计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2017,26(12):207−211 [doi: 10.15888/j.cnki.csa.006152]

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

基于 3D-CNN 的暴力行为检测①

周　智, 朱　明, Yahya Khan

(中国科学技术大学 信息科学技术学院, 合肥 230022)

摘　要: 大量的研究行为识别方法集中在检测简单的动作, 如: 步行, 慢跑或者跳跃等; 针对于打斗或者动作复杂的攻击性行为则研究较少; 而这些研究在某些监控场景下非常有用, 如: 监狱, 自助银行, 商场等. 传统的暴力行为识别研究方法主要利用先验知识来手动设计特征, 而本文提出了一种基于 3D-CNN 结构的暴力检测方法, 通过三维深度神经网络直接对输入进行操作, 能够很好的提取暴力行为的时空特征信息, 从而进行检测. 从实验结果可以看出, 本文方法能较好地识别出暴力行为, 准确率要高于人工设计特征的方法.

关键词: 动作识别; 暴力检测; 深度学习; 卷积神经网络

引用格式: 周智,朱明,Yahya Khan.基于 3D-CNN 的暴力行为检测.计算机系统应用,2017,26(12):207–211. http://www.c-s-a.org.cn/1003-3254/6152.html

Violence Behavior Detection Based on 3D-CNN

ZHOU Zhi, ZHU Ming, Yahya Khan

(School of Information Science and Technology, University of Science and Technology of China, Hefei 230022, China)

Abstract: A large number of research behavioral methods are focused on detecting simple actions such as walking, jogging, or jumping, while less research is on violence or aggressive behavior, but these studies are useful in some surveillance scenarios, such as: Prison, self-help banks, shopping malls and so on. Traditional methods of violent behavior recognition research mainly use a priori knowledge to manually design features. In this paper a violence detection method based on 3D-CNN structure is proposed. The three-dimensional deep neural network directly manipulates on the input, which can be a good extraction of violent behavior of time and space characteristics of information. It can be seen from the experimental results that this method can identify the violent behavior better than the characteristics of hand-craft features.

Key words: action recognition; violent detection; deep learning; convolution neural network

随着监控系统的大量使用, 视频数据出现爆发性的增长. 监控系统的作用是进行目标检测以及异常行为检测. 随着数据的急剧增长, 传统的依靠人工监控已愈发困难, 且效率低下. 因此, 依靠人工智能的监控系统的研究成为了热点, 其中, 对于人的暴力行为的检测是重要的研究方向.

由于暴力行为的动作比起简单的跑, 跳行为[1,2]要复杂很多, 所以也是相关研究中的难点. 目前, 针对于暴力行为检测, 许多通用的方法都是通过特征点提取, 比如说光流、梯度、颜色等, 在使用分类器如 SVM, HMM 等, 进行暴力检测. Nam[3]等人提出火焰、血液等特征来检测暴力行为. Bermejo[4]等人利用 STIP 对暴力行为进行分类. Tai[5]等人对光流向量进行计算, 进而检测暴力行为. Martin[6]等人使用多尺度的局部二相模式直方图进行暴力检测. Wang[7]使用了基于轨迹分析的暴力行为识别方法. 综上所述, 传统的暴力行为识别主要是采用基于人工设计特征的方法, 虽然识别准确率较高, 但是也具有某些缺陷, 如: 耗时较高, 易受噪声

① 基金项目: 中科院先导项目课题 (XDA06011203)
收稿时间: 2017-03-18; 修改时间: 2017-04-10; 采用时间: 2017-05-08

干扰依赖特定数据集等. 近年来, 以 CNN[8,9]为代表的深度学习算法取得了快速发展, Ji[10]等人首次提出了一种时空卷积神经网络用来进行人体动作识别. Karpathy[11]等人建立了一百万视频的行为分析数据集, 通过多种 CNN 结构训练视频, 进而来判断行为类别.

针对于此, 本文采用基于 3D-CNN 的暴力行为识别方法. 该方法基于深度学习, 无需手动提取特征, 通过 3D-CNN 模型自动学习特征, 识别暴力行为.

1 3D-CNN 模型

1.1 CNN 和动作表示

一般, 在视频中应用 CNN 的一个简单的方法是对每一帧图片用 CNN 来识别, 如图 1 所示. 但是传统的 2D-CNN 结构没有考虑时间维度上的特征信息. 因此, Ji[8]等人首次提出的 3D-CNN 模型用来进行动作识别. 通过在 CNN 的卷积层进行 3D 卷积, 从而能够在空间以及时间维度上都能学习有用的特征, 如图 2 所示.

3D-CNN 是将视频中的连续帧作为一个时空立方体, 以此作为 CNN 网络的输入, 用 3D 卷积核对时空立方体进行操作, 从而提取空间和时间上的特征信息. 选取不同的卷积核对立方体进行卷积, 就能得到多种时空特征.

1.2 3D-CNN 模型结构

Tran[12]等人提出了一种 3D 深度卷积神经网络的框架-C3D 模型. 本文提出了将 C3D 模型运用于暴力行为检测的方面, 并且在原始的 C3D 进行了改进, 从而能更有效地检测暴力行为. 本文模型结构如图 3 所示, 模型共有 8 个卷积层, 5 个最大池化层, 2 个全连接层, 最后加上一个 SoftMax 层. 所有的 3D 卷积核大小都是 3*3*3, 时间和空间维度的步长都为 1, Padding 为 1. 每个卷积层的卷积核个数可以在图 2 中看出. 每个池化层的滤波器大小都是 3 维的, 除了 Pool1 的滤波器大小是 1*2*2, 其他的 Pool 层滤波器大小都是 2*2*2.

网络的输入视频大小是 171*128*16, 通过对输入进行中心裁剪得到尺寸为 112*112*16. 在 Conv1a 层中采用 3*3*3 大小的卷积核作用输入层, 卷积核步长为 1*1*1, 激活函数为 ReLu 函数. 选取 64 种不同的卷积核, 这样共得到 64 个 Feature Map. 其计算过程如下:

$$x_{ij}^{xyz}=sigm\left(\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1}\omega_{ijm}^{pqr}x_{(i-1)}^{(x+p)(y+q)(z+r)}+b_{ij}\right)$$

在上述公式中, $sigm(\bullet)$表示 sigmoid 函数, ω_{ijm}^{pqr}表示卷积核与上一层的第 m 个特征图的连接权重, P_i、Q_i、R_i 是权重立方体的维度.

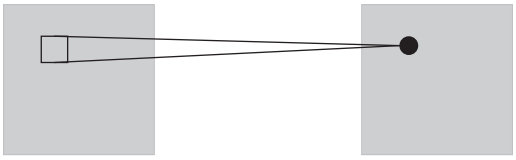

图 1 2D 卷积

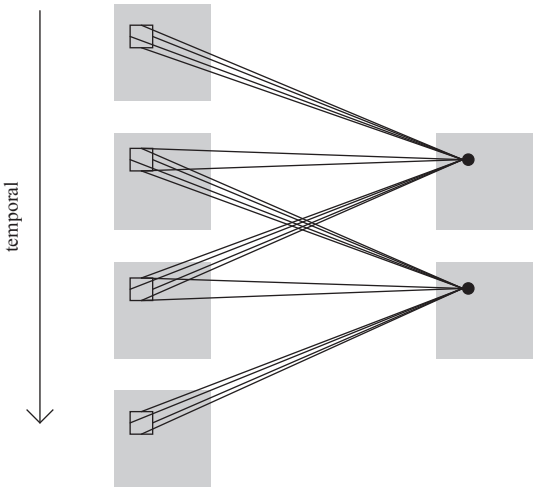

图 2 3D 卷积

在卷积层 Conv1a 后面是降采样层 Pool1, 采用 2*2*1 大小的滤波器对 Conv1a 层的每个特征图进行降采样, 步长为 2*2*1, 这样做可以使特征图变小, 简化网络的计算复杂度. 其计算过程如下所示:

$$a_j^l=\beta_j^l down\left(a_j^{(l-1)}\right)+b_j^l$$

其中, $down(\bullet)$表示降采样函数, 一般采取最大池化或平均池化两种操作方式, 通过降低上一层特征图的分辨率从而提高模型的鲁棒性. 在这里, 使用的降采样方式是最大池化, 选择图像区域的最大值作为该区域池化后的值, β_j^l表示连接的权重, b_j^l是池化后的偏置项.

同样地, 卷积层 Conv2a 和池化层 Pool2 所采用的的连接方式和计算方式的原理与 Conv1a 和 Pool1 相同, Feature Map 个数为 128 个. 随后的 3 个层数都是两个卷积层后面加一个池化层, Feature Map 个数分别为 256, 512 以及 256 个. 在 Pool5 层后面有两个全连接层, 全连接层神经元个数为 512 个和 100 个, 全连接层后面都接有一个 dropout 层来减轻网络过拟合, 最后一层是 SoftMax 层来进行分类.

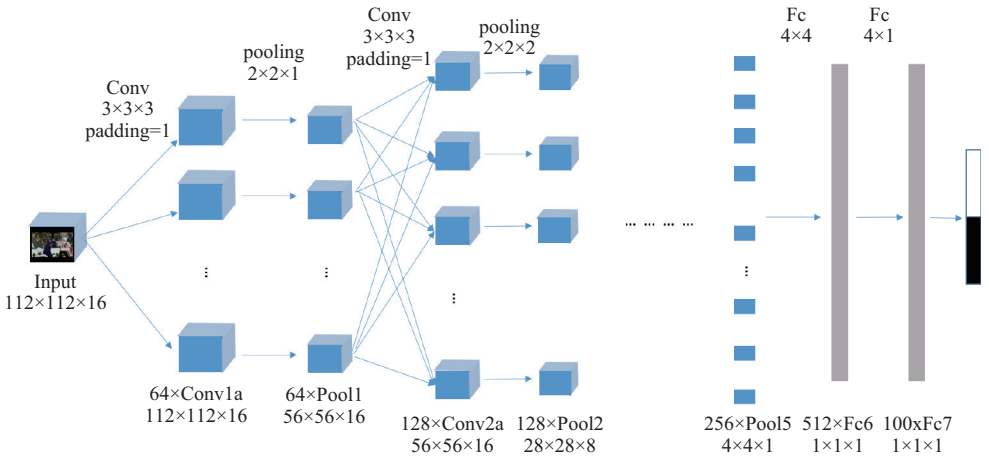

图 3 本文 3D-CNN 模型

1.3 模型融合

一般来说, 不同的输入可以训练得到不同的模型, 其预测的结果是不同的. 因此在本文中考虑不同的模型之间的组合会对结果产生影响. RGB 图主要反映图像的表观信息, 故可以提取图像的其它信息来更好地反映图像内容, 并以此作为模型的输入, 通过不同的输入构造多个不同的 3D CNN 模型, 在分类阶段, 进行模型融合, 计算每个模型的输出, 通过求平均等方法得到最终的预测结果.

光流信息能很好地反映运动目标的方向及速度信息, 可以通过提取图像的光流信息, 得到光流图谱.

图 4 光流图谱

在本章中采用 Lucas-Kanade 算法对运动区域计算光流, 该方法是稀疏光流的一种, 时间消耗较低. 该算法假设光流是局部平滑的, 即某像素点固定大小的邻域内光流不变. 令 $m\times m$ 固定区域内的光流为(μ,ν), 满足光流约束条件$I_x\mu+I_y\nu+I_t=0$, 可得:

$$\begin{bmatrix} I_{x_1} & I_{y_1} \\ I_{x_2} & I_{y_2} \\ \vdots & \vdots \\ I_{x_n} & I_{y_n} \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix} = - \begin{bmatrix} I_{t_1} \\ I_{t_2} \\ \vdots \\ I_{t_n} \end{bmatrix}$$

上式中, n 表示区域内的像素点数目 ($n=m^2$), I_x 和 I_y 表示区域内的光流变量的空间梯度, I_t 为区域内光流变量的时间梯度. 求解上述方程:

$$\begin{bmatrix} \mu \\ \nu \end{bmatrix} = \begin{bmatrix} \sum I_{x_i}^2 & \sum I_{x_i} I_{y_i} \\ \sum I_{x_i} I_{y_i} & \sum I_{y_i}^2 \end{bmatrix}^{-1} \begin{bmatrix} -\sum I_{x_i} I_{t_i} \\ -\sum I_{y_i} I_{t_i} \end{bmatrix}$$

由上式得到了光流的水平速度和垂直速度(μ,ν), 通过像素点的光流速度能够绘制光流图谱, 如图 4 所示. 在输入数据预处理完成之后, 我们将连续的 16 帧原始图片以及对应的 16 帧光流图片, 分别作为两个 3D-CNN 模型的输入, 经过训练之后, 分别得到两个最终的网络模型, 在测试时, 每个模型的 SoftMax 层都会输出每种行为对应的概率, 采用求平均值的方法, 计算最终的概率, 计算公式如下所示:

$$P(x)=\sum_{i=1}^{N}\frac{1}{N}P(x;M_i)$$

其中, x 表示输入的连续帧图片, $P(x;M_i)$表示第 i 种模型对每种行为的判定概率, $P(x)$ 表示模型融合之后的最终判定概率, 判定概率最大的行为就被认为是最终的输出结果.

2 实验结果与分析

在本节中, 为了评估模型的有效性, 我们在暴力行为数据集 HockeyFight 上进行测试, HockeyFight 数据集包含 1000 个冰球比赛的片段, 其中包括暴力视频和正常比赛视频各 500 个片段, 如图 5 所示. 同时, 我们也采用自己准备的 ATM 数据集进行实验, ATM 数据集同样也包含 1000 个 ATM 机取款的片段, 其中也包括抢劫暴力视频和正常取款视频, 如图 6 所示. 以下是实验结果的说明.

HockeyFight 数据集: 包括 1000 个视频片段, 每个片段包含连续 32 帧图片, 我们以连续 16 帧作为一个

样本, 共有 2000 个样本. 我们随机选择 800 个打斗样本和 800 个正常样本, 作为训练集, 剩余的作为测试集. 我们设置初始学习率为 0.03, batchsize 为 30, 每次随机批处理 30 个片段, 分别在不同迭代次数下, 进行了对比实验, 如表 1 所示.

图 5 HockeyFight 数据集视频片段

图 6 ATM 数据集视频片段

从表 1 可以看出, 在迭代次数为 8000 时, 检测准确率最高, 当迭代次数低于 8000 次的时候, 模型训练不够充分; 当高于 8000 次的时候, 模型会出现过拟合, 准确率都会下降. 图 7 表示本文 3D-CNN 模型的 ROC 曲线图, 可以看出本文模型能够有效地检测出视频中的暴力场景.

表 1 C3D 模型在 HockeyFight 数据集的准确率

学习率	迭代次数	准确率(%)
0.03	2000	89.9
0.03	3000	91.1
0.03	4000	91.7
0.03	5000	92.8
0.003	8000	93.8
0.003	10000	92.3
0.003	15000	91.3

同时, 为了进一步验证模型的有效性, 我们与多种手工提取特征的算法进行对比, 文献[4]中提出了两种行为特征描述子 STIP 和 MoSIFT, 并且在 Hockey 数据集上进行验证, 结果如图 8 所示.

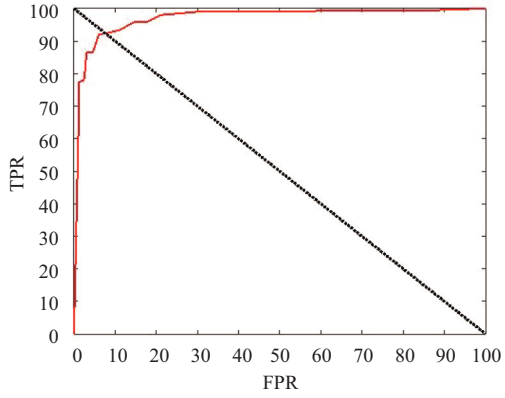

图 7 HockeyFight 数据集 ROC 曲线图

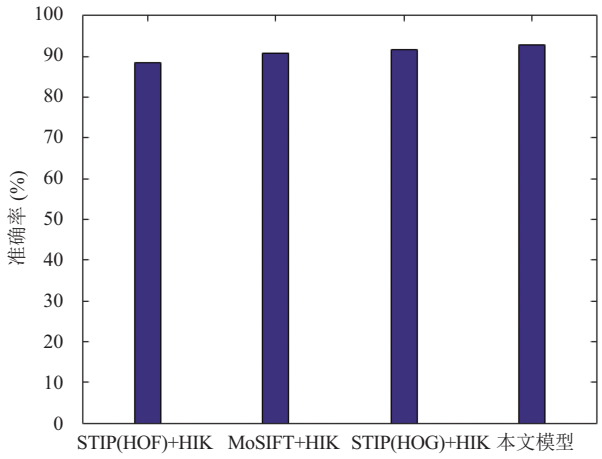

图 8 算法准确率对比

图 8 中, 本文模型在 HockeyFight 数据集上的准确率达到了 93.8%, 而文献[4]中的三种方法的准确率分别为 88.6%, 90.9%, 91.7%, 由上可知, 本章提出的模型检测准确率高于提出的所有方法. 此外, STIP 和 MoSIFT 特征都属于手工提取特征, 本文利用深度学习的方法, 直接对输入进行操作, 无需依靠经验手工提取特征, 并且耗时也要比 STIP 和 MoSIFT 等传统特征要好.

考虑到不同的模型以及模型组合对实验结果的影响, 我们也做了相关实验, 对比模型分别为: RGB 图像训练的模型, RGB+光流图 (FLOW) 训练的模型. 分别选取准确率最高的数据进行对比, 如表 2 所示.

从表中可以看出, RGB+FLOW 模型融合在一定程度上可以提高准确率, 最高准确率达到了 94.4%, 要高于使用 RGB 图像构建的 3D-CNN 模型的最高准确率.

可见合适的模型融合能够有效地提高识别准确率. 本文中只对比了 RGB+FLOW 模型的融合, 事实上, 也可以选择其它合适的模型进行组合.

表 2 模型组合在 HockeyFight 数据集的准确率

迭代次数	RGB模型准确率(%)	RGB+FLOW模型准确率(%)
2000	89.9	89.5
3000	91.1	91.3
4000	91.7	91.9
5000	92.8	93.4
8000	93.8	94.4
10000	92.3	92.1
15000	91.3	91.4

ATM 数据集: 场景是 ATM 机自助取款银行. 我们以连续 16 帧作为一个样本, 数据集中包含了 1500 个训练样本, 其中打斗样本 700 个, 正常样本 800 个; 500 个测试样本, 其中打斗样本 200 个, 正常取款样本 300 个. 我们设置学习率为 0.3, batchsize 为 20. 表 3 所示本文方法在 ATM 数据集上的实验结果.

表 3 C3D 模型在 ATM 数据集的准确率

迭代次数	准确率(%)
500	91.7
1000	93.9
1500	96.8
2000	91.5
2500	91.7

我们可以看出, 在迭代次数为 1500 次的时候, 准确率最高达到了 96.8%. 此外, 我们也采用 STIP(HOG) 方法对 ATM 数据集进行了验证, 选取效果最好的准确率, 将结果与本文方法作为对比, 如表 4 所示.

表 4 本文算法与 STIP 比较

算法	准确率(%)
STIP(HOG)+HIK	91.4
3D-CNN	96.8

通过表 3 和表 4 可以看出, 我们提出的算法在 ATM 数据集上的表现也要好于 STIP 算法, 因此, 本文的算法在暴力行为检测中要优于文献提出的三种手工设计特征: STIP(HOG), STIP(HOF), MoSIFT.

3 结语

本文提出了一种基于 3D-CNN 的暴力行为检测方法, 与传统的基于人工合计特征的暴力行为检测相比, 本文基于 3D 卷积神经网络自动提取时空特征, 检测效果要好于手工设计的特征, 也要好于 2D 维度的 CNN 模型. 另外, 本文方法还对不同模型的组合进行了对比实验, 实验结果表明合适的模型组合能有效地提高检测准确率. 随着相关视频数据的增长, 基于 3D-CNN 的方法在检测精度方面将更具优势.

参考文献

1 胡琼, 秦磊, 黄庆明. 基于视觉的人体动作识别综述. 计算机学报, 2013, 36(12): 2512–2524.

2 郑胤, 陈权崎, 章毓晋. 深度学习及其在目标和行为识别中的新进展. 中国图象图形学报, 2014, 19(2): 175–184.

3 Nam J, Alghoniemy M, Tewfik AH. Audio-visual content-based violent scene characterization. Proc. of 1998 International Conference on Image Processing. Chicago, IL, USA. 1998. 353–357.

4 Nievas EB, Suarez OD, García GB, *et al*. Violence detection in video using computer vision techniques. Proc. of the 14th International Conference on Computer Analysis of Images and Patterns. Seville, Spain. 2011. 332–339.

5 Martin V, Glotin H, Paris S, *et al*. Violence detection in video by large scale multi-scale local binary patterns dynamics. MediaEval 2012 Workshop. Pisa, Italy. 2012.

6 Hassner T, Itcher Y, Kliper-Gross O. Violent flows: Real-time detection of violent crowd behavior. Proc. of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Providence, RI, USA. 2012. 1–6.

7 Wang H, Kläser A, Schmid C, *et al*. Action recognition by dense trajectories. Proc. of 2011 IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO, USA. 2011. 3169–3176.

8 Geng YY, Liang RZ, Li WZ, *et al*. Learning convolutional neural network to maximize Pos@Top performance measure. arXiv:1609.08417, 2016.

9 Li QF, Zhou XF, Gu AH, *et al*. Nuclear norm regularized convolutional Max Pos@Top machine. Neural Computing and Applications, 2016: 1–10, doi: 10.1007/s00521-016-2680-2.

10 Ji SW, Xu W, Yang M, *et al*. 3D convolutional neural networks for human action recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221–231. [doi: 10.1109/TPAMI.2012.59]

11 Karpathy A, Toderici G, Shetty S, *et al*. Large-scale video classification with convolutional neural networks. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 1725–1732.

12 Tran D, Bourdev L, Fergus R, *et al*. Learning spatiotemporal features with 3D convolutional networks. Proc. of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 4489–4497.