=\boldsymbol{X}$。

4.4 融合注意力机制的时空图卷积(AST-GCN)

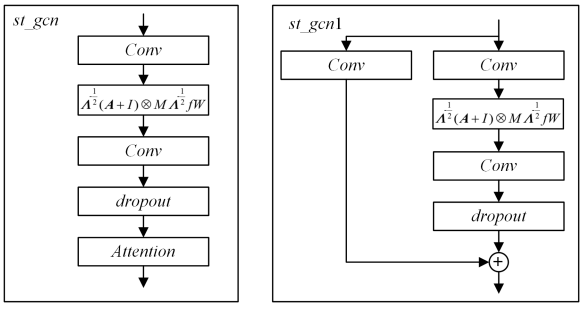
首先对行为视频片段执行姿势估计，以获取骨架时序-空间信息表示矩阵，应用多层时空图卷积分析挖掘自然连接关节点的空间变化以及不同时刻相同关节位置的时间变化，逐步生成更高级别的特征。整个多层时空图卷积和 softmax 分类器组成网络结构如表 1 所列。

表 1 AST-GCN 网络结构

Table 1 AST-GCN network structure

网络层	输出张量
st_gcn	(*batch* * 2, *T*, 18, 64)
st_gcn	(*batch* * 2, *T*, 18, 64)
st_gcn	(*batch* * 2, *T*, 18, 64)
st_gcn	(*batch* * 2, *T*, 18, 64)
*st_gcn*1	(*batch* * 2, [*T*/2], 18, 128)
st_gcn	(*batch* * 2, [*T*/2], 18, 128)
st_gcn	(*batch* * 2, [*T*/2], 18, 128)
*st_gcn*1	(*batch* * 2, [[*T*/2]/2], 18, 256)
st_gcn	(*batch* * 2, [[*T*/2]/2], 18, 256)
st_gcn	(*batch* * 2, [[*T*/2]/2], 18, 256)
shape	(*batch*, 2, [[*T*/2]/2], 18, 256)
mean	(*batch*, [[*T*/2]/2], 18, 256)
conv	(*batch*, [[*T*/2]/2], 18, 4)
avg_pool	(*batch*, 1, 1, 4)
shape	(*batch*, 4)
softmax	4

图 9 中，*st_gcn* 和 *st_gcn*1 为时空图卷积中的基础结构，*shape* 为维度转换操作，*mean* 是在人数维度求特征均值，*conv* 为 1×1 卷积层，*avg_pool* 为平均池化层，最后通过标准 *softmax* 分类器得到 4 种暴力行为的置信度，[]为向上取整。



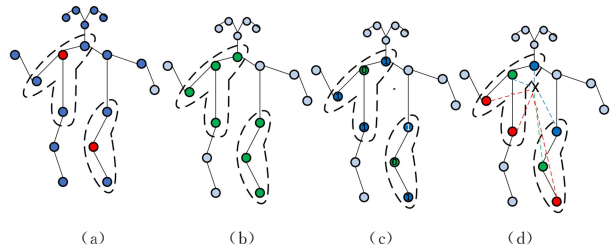
(a)st_gcn structure (b)st_gcn1 structure

图 9 单层时空图卷积结构

Fig. 9 Structure of single layer AST-GCN

4.4.1 分区策略

合适的分区策略对邻接矩阵的构建至关重要。在人体行为识别上的应用中，时空图卷积主要包含 3 种分区策略，同时对于单帧的分区策略，可自然地扩展到时空域[4,8]。如图 10 所示，虚线圈表示卷积核 $D=1$ 的感受野。图 10(a)表示提取的骨架信息，人体各关节点标记为蓝点；图 10(b)表示单标签分区策略，将节点感受野为 1 的领域集内的所有节点均标记为同一标签(绿色)；图 10(c)表示距离分区策略，节点领域分为两个子集，即距离为 0 的根节点本身(绿色)以及距离为 1 的相邻节点(蓝色)；图 10(d)表示空间配置分区策略，以根节点(绿色)到骨架重心的距离为基准，相邻各节点中与骨架重心(黑色十字)的距离小于该基准的为向心节点(蓝色)，大于则为离心节点(红色)。



(a) (b) (c) (d)

图 10 分区策略(电子版为彩图)

Fig. 10 Partition strategies

4.4.2 可学习的边缘重要性加权

在不同动作情况下，各关节点对行为识别的重要性有所不同，动态建模时应该具有不同的重要性，因此在图卷积中增加一个可学习的掩码 M[8]。掩码作为可学习的重要性权重，通过缩放节点特征来调整节点对其相邻节点的贡献。掩码 M 与节点的邻接矩阵相乘，使用 $(\boldsymbol{A}+I)\otimes M$ 作为输入，其中 $\otimes$ 表示两个矩阵之间的元素乘积，掩码 M 初始化为一个全 1 矩阵。添加可学习的重要性加权可以进一步提高时空图卷积的识别性能。

4.5 通道注意力模块

在目标检测、分类等任务中，注意力机制的有效性得到了证明，并被逐渐引入到行为识别领域中[18]。时空图卷积能够有效地提取空间特征与时序特征，为使网络自动学习特征图通道之间的相关性和重要性，本文在 st_*gcn* 基础网络中引入通道注意力模块 SENet (Squeeze-and-Excitation Networks)[17]。如图 11 所示，SENet 主要通过压缩(Squeeze)和激励(Excitation)两个操作来实现通道特征重要性的学习。首先通过全局池化 GAP 实现特征压缩，通过全连接层 FC 对特征降维，ReLU 激活层学习特征通道间的非线性关系，然后 FC 对特征图进行升维。最后将 Sigmoid 激活函数得到的权重与输入特征图进行乘积操作，从而实现特征权值重分配。通道注意力模块 SENet 通过对各通道的依赖性进行建模，自适应调整各通道的特征响应值，加强有用特征并抑制无用特征，从而提升网络的特征描述能力。

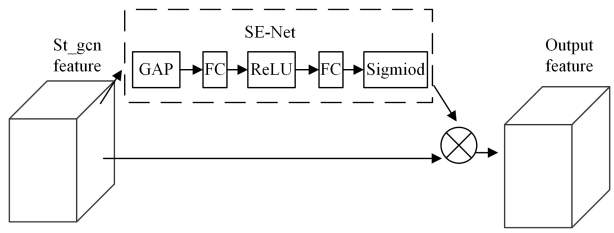


图 11 Attention 模块

Fig. 11 Attention module

5 实验分析

5.1 航拍暴力行为数据集构建

依据暴力行为识别流程，分别构建用于检测暴力行为关键帧的航拍暴力图片（关键帧）数据集和基于 AST-GCN 的暴力行为识别的航拍暴力视频数据集。为了构建航拍暴力数据集，结合领域专家意见以及参考 AVI 数据集[21]中暴力行为的分类，本文中主要包含扼杀、锤击、脚踢、射击 4 类典型的暴力行为，运用无人机对暴力行为进行拍摄以制作数据集。航拍暴力行为以 20～26 岁的健康受试者 10 人为数据采集对象，随机将其分成 5 组。每组测试者依次完成 4 种暴力行为，无人机定高 4m，距离动作行为人依次 3m，5m，7m 进行航拍，相机和垂直面夹角为 45°，分辨率为 640×480，帧率为 30 fps，以此完成暴力行为视频的录制。

5.1.1 航拍暴力行为关键帧数据集

暴力行为关键帧数据集由暴力行为典型动作帧构成，用于视频暴力行为初定位。首先，在录制的暴力行为视频数据集中提取具有典型暴力行为动作的图片帧。无人机航拍进行视频采集，测试者依次执行指定类别的行为动作，每个动作持续 90 s。视频数据中相邻帧图像差别很小，为避免数据冗余，在每秒 30 帧数据中间隔采样 10 帧图像并完成人体区域裁剪，将结果作为关键帧数据集。典型暴力帧图像如图 12 所示，详细信息如表 2 所列[1)]。

(a)扼杀 (b)锤击

(c)脚踢 (d)射击

图 12 典型暴力动作帧

Fig. 12 Typical violent behavior frame

表 2 航拍暴力行为关键帧数据集

Table 2 Key frame data set of aerial violent behaviors

行为	分辨率	关键帧数量
扼杀 strangling	640×480	1300
锤击 hammering	640×480	1300
脚踢 kicking	640×480	1300
射击 shooting	640×480	1300

5.1.2 航拍暴力行为视频数据集

航拍暴力行为视频数据集由暴力视频片段组成，用于暴力行为序列的识别。数据集包含 4 类行为，共有 1600 个视频片段，每个视频帧率为 30 fps，数据集详细信息如表 3 所列。

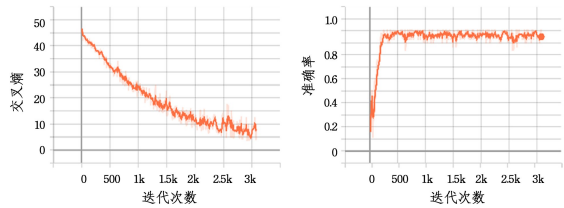
表 3 航拍暴力行为视频数据集

Table 3 Aerial violence video data set

行为	视频帧率/fps	分辨率	视频片段数
扼杀 strangling	30	640×480	400
锤击 hammering	30	640×480	400
脚踢 kicking	30	640×480	400
射击 shooting	30	640×480	400

5.2 暴力行为关键帧定位实验

自制航拍关键帧数据集划分为 4500 张训练集和 2000 张测试集。训练实验平台为 NVDIA GeForce GTX 1050 Ti，使用 Tensorflow1.13 框架作为训练模型，Batchsize 设置为 32，初始学习率 lr 设置为 0.0001，同时采用退化学习率方法。训练时交叉熵损失和准确率曲线如图 13 所示。



(a)交叉熵 (b)准确率

图 13 训练的交叉熵和准确率曲线

Fig. 13 Curve of trained cross entropy and accuracy

广场、车库和建筑物附近等典型场景的检测效果如图 14 所示，其中，每行代表一个典型动作，1－4 行分别为扼杀、锤击、脚踢和射击。每列为一个典型场景，1－3 列分别为广场、车库和楼房附近。

图 14 暴力行为关键帧检测结果

Fig. 14 Key frame detection results of violent behaviors

5.3 时序暴力行为识别实验

通过文献调研和对视频片段进行分析可知，当视频帧率为 30 fps 时，每种暴力行为从开始到完成约需 30～40 帧。为保证提取暴力行为图像帧的有效性，且包含行为动作的主要过程，本文有间隔地在关键帧前后提取 30 帧视频图像。

使用人体姿态估计算法得到骨架序列，构建骨架时序-空间信息表达矩阵，用于代表整个暴力行为特征，并将其用于

1) https://github.com/qiuqiu-lwf/A-ST-GCN

训练和预测。首先根据航拍采集数据集，并将其划分为训练集和测试集，分别为1400个和200个视频片段，用于模型训练与验证。最终识别精度ACC定义为正确识别视频片段数与测试集总数之比。

如表4所列，本文针对不同策略分析了其对识别准确率的影响。首先考虑邻接矩阵构建时不同分区策略的影响，在3种分区策略下，空间配置分区策略能够更好地获取骨骼间的信息，识别精度最优。进一步地，针对策略3，在其基础上加入边缘重要性加权，从而提升识别性能。最后，考虑引入的注意力机制，通过增加通道特征描述能力来提升识别准确率。本文最终采用策略5。同时，针对航拍暴力行为识别实时性进行分析。暴力行为单次关键帧检测在Jetson TX2上完成，使用CUDNN和TensorRT加速，在地面监控中心通过时空图卷积完成多帧暴力行为骨架信息分析。实际场景下的时间性能测试如表5所列。

表4 不同策略下的识别准确率

Table 4 Recognition accuracy under different strategies

（单位：%）

方法	准确率
策略1(单标签分区)	79.5
策略2(距离划分)	81.5
策略3(空间配置分区)	82.0
策略4(策略3+可学习的边缘重要性加权)	85.5
策略5(策略4+注意力机制)	**86.6**

表5 真实场景测试

Table 5 Real scene test

（单位：ms）

场景	关键帧检测时间	骨架图时序分析时间
广场	160	61.8
车库	160	64.0
楼房旁	150	66.5

针对暴力行为识别，为避免训练时的过拟合问题，本文在时空图卷积网络中加入Dropout层随机失活神经元，对其参数进行调整。设置参数寻优范围为0.5～0.9，分别进行5次测试并取5次结果的均值，具体结果如图15所示，其中ACC为识别精度。Dropout取值为0.7时得到最高的识别精度为86.6%。

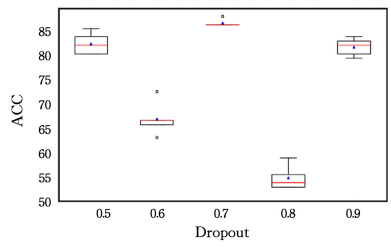


图15 不同Dropout下的识别精度

Fig. 15 Recognition accuracy under different Dropouts

最后与已有行为识别相关算法进行性能对比，在自建数据集上使用不同算法进行实验。如表6所列，本文提出的AST-GCN较其他算法有一定提升，与Chen[4]提出的算法相比识别准确率提升了1.1%，最终识别精度可达86.6%。

表6 相关算法实验对比

Table 6 Experimental comparison of related algorithms

（单位：%）

模型	准确率
Fernando等[22]提出的模型	74.0
Shahroudy等[23]提出的模型	78.5
Kim等[15]提出的模型	81.0
Chen[4]提出的模型	85.5
本文模型	86.6

结束语 本文设计了一种基于融合注意力机制的时空图卷积AST-GCN的航拍视频暴力行为识别方法。首先通过人体姿态估计获取人体关节点坐标，根据人体关节点位置进行人体区域裁剪，获取人体区域特征图，将其作为关键帧检测识别网络的输入，实现视频暴力行为关键帧检测；然后根据关键帧位置有间隔地提取前后多帧暴力行为时序图像，构建骨架时序-空间信息表达矩阵；最后在时空图卷积网络中引入通道注意力机制，获取每个特征通道的权重，融合完成基于AST-GCN的航拍视频暴力行为识别。本文方法对于航拍视频监控和人机交互等应用实现具有重要的工程价值和科学意义。

本文对未来工作的展望：

(1)使用更有效的注意力机制，从而实现更为高效的特征描述；

(2)无人机航拍监测中相机抖动、低分辨率、相似动作干扰和续航等问题也是后续研究中面临的重大挑战。

参 考 文 献

[1] MA Y X, TAN L, DONG X, et al. Behavior Recognition For Smart Surveillance[J]. Journal of Image and Graphics, 2019, 24(2): 282-290.

[2] DOROGYY Y, KOLISNICHENKO V, LEVCHENKO K. Violent crime detection system[C]// 2018 IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT). IEEE, 2018: 352-355.

[3] LI T, LIU J, ZHANG W, et al. UAV-Human: A Large Benchmark for Human BehaviorUnderstanding with Unmanned Aerial Vehicles[C]// 2021 Conference on Computer Vision and Pattern Recognition (CVPR). 2021: 16266-16275.

[4] CHEN L. Violent behavior monitoring in aerial scenes based on human pose estimation[D]. Mianyang: Southwest University of Science and Technology, 2020.

[5] SONG W, ZHANG D, ZHAO X, et al. A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks [J]. IEEE Access, 2019, 7: 39172-39179.

[6] HE L, SHAO Z P, ZHANG J H, et al. Review of Deep Learning-based Action Recognition Algorithms [J]. Computer Science, 2020, 47(6A): 139-147.

[7] TIAN Z S, YANG L K, FU C Y, et al. Human action recognition based on multi-antenna FMCW radar [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2020, 32(5): 779-787.

[8] YAN S, XIONG Y, LIN D. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition[C]// Thirty-second AAAI Conference on Artificial Intelligence

(AAAI). 2018:7444-7452.

[9] YAO G,LEI T,ZHONG J. A Review of Convolutional-Neural-Network-Based Action Recognition[J]. Pattern Recognition Letters,2018,118:14-22.

[10] GAO C Q,CHEN X. Deep learning based action detection:a survey [J]. Journal of Chongqing University of Posts and Telecommunications(Natural Science Edition),2020,32(6):991-102.

[11] FANG H S,XIE S,TAI Y W,et al. RMPE:Regional Multi-person Pose Estimation[C]// 2017 International Conference on Computer Vision (ICCV). 2017:2353-2362.

[12] CAO Z,SIMON T,WEI S,et al. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields[C]// Computer Vision and Pattern Recognition. 2017:1302-1310.

[13] LIU J,SHAHROUDY A,PEREZ M,et al. Ntu rgb+d 120:A large-scale benchmark for 3d human activity understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2019,42(10):2684-2701.

[14] LI M H,XU H J,SHI L X,et al. Multi-person Activity Recognition Based on Bone Keypoints Detection[J]. Computer Science,2021,48(4):138-143.

[15] KIM T S,RRITER A. Interpretable 3D Human Action Analysis with Temporal Convolutional Networks[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). 2017:20-28.

[16] CHENG K,ZHANG Y,HE X,et al. Skeleton-based action recognition with shift graph convolutional network[C]// 2020 Conference on Computer Vision and Pattern Recognition (CVPR). 2020:183-192.

[17] HU J,SHEN L,ALBANIE S,et al. Squeeze-and-Excitation Networks[J]. IEEE Transactions on Pattern Analysis Machine Intelligence,2017,42(8):2011-2023.

[18] SHI L,ZHANG Y,CHENG J,et al. Skeleton-based action recognition with multi-stream adaptive graph convolutional networks[J]. IEEE Transactions on Image Processing,2020,29:9532-9545.

[19] MARKOVITZ A,SHARIR G,FRIEDMAN I,et al. Graph Embedded Pose Clustering for Anomaly Detection[C]// 2020 Conference on Computer Vision and Pattern Recognition(CVPR). 2020:10536-10544.

[20] LIN M,CHEN Q,YAN S. Network In Network[C]// 2014 International Conference on Learning Representations(ICLR)[J]. arXiv:1312. 4400,2013.

[21] SINGH A,PATIL D,OMKAR S. Eye in the Sky:Real-time Drone Surveillance System (DSS) for Violent Individuals Identification using ScatterNet Hybrid Deep Learning Network[C]// 2018 Conference on Computer Vision and Pattern Recognition (CVPR). 2018:1629-1637.

[22] FERNANDO B,GAVVES E,ORAMAS J,et al. Modeling video evolution for action recognition[C]// 2015 Conference on Computer Vision and Pattern Recognition (CVPR). 2015: 5378-5387.

[23] SHAHROUDY A,LIU J,NG T T,et al. Ntu rgb+d:A large scale dataset for 3d human activityanalysis[C]// 2016 Conference on Computer Vision and Pattern Recognition (CVPR). 2016:1010-1019.

SHAO Yan-hua, born in 1982, Ph.D, lecturer, is a member of China Computer Federation. His main research interests include computervision and machine learning.

LI Wen-feng, born in 1997, postgraduate. His main research interests include computervision and so on.

(责任编辑:喻藜)