

基于深度学习的校园欺凌行为检测研究

符水波¹, 李 刚²

(1.宁波市镇海蛟川书院, 浙江 宁波 315200; 2.宁波大学 信息科学与工程学院, 浙江 宁波 315211)

摘要: 为通过校园监控网络实时发现校园欺凌事件, 提出采用人体姿态检测模型与递归神经网络相结合的方法来分析视频中的每一对个体行为, 并通过一个完全连接的网络来识别一系列视频帧的行为, 进而判断是否存在校园暴力行为。此外, 还建立一个包含了来自校园不同地方的3 000多个视频片段数据库, 并对视频中的运动行为进行标注, 以用于训练及测试。通过对照实验发现: 与传统方法相比, 本文设计算法的实时分析准确率更高。

关键词: 运动检测; 动作识别; 深度学习

中图分类号: TP391.4

文献标志码: A

文章编号: 1001-5132 (2020) 03-0057-06

“校园欺凌现象”由来已久, 它是校园冲突的表现之一, 自有学校以来, 学生之间的冲突未曾断绝。零星、轻微、小范围、不触碰道德及法律底线的校园冲突只需按照校规校纪处置便可解决; 但当欺凌现象范围越来越广, 程度越来越严重, 影响了正常的学校教学秩序, 甚至超出了人类的道德良知底线及法律法规底线时, 这一现象就成为了问题^[1]。在2017年3月3日召开的全国政协大会上, 浙江省镇海中学学生以“模拟提案”方式提交的调查报告《关于完善现有校园欺凌预防和处理体系的提案》表明, 校园欺凌已经引发社会各方关注。事实上, 很多校园欺凌甚至发生在初中校园^[2]。经常发生的暴力事件不仅会对受害者本身造成身心方面的伤害, 更会阻碍其他无辜学生的正常学习生活, 甚至引发大规模的公民恐慌。

对于校园欺凌现象, 事后调查并提供医疗辅助工具也会有弥补和修正作用, 但更重要的是提供早期预警, 以便在发生事件时能及时阻止事态的进一步发展。因此, 有必要对校园中欺凌行为进行实时监视和检测。目前视频监控设备在各个校园实现了较高的普及率, 有助于通过监控系统来实现此目标, 但单纯人工检测不仅工作量大, 且难免出现疏漏。为此, 本文针对事先预防和事中监控

两个环节开展行为研究, 以达到拉长监控、制止校园欺凌的时间链条的目的。

在过去的几十年中, 许多研究者致力于行动识别和暴力检测, 通过分析每个簇的密度和形状来检测异常运动模式。一些研究人员建议使用面部表情、头部、手部和身体运动以及语音跟踪来检测愤怒情绪及攻击行为。对于视频, 建议使用多个摄像机来跟踪人员以获得3D信息, 并提取关于肢体运动的特征。对于音频, 可对语音、尖叫、言语和踢对象的实例进行分类。一些研究还使用由运动的方向和大小组成的加速度矢量来进行检测。

同样, 基于光流法的判决方式也已被广泛使用。一些研究者提出结合背景差分 and 光流法的检测算法对图像背景差分提取前景寻找特征点, 利用光流法预测特征点位置, 将特征点以光流运动方向为依据划分后处理数据, 得到累积加速度并加以判断^[3]。也提出了两级聚合的方法, 通过自然和分析的方式检测暴力行为, 从连续帧中提取运动像素和运动区域, 然后计算用于分类的光流向量^[4]。也提出了一种高斯光流模型来提取候选暴力区域, 这些区域被自适应地建模为偏离场景中观察到的人群正常行为, 然后通过对候选暴力区域进行密集采样来执行暴力检测^[5]。也提出了一种结

合光流法与信息熵的人群突发事件检测与判断算法,通过LK光流算法从视频中提取运动物体的特征点,并得到特征点的位置信息,根据位置信息计算速度和加速度,分析运动强度,检测突发事件^[6]。

然而,许多方法容易受到背景噪声等因素影响。例如,当2名学生在校园里与许多其他学生一起经过时,光学流动很容易被背景分散。为了改善校园中的暴力行为检测的效率,本文使用深度学习方法来提取身体骨骼,并通过骨骼间的交互运动判定是否有暴力行为发生。在新方法中,改进了骨架的提取模式以获得更高的准确性。同时提出两个递归神经网络来模拟暴力行为,并用一个完全连接的网络来识别一系列视频帧的暴力行为。根据实验结果可以发现,与传统方法相比,新方法在实时表现等方面做得更为优秀。

1 检测方法

1.1 模型构建

通过构建OpenPose算法模型来帮助检测暴力事件。首先,从监控系统中获取视频,提取身体上的关键点;接着,将这些关键点视作连续运动的样本,并在每个个体的一系列图像间匹配这些关键点,为区分不同图像中的相同个体,使用不同颜色表示不同个体,并通过检查和合并错误关键点来优化OpenPose程序,以获得更高的检测精度;然后将每20个图像的位置信息聚合形成数据集,并进行标准化;最后使用数据集通过Softmax交叉熵损失来训练递归神经网络,以检测暴力行为发生与否。相应流程如图1所示。

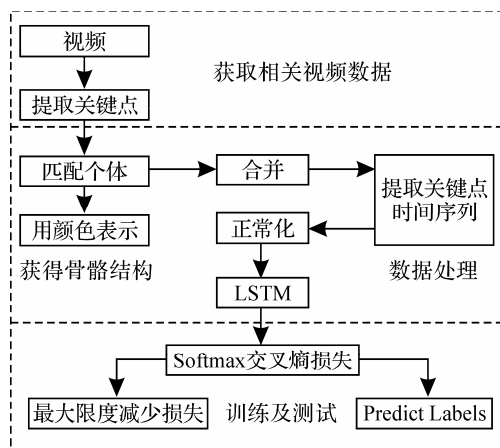


图1 关键点的规范化流程

1.2 数据建立

OpenPose算法是输入大小为 $w \times h$ 的彩色图像,并记录输入图像中所检测到每个人身体关键点的2D坐标,例如左右肩、手腕、膝盖等^[7]。该架构能够以相对高的准确度实时识别图像中的这些关键点,并将其一一对应。原始图像的示例和相应结果如图2所示。

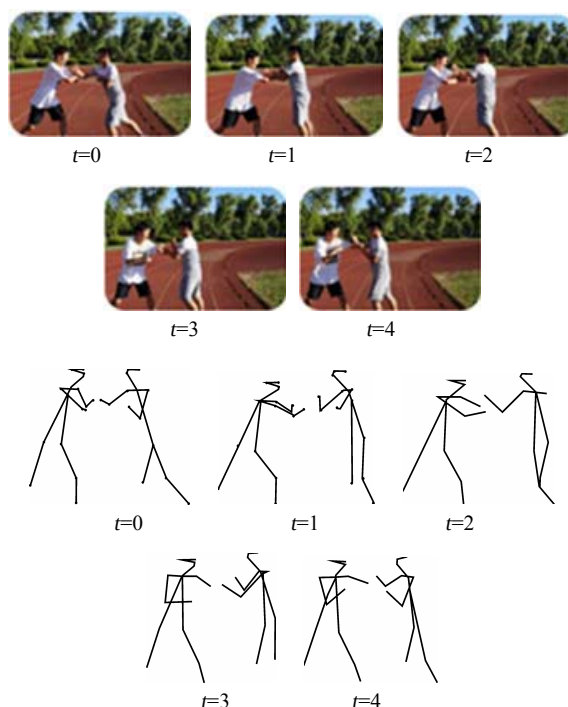


图2 原始图像及人体骨骼详解

2 检测数据处理

2.1 匹配关键点

我们使用特定颜色在一系列图像中表示一个特定的实验对象,以便清晰呈现对象间的交互(图2)。在上一步中,首先得到了每个人的骨架结构,接着需要在一系列图像中匹配同一个人的对应关键点,以便识别该人的行为。如果视频帧速率足够高,就可以将个体的位置信息存储在矩阵中,并将匹配问题视为优化问题。

表1定义了表示方法,其优化目的是找到最佳 I (转置矩阵)来最小化匹配成本 $\min_I |A_f - A_{f+1} \times I|$ 。但是,目标函数和约束函数不是凸函数。此外,当检测到的个体数量不同时, A_f 和 A_{f+1} 可以具有不同的行数,因此不能使用凸优化,虽然也可以通过枚举找到最好的结果,但在时间效率上会大打折扣。

扣. 可考虑到 2 个相邻帧中的同一个人几乎没有位置变化, 只是枚举 f 帧中某对象(如个体 i), 并且在帧 $f+1$ 中找到对象 j , 并使得 $D((x_{i,f}, y_{i,f}), (x_{i,f+1}, y_{i,f+1}))$ 最小.

表 1 用作人物匹配的符号

函数	说明
$(x_{i,f}, y_{i,f})$	帧 f 中的个体 i 的像素位置, 其中, 将左右肩的中心视为个体 i 的位置
A_f	包含帧 f 中所有个体关键点信息的矩阵, 其中每一行包含一个个体关键点的像素位置. 其中, 行和个人之间的对应关系是随机的
I	排列矩阵中每行每列中只有一对条目相等
$D((x_{i,f_1}, y_{i,f_1}), (x_{i,f_2}, y_{i,f_2}))$	计算帧 f_2 中各个 i 帧 f_1 和个体 j 之间的欧几里德距离的函数

函数 $D(\cdot)$ 定义如下:

$$D((x_{i,f}, y_{i,f}), (x_{i,f+1}, y_{i,f+1})) = (x_{i,f} - x_{i,f+1})^2 + (y_{i,f} - y_{i,f+1})^2. \quad (1)$$

然而, 如果最小的 $D((x_{i,f}, y_{i,f}), (x_{i,f+1}, y_{i,f+1}))$ 仍然大于阈值, 即个体 j 在之前找不到最佳匹配框架, 我们便认为个人 j 第 1 次出现在第 $f+1$ 帧中, 并赋予这个对象一种新的颜色.

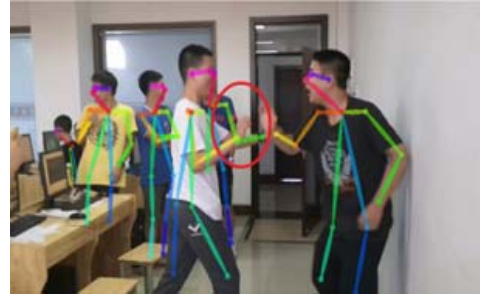
在此过程中, 我们将左肩和右肩的中心视为研究对象的位置, 以避免一些干扰因素(如突然伸出手). 考虑到左右肩的运动范围相对较小, 为此选择肩部中心的位置作为匹配的关键点.

2.2 优化关键点

当匹配相邻的帧时, 我们发现 OpenPose 算法检测到的学生关键点通常都有误差. 例如图 3 所示, 属于同一个人的关键点有时可能被检测为 2 个独立个体的关键点.

这样的流量显然会降低打斗行为检测的准确性, 因此有必要在进一步计算之前优化身体骨架, 于是需要提出一个算法来解决此问题. 通过合并实际属于一个人但被错误地分配给几个人的点, 并将合并的关键点存储为一个人.

为确保每个关键点属于正确的人, 使用适应度函数 $F(\cdot)$ 来计算关键点和个体之间的适应度.



(a) OpenPose 中的示意图



(b) 骨架优化示意图

图 3 优化 OpenPose 算法关键点

适应度函数定义如下:

$$F(x, y, z) = \Sigma((bl(x) - bl(y)) \times bl(z) \times (x - z)^2) / \Sigma((bl(x) - bl(y)) \times bl(z)), \quad (2)$$

其中, $x = A_f[i, :]$, $y = A_{f+1}[j, :]$, $z = A_{f+1}[k, :]$, 而 i, j, k 是研究对象的身份. 当缺少某些关键点时, 在 x, y, z 的相应条目中表示为零. 因此, $F(x, y, z)$ 基本上是 y 的缺失部分处的 x 和 z 之间的加权平均值, 详细符号解释见表 2. 在不失一般性情况下, 可假设帧 f 中的个体 i 通过在前节中讨论的关键点匹配方法与帧 $f+1$ 中的个体 j 匹配. 在此函数中, 假设每个符号关键点帧 f 中的个体是完整的, 在下一帧中, 如果已经在帧 f 中获得个体 i , 并且在帧 $f+1$ 中配对个体 j , 但是被检测到的 j 的关键点明显小于 i , 则可在帧 $f+1$ 中找到另一个人 k , 使得最小化 $F(x, y, z)$ (其中, $z = A_{f+1}[k, :]$); 如果 $F(x, y, z)$ 小于设定阈值, 则将单个 k 的关键点合并到帧 $f+1$ 中个体 j 的关键点; 之后, 如果个别 j 的关键点仍明显小于 i , 就重复此过程.

表 2 用于确定估价的符号

函数	说明
$bl(x)$	其函数采用数值向量 x 作为输入, 并返回布尔向量作为具有相同大小的输出. 如果 $x_i > 0$, 则 $bl(x)_i = 1$, 如果 $x_i = 0$, 则 $bl(x)_i = 0$

续表

函数	说明
-	接输入 2 个布尔向量并返回布尔向量的运算符. 例如, 定义 $R=x-y$, 其中, x 和 y 是 2 个布尔矢量, 如果 $x_i=0$, 则 $R_i=0$, 否则, $R_i=x_i-y_i$
$\times$	一个输入 2 个布尔数值向量并返回数值向量作为输出的运算符. 布尔向量中的假条目被视为 0, 而真条目被视为 1. 例如, 可定义 $R=x \times y$, 其中, x 和 y 是 2 个布尔矢量, $R_i=x_i y_i$. 换句话说, 操作在逻辑操作中实现“and”
Σ	输入实数向量并将综合作为输出返回的运算符
$F(x,y,z)$	输入 3 个数值向量 x 、 y 、 z , 并返回实数值作为输出的函数. 如果合并 y 和 z , 则是实际值与 x 相比的适应度

对于第 0 帧, 可以检查是否正确分配了个人的关键点, 如果没有, 则可以手动将错误分离的关键点合并到正确的人, 然后使用 $F(\cdot)$ 函数来固定帧 1 作为起始, 使用先前固定的帧来固定所有视频帧的下一帧. 核心代码实现如下所示.

```

B=Af, the list of people in frame f;
C=Af+1, the list of people in frame f+1;
for j-th row in C do
    if some rows in B are not paired then
        match row i in C with row j in B according
to Eq.(2);
        z=B[i,:];
        y=C[j,:];
        if D(i,j)≤threshold then
            assign individual j in i's color;
            if j's keypoints are less than i's, then
                find the individual k in C that can best
complement the missing keypoints by minimizing
function F(·) according to Eq.(2);
                merge k's keypoints to j's;
            end
        else
            assign individual j a new ID;
        end
    else
        assign individual j a new ID;
    end
end
end

```

2.3 数据标准化

当监控大型场景时, 一些靠近监控摄像头的人在画面中会显得很大, 并且不同位置将影响个体的大小以及它们在每帧中像素方面的运动幅度, 从而降低判定准确度. 因此, 需要处理相关数据, 以尽量减少对检测算法的干扰. 在检查个体的打斗状态时, 为了标准化每对个体的运动信息, 将左侧个体的臀部中心视为原点, 将右侧个体的臀部中心视为点(1,0)进行归一化 x 轴. 然后使用较高人的肩膀和臀部之间的距离作为长度 0.5 来规范化 y 轴. 简而言之, 可以将较高人的髋肩距离定义为 y_d , 将 2 个人之间的臀部距离定义为 x_d , 即可规范化关键点: $(x', y') = (x / x_d, y / y_d)$, 归一化结果如图 4 所示. 对于每对个体而言, 可以使用归一化值 (x_0, y_0) 来预测其打斗状态.

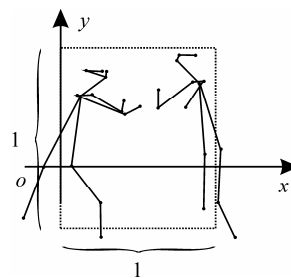


图 4 标准化关键点

2.4 动作建模

在匹配不同帧中的每个个体后, 使用递归神经网络模型检查每对个体是否正在打斗. 通过在 20 个连续帧中为每对个体选择肘部和手腕的标准化位置作为输入, 并反转帧的顺序来构建数据集.

在这种方法中, 使用由长短期记忆(LSTM)单位组成的递归神经网络(RNN)来对数据进行建模. 这样的 RNN 可以被认为是一系列 LSTM 单元, 每个单元将消息以链状传递给其后继者. RNN 具有从 18 个信息中学习的优势, 但是如果时间序列变长, 则没有长期记忆的 RNN 可能会忘记早期信息. 但幸运的是, LSTM 单元旨在明确地将长期信息保存在每个单元的状态向量中.

4 个神经网络层以一种非常特殊的方式进行交互. LSTM 的关键是类似于传送带的细胞状态, 信息可以通过皮带流动, 只有很小的线性相互作用^[8]. 同时, LSTM 确实能够移除或添加信息到细胞状态中, 并通过调节结构选择通过信息. 具体来说, 我

们将 h_{t-1} 定义为最后一个单元状态的输出, 并将 h_t 定义为当前时间单元状态的输出, 而 x_t 是输入, 选择长度为 S 时间段作为输入帧的数量, 变量 W 是 $S \times S$ 权重矩阵, 变量 b 是 $S \times 2$ 矩阵, 使用 sigmoid 函数 $\sigma()$ 激活单元格状态:

$$h_t = \sigma(W \times [h_{t-1}, x_t] + b).$$

使用一个完全连接的层, 用 h_s 作为输入, 以预测在这段时间内 2 个人是否在打斗, 通过最小化预测标签和地面实况标签之间的 softmax 交叉熵损失来训练模型, 图 5 则为相应的打斗模型。



图 5 打斗模型

3 实验

实验在具有 8.0 GB RAM, Intel 1.10 GHz CPU 的 PC 上进行, 并将人形检测算法与其他分类算法进行融合并比较效果, 其他方法包括广义递归神经网络(GRNN)、支持向量机(SVM)和 K-最近邻(KNN). 在训练过程中, 两类 SVM 是计算超平面, 它将点集分成两部分, 损失最小. 在实验过程中尝试了许多内核, 并最终选择具有更好性能的径向基函数内核. 实验中, 点是骨架的特征向量, 类标签表示实时状态. 测试中, 预测的置信度与点到平面的距离成比例. 在 KNN 中, 通过在最接近该查询点的 k 个训练样本中分配标签来对数据点进行

分类. 文中设 $k=5$, 并使用归一化距离来训练样本作为预测的置信度.

为构建数据集, 在校园内拍摄 40 个视频片段, 并提取 4 340 个数据片段, 每个数据包含 20 个清晰且连续的视频帧. 对于每种算法, 使用 3 110 个数据进行训练, 另外 1 230 个用于测试. 其中一半的训练片和一半的测试数据都是存在暴力行为的. 示例帧如图 6 所示.

基于骨架检测优势, 各种环境无关因素对判定几乎没有负面影响. 不同的分类算法具有不同的性能, 如图 7 所示, 本文提出的基于 RNN 的融合方法具有最大曲线下面积(AUC), 表明最佳性能. 图 7 中, 假阳性率(False Positive Rate, FPR)定义为 $FPR = FP / (FP + TN)$, 并且真阳性率(True Positive Rate, TPR)定义为 $TPR = TP / (TP + FN)$.

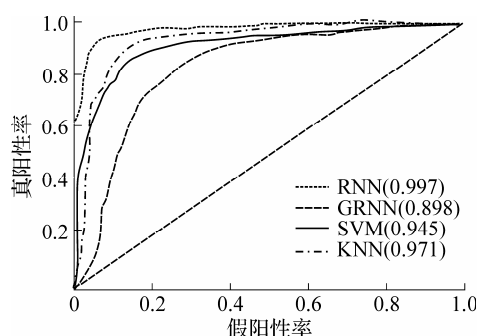


图 7 不同算法的 AUC 值比较

此外, 计算不同算法的平均运行时间如图 8 所示, 相对于其他算法, 基于 RNN 的融合算法更为



图 6 已在视频剪辑中使用过的场景

高效. 并且就平均时间来说, OpenPose 绘制正文关键点需要 0.142 s, 匹配个体并合并每个帧的对应部分需要 0.375 s, 可见新方法能够在每秒中进行至少一次判定.

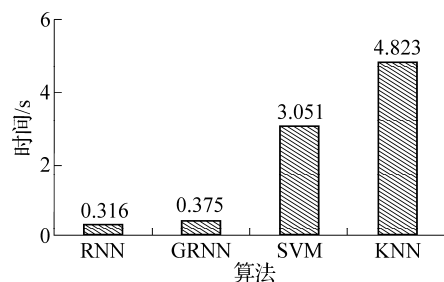


图8 运行每千次各算法所需时间柱状图

4 讨论

普通方法是基于人体关键点信息进行分析处理, 与事件发生的场景无关. 因此, 在一个场景中进行训练后, 新方法可以扩展到校园内的众多其他监控摄像机. 相应地, 在给定足够大数据的情况下, 新方法还可以检测其他攻击性操作. 实验结果显示, 基于数据库的训练, 新方法的准确性相比其他方法更高, 可以更为有效地判定是否有暴力行为的发生, 从而为遏制校园欺凌的进一步发展, 为防患于未然提供了技术保障. 未来可以通过其他信息(包括面部特征)进一步优化匹配过程, 以更准

确地匹配来自一系列图像的个体; 还可以引用面部/肩部方向的变化, 个体与摄像机的距离等参数, 将它们作为判断个体是否在打斗的参数, 从而进一步提高检测的准确率.

参考文献:

- [1] 徐俊丽. 初中校园欺凌现状调查及对策研究[D]. 开封: 河南大学, 2019.
- [2] 朱悦. 初中生法律信仰培育策略议析——以未成年人霸凌案为例[J]. 中小学德育, 2017(9):54-56.
- [3] 高鹏辉, 赵武峰, 沈继忠. 结合背景差分与光流法的人群状态突变检测[J]. 浙江大学学报(工学版), 2018, 52(4):649-656.
- [4] 王洪雁, 周梦星. 基于光流及轨迹的人群异常行为检测[J/OL]. [2020-03-23]. <https://doi.org/10.13229/j.cnki.jdxbgxb20190665>.
- [5] 胡瑞卿, 田杰荣. 基于光流法的运动目标检测算法研究[J]. 电子世界, 2019(5):58-61.
- [6] 邓立, 沈继忠, 高鹏辉. 结合光流法与信息熵的人群突发事件检测与判断[J]. 浙江大学学报(理学版), 2019, 46(1):39-47.
- [7] 杨明健, 黎镜林, 郭锐坤, 等. 基于 OpenPose 的人体睡姿识别实现与研究[J]. 物理实验, 2019, 39(8):45-49.
- [8] 王琛玮. 基于 Android 平台的心糖系统设计与实现[D]. 兰州: 兰州大学, 2017.

School bullying events detection based on deep learning

FU Shuibo¹, LI Gang²

(1.Zhenhai Jiaochuan Academy of Ningbo, Ningbo 315200, China;

2.Faculty of Electrical Engineering and Computer Science, Ningbo University, Ningbo 315211, China)

Abstract: The purpose of this study is to detect school bullying events in real time by monitoring network in campus. We establish a human posture model combined with a recursive neural network to detect and analyze the individual behaviors of paired students in the video, and an algorithm to identify bullying behaviors through a fully connected network. In addition, we build a database of more than 3 000 video clips from different parts of a campus and annotate the motion behaviors for training and testing purposes. Experimental results show that the presented algorithm is better than several commonly applied methods in accuracy and running speed.

Key words: motion detection; behavior identification; deep learning

(责任编辑 章践立)