doi :10.3969/j.issn.1671-1122.2017.12.010

# 一种基于三维卷积网络的暴力视频检测方法

宋伟[1]，张栋梁[1]，齐振国[2]，郑男[1]

（1. 中央民族大学信息工程学院，北京 100081；2. 北京交通大学电子信息工程学院，北京 100044）

**摘　要：**随着内容分发网络和视频转码技术的发展，网络流量呈现视频化趋势，互联网中充斥着各种非法特殊视频，危害社会公共安全，急需有效的检测算法。为探索深度学习理论在特殊视频检测上的应用，文章提出采用三维卷积网络框架进行暴力视频检测。相比于传统手工特征和 2D 卷积网络，该方法可以较好地保护视频帧序列在时间维度上运动信息的完整性，实现对暴力视频时空信息的有效表征。在暴力视频数据集 Hockey 上进行实验，取得了 98.96% 的准确率。实验结果表明该方法能够有效地检测暴力视频内容。

**关键词：**暴力视频检测；三维卷积网络；特殊视频

**中图分类号：**TP309.1　　**文献标识码：**A　　**文章编号：**1671-1122（2017）12-0054-07

中文引用格式：宋伟，张栋梁，齐振国，等 . 一种基于三维卷积网络的暴力视频检测方法 [J]. 信息网络安全，2017（12）：54-60.

英文引用格式：SONG Wei, ZHANG Dongliang, QI Zhenguo, et al. A Violent Video Detection Method Based on 3D Convolutional Networks[J]. Netinfo Security, 2017(12):54-60.

# A Violent Video Detection Method Based on 3D Convolutional Networks

SONG Wei[1], ZHANG Dongliang[1], QI Zhenguo[2], ZHENG Nan[1]

(1.*School of Information Engineering*, *Minzu University of China*, *Beijing* 100081, *China*; 2. *School of Electronic Information Engineering*, *Beijing Jiaotong University*, *Beijing* 100044, *China*)

**Abstract:** With the development of content distribution network and video transcoding technology, network traffic has a trend of being dominated by the video, and there are varieties of illegal special videos flooded the internet, endangering the social public security, so the effective detection algorithm is of great necessity. In order to explore the application of deep learning theory on special video detection, this paper proposes the use of 3D convolutional networks for violence video detection. Compared with traditional manual features and 2D convolutional networks, this method can well protect the motion information integrity of video frames in the time dimension, and realize the efficient characterization of spatio-temporal information. The experiment was carried out on the violent video dataset Hockey, achieving 98.96% accuracy. The results show that the method can effectively detect the violent contents of video.

**Key words:** violent video detection; 3D convolutional networks; special video

**收稿日期：**2017-9-1

**基金项目：**国家自然科学基金 [61503424]

**作者简介：**宋伟（1983—），男，湖北，讲师，博士，主要研究方向为图像处理、视频内容识别；张栋梁（1991—），男，山东，硕士研究生，主要研究方向为视频内容检测、视频行为识别；齐振国（1989—），男，山西，博士研究生，主要研究方向为信号处理、机器学习；郑男（1994—），女，山西，硕士研究生，主要研究方向为图像处理。

**通信作者：**宋伟 sw_muc@126.com

音频，但是在监控视频中音频是无法获取的

## 0 引言

通常，暴力内容检测指的是检测电影或电视中的暴力场景。随着维护社会公共安全需求的增加和安防监控技术的发展，智能视频监控也成为暴力内容检测的一个重要应用方向。当前，暴力内容检测面临严峻挑战，相比于视频行为识别，暴力内容检测的研究较为滞后，研究方法较为单一，不能满足行业日益增长的发展需求。

近年来，深度学习在视频分析领域得到了广泛的关注，尤其在视频行为识别方面，相关的算法研究工作十分火热。KARPATHY[1] 等人将卷积神经网络应用到视频分类上，并在其构建的大规模行为视频数据集 Sports-1M 上进行了实验。SIMONYAN[2] 等人提出了一个分别采用静态帧数据流和光流数据流的双路卷积网络结构进行视频行为识别，该网络能够捕捉帧间完整的外观信息和运动信息，首次成功验证了深度特征相比于传统特征的有效性。JI[3] 等人将传统卷积操作拓展到时间维度，提出了一种三维卷积网络用于视频行为识别；TRAN[4] 等人提出一种简单但有效的三维卷积网络学习视频的时空特征，该网络采用三维卷积和三维池化操作，在行为识别、相似行为标注、场景分类等数据集上取得了较好的实验效果；WANG[5] 等人提出了 TSN 网络，采用稀疏采样策略对视频长时序结构进行建模，进一步提高了识别结果。

然而，在视频暴力内容检测方面，采用深度学习方法的相关研究较少，大部分的研究工作是采用“特征 + 机器学习”，在获取一定准确性的同时，带来了操作步骤繁琐、特征抽取的时间消耗高、不能高效执行等诸多问题。针对这些问题，本文首次将采用三维卷积和三维池化的三维卷积网络应用到视频暴力内容检测上。

本文首先介绍了暴力视频检测的相关工作，然后介绍了三维卷积网络及相关实验设计，对实验结果进行了分析，最后分析了本文所提方法的优缺点，并展望下一步工作。

## 1 相关工作

传统的暴力内容检测方法主要基于多模态分类策略，有以下两种方式：1）基于单一暴力音视频特征；2）基于暴力音视频多模态特征融合。

在基于单一暴力音视频特征方面。PFEIFFER[6] 等人采用单一音频特征检测暴力事件，但由于音频特征经常掺杂不同噪声，会导致该方法误判率较高。CHENG[7] 等人采用基于高斯混合模型（Gaussian Mixture Model，GMM）和隐马尔科夫模型[8]（Hidden Markov Model，HMM）的分级方法检测音频中的枪击、爆炸和刹车事件。GIANNAKOPOULOS[9] 等人同样也提出了一种基于音频特征的暴力内容检测方法，从时域和频域提取多种音频特征。CLARIN[10] 等人介绍了一种采用 Kohonen 自组织映射网来检测视频每一帧上的肤色和血色像素，同时借助运动强度分析方法来检测视频中的血腥暴力行为。

在基于暴力音视频多模态特征融合方面，NAM[11] 等人提出采用火焰、血色等视觉特征以及音频特征进行暴力场景检测，是最早提出暴力视频内容检测的研究者之一。GONG[12] 等人提出一种借助低维视听觉特征和高维音频特征鉴别电影中潜在暴力内容的方法。LIN[13] 等人协同训练音视觉特征分类器，提出了一种弱监督暴力检测方法用于电影暴力场景检测。GIANNAKOPOULOS[14] 等人使用音频特征的统计特性及视频中运动方向的平均方差，联合 $K$ 近邻分类器检测视频暴力内容。

血液 和 火焰，局限性

以上研究工作多集中在依赖音频特征和血色等颜色特征来检测暴力内容。这些特征在电影中暴力内容检测上很有效，然而在现实的视频监控领域，音频和血腥画面很少出现，所以之后的研究多集中在视觉特征上。

近年来，HOG（Historgrams of Oriented Gradient）、HOF（Historgtrams of Optical Flow）、SIFT（Scale-Invariant Feature Transform）等局部特征描述子的流行促进了特征点提取方法和特征表示方法的发展。DATTA[15] 等人结合运动轨迹信息和人的肢体方向信息检测拳击、踢打、撞击等暴力事件。HASSNER[16] 等人提出了一种 ViF 描述子用于拥挤场景下的实时暴力内容检测。DENIZ[17] 等人提出一种采用极限加速模式作为主要特征快速检测暴力视频的方法。BERMEJO[18] 等人采用时空兴趣点 [19]（Spatio-temporal Interest Points，STIPs）和运动尺度不变特征变换 [20]（Motion Scale Invariant Feature Transform，MoSIFT），结合词袋模型（Bag-of-Words，BoW）和支持向量机 [21]（Support Vector Machine，SVM）进行暴力内容检测，但是这种方法仅计算检测到的兴趣点周围区域，判别力并不是很强，且词袋

模型忽视了特征间的空间关系。XU[22] 等人用稀疏编码方法替代词袋模型，进一步提高了 MoSIFT 在暴力视频检测上的表现效果。ROTA[23] 等人采用 iDT 方法 [24] 获取特征码本，定义人际空间关系描述相互间的暴力行为，该方法的不足在于过度依赖行人检测器且只分析两个人之间的交互行为，泛化能力欠佳。ZHANG[25] 等人将运动信息加入到改进的韦伯局部描述子 [26]（Weber Local Descriptor，WLD），提出 MoIWLD 算法，并结合稀疏表示分类器 [27]（Sparse Rrepresentation-based Classification，SRC），在暴力视频检测上取得了较好效果。

随着深度学习在人体行为识别领域上的广泛应用，有研究者将相关方法应用在视频暴力内容检测上。DING[28] 等人采用 9 层的三维卷积网络进行暴力视频内容检测，在 Hockey 数据集上取得了 91% 的检测准确率。该卷积网络采用三维卷积，但池化方法仍采用二维池化方法，不能有效地保护视频中的运动信息。DAI[29] 等人将基于双流卷积神经网络结合长短时序网络 [30]（Long Short-term Memory，LSTM），最终通过 SVM 分类的方法应用到暴力场景检测，实验结果表明该方法优于传统方法。ZHOU[31] 等人基于 TSN 网络结构，构建了新型卷积网络结构 FightNet 和一个暴力交互数据集 VID，将在 VID 上的预训练模型用在暴力视频数据集上，在 Hockey 数据集上取得了 97% 的准确率。

鉴于现有传统方法在检测结果上的不足和深度学习方法在视频检测上的高效，本文将采用三维卷积和三维池化的三维卷积网络应用在暴力视频内容检测上，并结合支持向量机和极限学习机等分类器，寻求更好的视频暴力特征表征方法。

自己写综述的时候，可以借鉴这种手法

## 2 三维卷积网络

卷积神经网络有三个结构上的特性：局部感受野、权值共享以及降采样。这些特性使得卷积神经网络具有一定程度上的平移、缩放和扭曲不变性，在图像识别领域取得了突破性进展。然而，卷积神经网络受限于二维输入，不能有效地处理视频帧序列。视频作为图像在时间上的序列组合，不仅包含空间维度的信息，而且包含时间维度的信息。采用卷积神经网络处理视频帧序列，本质上是对视频帧图像在空间维度上进行二维卷积和二维池化操作，无法表征视频时间维度上的信息。鉴于此，本文将卷积操作延伸到时间维度上，引入三维卷积网络模型，对视频帧序列进行三维卷积和三维池化操作，有效地表征视频时间维度上的信息，提高了处理视频帧序列问题的效率和准确率。

### 2.1 三维卷积

视频帧间具有一定的时间相关性，包含大量的运动信息。为了有效地对视频帧序列进行处理，就必须尽可能表征视频时间维度上的信息。三维卷积就是在空间进行卷积操作的同时，在时间维度上也进行卷积操作，三维卷积中的三维是指高度、宽度和时间维度，可以同时表征视频的时空动态信息。

三维卷积核和三维卷积操作的示例图分别如图 1 和图 2 所示。

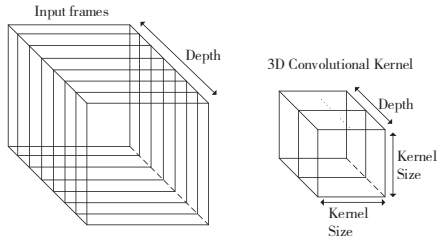


图 1 三维卷积核示例图

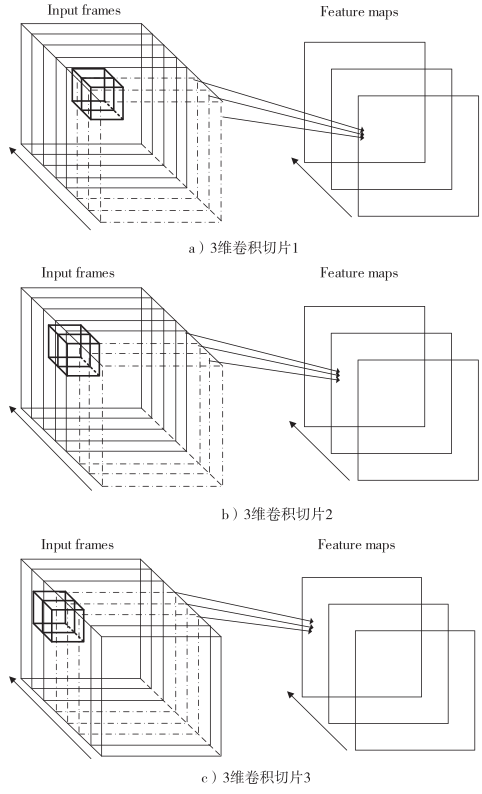


图 2 三维卷积操作示例图

三维卷积的公式如下：

$$v_{ij}^{xyz}=f\left(\sum_{d}\sum_{h=0}^{H_i-1}\sum_{w=0}^{W_i-1}\sum_{t=0}^{T_i-1}w_{ijd}^{hwt}v_{(i-1)d}^{(x+h)(y+w)(z+t)}+b_{ij}\right) \quad (1)$$

公式（1）中，$i$ 表示当前所在的卷积层，$j$ 表示该层的特征映射数量，$v_{ij}^{xyz}$ 表示在第 $i$ 层第 $j$ 个特征映射上 $(x,y,z)$ 位置处的激活值。$f(\cdot)$ 表示激活函数，其中，$H$，$W$，$T$ 分别表示三维卷积核的高度、宽度和时间维度上的大小。$w_{ijd}^{hwt}$ 表示卷积核的权重，$v_{(i-1)d}^{(x+h)(y+w)(z+t)}$ 表示第 $i$–1 层第 $d$ 个特征映射在 $(x,y,z)$ 处的激活值，$b_{ij}$ 表示偏置向量。

分别采用二维卷积和三维卷积处理视频序列，卷积操作前后的对比见图 3。

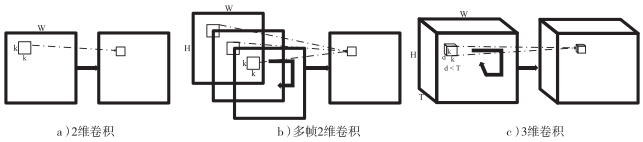


图 3 二维卷积和三维卷积的对比

从图 3 可以看出，在单帧图像上进行二维卷积操作，输出是一幅图像，在视频帧序列上进行二维卷积操作，输出仍然是图像，而对视频帧序列进行三维卷积操作，由于三维卷积可以表征时域信息，输出是一个帧序列。通过对比可以看出，三维卷积操作较好地兼顾了视频时域的上下文信息，对时空信息有相对较强的表征能力。

### 2.2 三维池化

为了减少模型计算量和避免出现过拟合现象，经过三维卷积操作获得特征之后，需要进行池化操作。同三维卷积类似，池化层也需要扩展到三维。三维池化不仅可以保持二维池化的空间平移不变性，在时间维度上也具有一定的不变性。经过三维池化处理后，网络在时间维度上的变化更加鲁棒。

三维卷积网络是将三维卷积和三维池化应用到卷积神经网络，借助三维卷积和三维池化操作，学习视频帧序列的运动信息，较好地保护了视频输入信号的时域信息。

## 3 实验及结果分析

本文是在 Ubuntu14.04 系统下，基于 Matlab R2015b 和深度学习框架平台 Caffe 完成的。

### 3.1 数据集

为了检验三维卷积网络在暴力视频内容检测上的有效性，本文在暴力视频检测公开数据集 Hockey 上进行了相关实验。Hockey[18] 数据集是由 Bermejo 等人在 2011 年构建，包含美国国家曲棍球联盟（National Hockey League，NHL）曲棍球比赛的 1000 个视频片段。数据集中的 500 个视频被人工标注为暴力视频，剩余 500 个视频被标注为非暴力视频。每个视频时长 2 秒，包含 41 帧（图像分辨率为 360×288）。本文随机选取 300 个暴力视频和 300 个非暴力视频作为训练集，剩余视频数据作为测试集。图 4 是从视频集中截取的部分帧画面，第一行是非暴力关键帧，第二行是暴力关键帧。

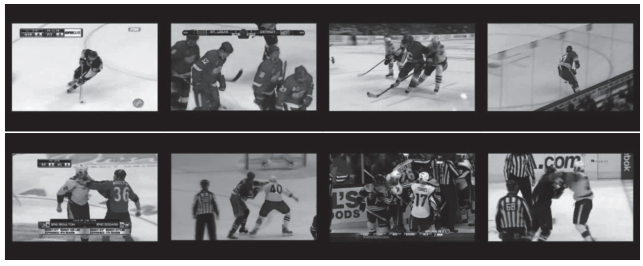

图 4 Hockey 数据集示例

### 3.2 网络结构设计

本文将训练集和测试集视频划分成 16 帧长无重叠视频段作为网络的输入，将原始帧图像大小调整为 128×171 像素。视频段的大小定义为 $c\times l\times h\times w$，其中，$c$ 为通道数量，$l$ 为视频段包含的帧的数量，$h$ 和 $w$ 分别为帧的高度和宽度，所以本文采用网络的输入视频段大小为：3×16×128×171。网络包括 5 个三维卷积层（3D Conv1~3D Conv5），5 个三维池化层（3D Pool1~3D Pool5），2 个全连接层（fc6, fc7）和一个 Softmax 输出层。5 个三维卷积层的卷积核个数依次为 64，128，256，256，256，其中三维卷积核大小为 3×3×3。所有的三维卷积层都采用零值填充，且卷积核步长为 1，使得经过卷积操作后输入和输出的大小没有改变。所有的池化层都采用最大池化方法，步长为 1，且核大小为 2×2×2，为了保存视频段的时域信息，第一个池化层的核大小设置为 1×2×2。最终，输入视频段经过三维卷积网络处理后得到 2048 维的特征向量。本文采用的网络结构如图 5 所示。

网络参数选取方面，对于基础学习率，本文选取大小为 [0.0002, 0.004] 区间的学习率进行网络训练，学习率更新规则采用“step”方式，stepsize=10000，最大迭代次数设为 20000。限于实验平台配置，本文实验采取的 batch size 为 10。在迭代次数区间 [0,20000] 内，每隔 1000 次保

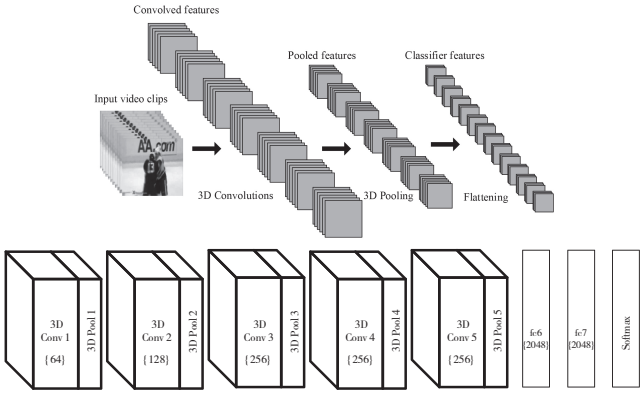


图 5 三维卷积网络结构图

存一次预训练模型，然后选取 loss 趋于稳定下的预训练模型进行特征提取。通过对比不同学习率下的最优检测准确率，选取最适合提取视频暴力内容特征的网络参数和预训练模型，具体的效果对比如表 1 所示。

表 1 不同学习率及迭代次数的最优检测准确率对比

| 学习率 | 迭代次数 | 检测准确率（%） | 学习率 | 迭代次数 | 检测准确率（%） |
|---|---|---|---|---|---|
| 0.0002 | 18000 | 98.01 | 0.0008 | 20000 | 98.85 |
| 0.0003 | 20000 | 98.60 | 0.0009 | 18000 | 98.83 |
| 0.0004 | 20000 | 98.64 | 0.001 | 19000 | 98.96 |
| 0.0005 | 19000 | 98.76 | 0.002 | 17000 | 98.35 |
| 0.0006 | 17000 | 98.84 | 0.003 | 20000 | 98.62 |
| 0.0007 | 19000 | 98.44 | 0.004 | 20000 | 97.32 |

从表 1 可以看出，当选取学习率为 0.001，迭代次数为 19000 时，视频段的检测准确率最高。因此，本文选用学习率为 0.001，迭代次数为 19000 的预训练模型进行后续实验。

### 3.3 分类器选择

当前，计算机视觉领域内的大量研究工作集中在采用深度学习技术提取图像或视频的特征向量并结合分类器对目标进行检测或识别，这些工作都取得了较好的效果。鉴于此，在直接使用三维卷积网络进行实验的同时，本文尝试采用“深度特征 + 分类器”的框架进行实验。用于特征分类的分类器有很多，其中性能比较好的主要有支持向量机（Support Vector Machine，SVM）和极限学习机 [32]（Extreme Learning Machine，ELM）等。因此，本文选取线性支持向量机和核极限学习机两种分类器进行实验。

为了探索对视频暴力内容的有效表征，本文将提取的深度特征和不同的分类器进行了组合，并对实验效果进行了对比。在 16 帧长的视频段经过三维卷积网络训练后，分别提取每个视频段的全连接层（fc6，fc7）特征和经 Softmax 层计算得到的 prob 层输出结果，对同一视频各视频段的 fc6 和 fc7 特征，利用 L2 范数对均值正则化，之后分别结合 SVM 和 ELM 分类器，并采用不同的核函数进行了实验，各核函数的表现效果如表 2 所示。

表 2 不同分类器的检测准确率对比

| 特征 \ 分类器 | SVM (Linear) | SVM (Polynomial) | SVM (RBF) | SVM (Sigmoid) |
|---|---|---|---|---|
| fc6 | 96.25% | 95.75% | 95.75% | 95.75% |
| fc7 | 96.00% | 96.25% | 96.25% | 96.25% |
| 特征 \ 分类器 | ELM (RBF) | ELM (Linear) | ELM (Polynomial) | ELM (Wavelet) |
| fc6 | 96.50% | 96.25% | 96.50% | 96.75% |
| fc7 | 96.25% | 96.25% | 96.25% | 96.25% |

从表 2 可以看出，采用 C3D+SVM 分类器方法时，使用 fc7 层特征取得了相对较好的表现效果。采用 C3D+ELM 分类器方法时，使用 fc6 层特征和 Wavelet 核函数取得了相对最好的表现效果。综合分析表 2 的准确率对比情况可知，采用 ELM 分类器的检测效果优于采用 SVM 分类器的检测效果。

对各视频段 prob 层输出的准确率，直接取其均值作为视频检测的准确率。本文将采用 fc6 层、fc7 层特征以及 prob 层输出结果的最优检测准确率进行了对比，具体的检测准确率对比如图 6 所示。

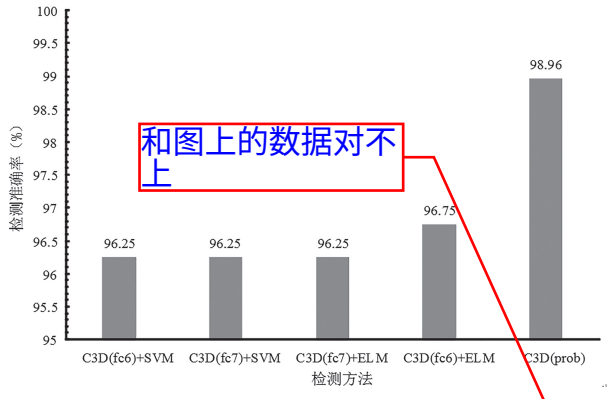


图 6 使用不同层特征及分类器的检测准确率对比

从图 6 可以看出，采用 C3D（fc6/fc7）+SVM 方法可以达到 96.25% 的检测准确率，采用 C3D（fc7）+SVM 方法可以达到 96.75% 的检测准确率，采用 C3D（prob）方法可以达到 98.96% 的检测准确率。相比于使用全连接层特征，使用 C3D 端到端训练取得了最好的效果。

结合表 2 和图 6 的对比实验结果，可以清晰地看出，采用三维卷积网络可以有效表征视频暴力特征。

### 3.4 实验结果分析

将本文方法同已有方法在 Hockey 数据集的检测准确

率进行了对比，对比结果如表 3 所示。

**表 3 各算法在 Hockey 数据集的检测准确率**

| 方法 | 检测准确率(%) |
|---|---|
| HOF+BoW[18] | 88.6 |
| HOG+BoW[18] | 91.7 |
| MoSIFT+BoW[18] | 90.9 |
| MoSIFT+KDE+Sparse Coding[22] | 94.3±1.68 |
| MoIWLD+KDE+SRC[25] | 96.8±1.04 |
| 3D-CNN[15] | 91 |
| **C3D+SVM** | **96.25** |
| **C3D+ELM** | **96.75** |
| FightNet[31] | 97.0 |
| **C3D(prob)** | **98.96** |

从表 3 可以看出，采用三维卷积网络框架的方法取得了当前最好的识别准确率，相比于文献 [25]，本文采用 C3D+SVM 方法和 C3D+ELM 方法，在检测准确率上分别有 0.55% 和 0.05% 的下降，采用 C3D(prob) 方法在检测准确率上有 2.16% 的提升。可以看出，本文方法整体检测效果要优于传统方法。相比于文献 [15]，本文采用 C3D+SVM 方法和 C3D+ELM 方法，在检测准确率上分别有 5.25% 和 5.75% 的提升，采用 C3D(prob) 方法在检测准确率上有 7.96% 的提升。文献 [15] 的网络结构采用的是三维卷积和二维池化，而本文采用三维卷积和三维池化，性能明显占优。相比于文献 [31]，本文采用 C3D+SVM 方法和 C3D+ELM 方法，在检测准确率上分别有 0.75% 和 0.25% 的下降，采用 C3D(prob)方法在检测准确率上有 1.96% 的提升。然而文献 [31] 采用的是在训练样本数为本文 5 倍左右的混合数据集 VID 上得到的预训练模型，因此不能直接对比测试结果。

通过分析表 3 的检测准确率结果，可以看出本文采用的三维卷积网络方法整体优于现有传统方法和深度学习方法。本文实验结果证明了采用三维卷积网络方法检测暴力视频内容的有效性。

## 4 结束语

本文将基于三维卷积网络的方法用于暴力视频内容检测，采用的三维卷积网络运用三维卷积和三维池化，较好地保护了动作在时间维度上运动信息的完整性，实现了对暴力视频时空信息的有效表征。在采用三维卷积网络检测暴力视频内容的同时，本文还尝试将采用三维卷积网络提取的深度特征同支持向量机及极限学习机相结合，对比采用不同分类器的检测效果，验证了“深度特征 + 分类器”方法的可行性。实验表明，本文方法在暴力视频公开数据集 Hockey 上取得了较好的检测准确率。考虑到本文方法采用 16 帧长的视频段作为网络输入，在处理长视频问题上会影响部分动作的完整性，因此下一步的工作将集中在研究自适应帧长的深度网络结构上。（责编 吴晶）

**参考文献：**

[1] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale Video Classification with Convolutional Neural Networks[C]//IEEE. 2014 IEEE Conference on Computer Vision and Pattern Recognition, June 24-27, 2014, Columbus, Ohio, USA. New York: IEEE, 2014: 1725-1732.

[2] SIMONYAN K, ZISSERMAN A. Two-stream Convolutional Networks for Action Recognition in Videos[J]. Advances in Neural Information Processing Systems, 2014, 1(4): 568-576.

[3] JI S, XU W, YANG M, et al. 3D Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[4] TRAN D, BOURDEV L, FERGUS R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]//IEEE. 2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE, 2015: 4489-4497.

[5] WANG L, XIONG Y, WANG Z, et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition[C]//IEEE. European Conference on Computer Vision, October 8-16, 2016, Amsterdam, the Netherlands. Cham: Springer International Publishing, 2016: 20-36.

[6] PFEIFFER S, FISCHER S, EFFELSBERG W, Automatic Audio Content Analysis [C]//ACM. the fourth ACM International Conference on Multimedia, November 18-22, 1996, Boston, Massachusetts, USA. New York: ACM, 1996: 21-30.

[7] CHENG W H, CHU W T, WU J L. Semantic Context Detection Based on Hierarchical Audio Models[C]//ACM. the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval, November 07-07, 2003, Berkeley, California, USA. New York: ACM, 2003: 109-115.

[8] RABINER L R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition[J]. Readings in Speech Recognition, 1990, 77(2): 267-296.

[9] GIANNAKOPOULOS T, KOSMOPOULOS D, ARISTIDOU A, et al. Violence Content Classification Using Audio Features[C]//SETN. the 4th Helenic Conference on Advances in Artificial Intelligence, May 18-20, 2006, Heraklion, Greece. Heidelberg: Springer Berlin Heidelberg, 2006:502-507.

[10]CLARIN C, DIONISIO J, ECHAVEZ M, et al. DOVE: Detection of Movie Violence Using Motion Intensity Analysis on Skin and Blood[J]. PCSC, 2005(6): 150-156.

[11] NAM J, ALGHONIEMY M, TEWFIK A H. Audio-visual Content-based Violent Scene Characterization[C]//ICIP. 1998 International Conferenceon Image Processing, October 4-7, 1998, Chicago, Illinois, USA. New York: IEEE, 1998:353-357.

[12] GONG Y, WANG W, JIANG S, et al. Detecting Violent Scenes in Movies by Auditory and Visual Cues[J]. Advances in Multimedia Information Processing-PCM 2008(1): 317-326.

[13] LIN J, WANG W. Weakly-supervised Violence Detection in Movies with Audio and Video Based Co-training[J]. Advances in Multimedia Information Processing-PCM, 2009(1): 930-935.
[14] GIANNAKOPOULOS T, MAKRIS A, KOSMOPOULOS D, et al. Audio-visual Fusion for Detecting Violent Scenes in Videos[C]// SETN. the 6th Hellenic Conference on Advances in Artificial Intelligence, May 4-7, 2010, Athens, Greece. Cham: Springer International Publishing, 2010: 91-100.
[15] DATTA A, SHAH M, LOBO N D V. Person-on-person Violence Detection in Video Data[C]// ICPR. the 16th International Conference on Pattern Recognition 2002, August 11-15, 2002, Quebec City, Quebec, Canada. New York: IEEE, 2002: 433-438.
[16] HASSNER T, ITCHER Y, KLIPER-GROSS O. Violent Flows: Real-time Detection of Violent Crowd Behavior[C]//CVPRW. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop, June 16-21, 2015, Providence, RI, USA. New York: IEEE, 2015: 1-6.
[17] DENIZ O, SERRANO I, BUENO G, et al. Fast Violence Detection in Video[C]// VISAPP. 2014 International Conference on Computer Vision Theory and Applications, January 5-8, 2014, Lisbon, Portugal. New York: IEEE, 2014: 478-485.
[18] BERMEJO N E, DENIZ S O, Bueno G G, et al. Violence Detection in Video Using Computer Vision Techniques[C]//CAIP. International conference on Computer analysis of images and patterns, August 29-31, 2011, Seville, Spain. Heidelberg: Springer Berlin Heidelberg, 2011: 332-339.
[19] LAPTEV I, LINDEBERG T. Space-time Interest Points[C]//ICCV. the 9th International Conference on Computer Vision, October 13-16, 2003, Nice, France. New York: IEEE, 2003: 432-439.
[20]CHEN M Y, HAUPTMANN A. MoSIFT: Recognizing Human Actions in Surveillance Videos[J]. Annals of Pharmacotherapy, 2009, 39(1):150-152.
[21]CHANG C C, LIN C J. LIBSVM: a Library for Support Vector Machines[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(3): 27.
[22]XU L, GONG C, YANG J, et al. Violent Video Detection Based on MoSIFTFeature and Sparse Coding[C]// Acoustics, Speech and Signal Processing (ICASSP). 2014 IEEE International Conference on Acoustics, Speech and Signal Processing, May 4-9, 2014, Florence, Italy. New York: IEEE, 2014:3538-3542.
[23] ROTA P, CONCI N, SEBE N, et al. Real-life Violent Social Interaction Detection[C]//ICIP. 2015 IEEE Image Processing, September 27-30, 2015, QuebecCity, Quebec, Canada. New York: IEEE, 2015: 3456-3460.
[24] WANG H, SCHMID C. Action Recognition with Improved Trajectories[C]//ICCV. 2013 IEEE International Conference on Computer Vision, December 3-6, 2013, Sydney, Australia. New York: IEEE, 2013: 3551-3558.
[25]ZHANG T, JIA W, HE X, et al. Discriminative Dictionary Learning with Motion Weber Local Descriptor for Violence Detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 696-709.
[26]CHEN J, SHAN S, He C, et al. WLD: A Robust Local Image Descriptor[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9):1705-1720.
[27]WRIGHT J, YANG A Y, GANESH A, et al. Robust Face Recognition via Sparse Representation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2): 210-227.
[28]DING C, FAN S, ZHU M, et al. Violence Detection in Video by Using 3D Convolutional Neural Networks[M]. New York: Springer International Publishing, 2014.
[29] DAI Q, ZHAO R W, WU Z, et al. Fudan-Huawei at MediaEval 2015: Detecting Violent Scenes and Affective Impact in Movies with Deep Learning[EB/OL]. http://ceur-ws.org/Vol-1436/Paper29.pdf, 2017-9-12.
[30]GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to Forget: Continual Prediction with LSTM[J]. Neural Computation, 2000(1): 2451-2471.
[31]ZHOU Peipei, DING Qinghai, LUO Haibo, et al. Violent Interaction Detection in Video Based on Deep Learning[C]//Jiangsu Optical Society,Southeast University. The Optical Society of America. 6th Conference on Advances in Optoelectronics and Micro/Nano-Optics, AOM 2017, April 23-26, 2017, Nanjing, China. Bristol: IOP Publishing, 2017: 012044.
[32] HUANG G B, ZHU Q Y, SIEW C K. Extreme Learning Machine: Theory and Applications[J]. Neurocomputing, 2006, 70(1): 489-501.