一种新的基于三维卷积共生梯度直方图和多示例学习的特殊视频检测算法

宋 伟[1)] 任 栋[1)] 于 京[2)] 齐振国[2)]

[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 已有的基于梯度方向直方图信息的视频内容检测算法侧重在二维的视频帧上提取特征,忽略了视频内容在时间维度上的相关性.提取局部梯度间潜在的共生关系特征可一定程度上提高算法的检测准确率;同时,对相邻特征池化可有效减少特征降维过程中的信息丢失.基于此,利用视频帧间结构信息通过卷积运算构建共生梯度直方图的三维结构,然后对相邻特征池化实现描述特征的有效降维,解决了忽略帧间信息影响识别准确率以及高维度特征难以训练的问题;将视频特征映射到多示例学习中的示例和包,非常容易地实现了对不同长度视频的检测.在公开测试数据集 Hockey、Movie 上进行测试,实验结果显示,Hockey 数据集上算法的检测准确率高于现有最优算法 3%,Movie 数据集上的检测准确率高于现有最优算法 0.5%,验证了新特征与算法的有效性.

关键词 视频内容检测;梯度方向直方图;多示例学习;卷积;池化;极限学习机

中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2019.00149

A New Special Video Detection Algorithm Based on 3D Convolution CoHOG and MIL

SONG Wei[1)] REN Dong[1)] YU Jing[2)] QI Zhen-Guo[2)]

[1)](*School of Information Engineering, Minzu University of China, Beijing* 100081)

[2)](*School of Electronic Information Engineering, Beijing Jiaotong University, Beijing* 100044)

Abstract Existing video content detection algorithms based on gradient direction histogram information are focused on the features extracted from the single two-dimensional video frames, ignored the correlation of the video frames on the time dimension. The frames in the video are inseparable whole. All consecutive frames could express true and complete semantics. The extracted information contained in video is inaccurate if only consider key frames. The correlation contain semantic information of video, is import for video content detection. And the potential symbiotic relationship between local gradient direction features is beneficial to the improvement of the algorithm accuracy. Just as important, pooling used in the adjacent features can reduce high-dimensional feature dimension, avoid losing hidden action information. Constructed 3D Conv-CoHOG feature by using the hidden structure information in video frames on the time dimension, and extending two-dimensional CoHOG features to three-dimensional features. Pooling operation on neighboring features reduced feature dimension effectively. This algorithm solved the problems of recognition accuracy reduction because of the inter-frame information

收稿日期:2017-07-07;在线出版日期:2018-07-16.本课题得到国家自然科学基金(61503424,61331013)、国家留学基金、中央民族大学一流大学一流学科("图像工程")、中央民族大学青年教师科研能力提升计划项目资助.**宋 伟**,男,1983 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为图像处理、视频内容识别、多媒体网络.E-mail:songwei@muc.edu.cn.**任 栋**,男,1990 年生,硕士,中国计算机学会(CCF)会员,主要研究方向为模式识别、视频内容检测.**于 京**,男,1971 年生,博士,教授,主要研究领域为图像处理.**齐振国**,男,1989 年生,博士,中国计算机学会(CCF)会员,主要研究方向为机器学习、视频识别.

neglect and the high computing complexity caused by high-dimensional features. Mapping video features to instances and bags corresponding to multiple-instance learning, dealing with video content detection problems for different lengths of videos simply. In this article, we introduced field of research and the importance of video violence content detection firstly. Then summarized the achievements of previous research, classified the findings of the research. All algorithms are divided into 3 categories, based on multi-modal features of audio and video and fused color feature, based on fusion of different action features, and the content detection algorithm based on neural network and unsupervised feature extraction. The most important part of this article is the introduction of algorithmic structure. We introduced the concept of HOG features and the extraction process, compared the extraction difference between HOG, CoHOG and Conv-CoHOG, also compared the extraction difference between HOG and HOG3D, and proposed the new special video content detection algorithm 3D convolution CoHOG extended from Conv-CoHOG. We compared the difference between the proposed new feature and the old features, such as computational dimension, feature dimension, and the relationship between adjacent features. In part 3.2, we introduced the framework of the new algorithm. In part 3.3 to part 3.7, we introduced the construction of feature extraction unit, the quantization of three dimensional gradients, extraction of Co-HOG3D, extraction of Conv-CoHOG3D, and the training of multiple-instance learning algorithm model. In part 4.1, described the two databases used in this experiment. In part 4.2, showed parameter setting and evaluation criteria. Then we analyzed the experimental results. In stage of training data, we used three classifiers, each classifier has a variety of implementations. When testing, compared the results of different features, analyzed the reasons for the different results, and analyzed the effectiveness of the new feature. In the end, we put forward effective solution on special video content detection. The highest detection accuracy on hockey and movie sets illustrated the availability of the proposed new algorithm on the special video detection. 3% higher than the existing optimal algorithm on Hockey data set, 0.5% higher than the existing optimal algorithm on Movie data set.

Keywords special videos detection; histogram of oriented gradient; multiple instance learning; convolution; pooling; extreme learning machine

1 引 言

移动智能终端设备的广泛使用,极大促进了我国网民数量的快速增长.互联网发展统计报告指出,在 2017 年 12 月网民数量已超 7.72 亿,占总人口的 55.8%,其中有 7.53 亿人通过手机上网,占 97.5%①.截至 2017 年第三季度,3G/4G 基站累计达到 447.1 万个,占比达 74.0%.第四代移动通信技术(4G)的发展、公共免费 WIFI 的建设,为公民上网提供了快速的渠道,公民浏览的内容不再局限于静态网页,而是更多的享受视频内容所带来的直观感受.思科也对互联网视频做了详细的调研,指出视频流量占据了 90%以上的网络流量,伴随着手机等智能终端的快速普及,移动视频流量在 2012 年首次超过了总移动流量的 50%,2016 年占全部移动流量的 60%,到 2021 年将至少达到 78%②.

开放的互联网环境使得视频的传播更加容易,视频是多形式的信息表达介质,其直观性为市民获取信息提供了更便捷的方式,但是复杂性更易隐藏违法内容,为相关部门的监管制造了严峻的挑战[1].暴力视频主要是指:包含打斗等行为,这些内容伤害人体或着造成人体的疼痛,进一步引起观者的不适[2].网络视频因传播方便,更易包含暴力、血腥等

① 第 41 次《中国互联网络发展状况统计报告》. http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/201803/P020180305409870339136.pdf 2018,03,05

② Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016—2021 White Paper. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html 2017,2,7

画面，若不能有效管制，任其肆意散播，特别是对于辨别能力差的青少年，极易诱发盲目模仿，影响心智发育，甚至诱发犯罪.

暴力视频的无序传播危害极大，传统人工识别过滤等方式效率低下且过滤处理存在滞后性[3]，急需研究适用于机器自动识别快速检测且具有高准确率的系统. 视频是多模态特征的融合，包含复杂的动作信息和情感信息，因此，针对视频的特殊内容检测比基于单视频帧的检测更为复杂.

本文主要研究特殊视频中的暴力内容，寻找高效、准确的视频暴力内容检测算法，以期过滤网络中无序传播的暴力视频，从而清除相关特殊内容，创造干净有序的视频传播渠道. 同时研究成果也为智能监控等领域的发展提供助力，为用户提供高速纯净的网络环境.

2　相关工作

与数字图像相比，多模态、异构、多尺度的特点使得特殊视频的检测更加复杂，难度也更高. 以往暴力视频的研究主要有：(1) 基于多模态的音视频特征和相关颜色特征融合[4-7]；(2) 基于不同动作特征融合的暴力视频内容检测算法[8-21]；(3) 基于神经网络的无监督特征提取的内容检测算法[22-29].

视频暴力内容检测研究的初期，音频特征、图像特征及其组合得到较多的关注. Giannakopoulos 等人[4]在频域和时域空间共提取了 7 种音频特征以实现对视频暴力内容的检测. Clarin 等人[5]使用自组织映射识别视频中的皮肤和血液区域，通过对动作强度的计算，检测有流血且含大幅度动作的暴力内容. Nam 等人[6]使用音频特征对尖叫、爆炸等内容进行检测，融合图像特征检测出的火焰、血液等内容，实现视频的暴力内容检测. Lin 等人[7]利用多种特征分别训练动作、声音、颜色检测器，多种检测器共同实现暴力内容分类.

之后，Datta 等人[8]在视频中计算加速度运动特征，识别视频中的打斗场景，完成暴力内容的检测. Wang 等人[9]使用稠密轨迹、支持向量机(Support Vector Machine，SVM)和慢特征分析(Discriminative Slow Feature Analysis)实现暴力内容检测. Chen 等人[10]探索了局部动作描述算子及视觉词袋方法在视频暴力内容检测方面的应用，并且取得了理想的检测结果. Giannakopoulos 等人[11]得到音频特征的统计数据，同时在运动方向上进行平均方差的计算，使用 K 近邻分类器实现数据的分类.

近年来，在动作识别领域一些局部动作特征得到较多应用，如方向梯度直方图(Histogram of Oriented Gradient，HOG). 暴力内容的检测也对这类算子做了大量的研究. Nievas 等人[12]使用梯度方向直方图(HOG)、方向光流直方图(Histograms of Optical Flow，HOF)和动作尺度不变特征变换[13](Motion Scale Invariant Feature Transform，MoSIFT)作为动作特征，同时使用视觉词袋模型(Bag of Visual Word，BoVW)、支持向量机(SVM)进行特征的建模与识别. Kataoka[14]利用稠密轨迹的方法，加入扩展的 CoHOG 特征，使用融合的特征完成视频暴力内容的检测. Xu 等人[15]提取了动作的 MoSIFT 特征，特征降维方面利用了核密度估计(Kernel Density Estimation，KDE)和稀疏编码(sparse coding)，在公共测试数据集 Hockey 上的检测准确率为 94.3%. Peng 等人[16-17]使用稠密轨迹描述视频中人物运动的边界，以 CoHOG 等特征表征视频的动作信息. 文献[18-21]也使用 HOG、HOF、LTP(Local Trinary Patterns)、MoSIFT、LaSIFT 等动作特征或多种特征的组合来描述视频信息，检测效果相比于其它特征得到较大的提升.

随着深度学习理论在计算机视觉等多个领域的广泛应用[22-24]，Ding 等人[25]首次将 CNN 网络应用到视频暴力内容的检测，文章利用 9 层(7 隐层)的卷积神经网络进行暴力内容检测，在 Hockey 数据上的检测准确率为 91%. Wang 等人[26]提出的视频时序分割网络(Temporal Segment Network，TSN)，将完整长视频分为若干视频段，分别输入到时间和空间特征提取网络，最后将空间性判定和时间性判定融合得到最终类别. Zhou 等人[27]使用 TSN 构建了一种用于暴恐视频识别的网络框架 FightNet，在 Hockey 和 Movies 数据集上均取得高准确率. Feichtenhofer 等人[28]提出 Convolutional Two-Stream Network Fusion 将单帧的图像信息和帧与帧之间的变化信息进行融合，从而实现时空信息的互补. 此外，Tran 等人[29]提出的一种鲁棒高效的三维卷积网络模型(C3D)，通过三维卷积核可同时提取视频空间和时间特征，只需将所得特征输入到线性分类器就可以得到高准确率的识别结果，也为特殊视频检测提供了一种新思路.

文献[14，16-17，30]使用 Co-HOG 及其扩展特

征提取视频的动作信息，提取的梯度共生关系充分描述了相邻梯度的关系，但特征整体维度较高. 视频帧与视频帧之间存在的时空结构信息，帧平面中相邻特征点间存在的共生关系，若能准确描述，可以实质性的提高特征子的描述力. 在三维空间中，本文进一步扩展 Co-HOG 特征构建 Co-HOG3D 特征，通过提取相邻帧间的梯度联系，得到时间维度上相邻特征的相关信息. 因从三维空间提取动作特征，故特征维度也数倍增加，新特征在提高算法准确率的同时增加了特征计算的计算量，使模型训练的复杂度变高. Conv-CoHOG 特征是添加了卷积和池化操作对 CoHOG 特征的扩展，在充分提取特征信息的同时，保证了特征的相对低维度，具有较强的动作描述能力[31]. 基于以上分析，考虑到视频所具有的特殊结构以及内容的复杂性，利用多示例学习算法中示例、包概念与视频相关概念的对应性，本文构建了更适合视频暴力内容检测的新算子 Conv-CoHOG3D 及融合多示例学习的新算法. 主要贡献如下：

(1) 结合视频的特殊结构，构造词包模型，提出了使用多示例学习算法进行视频暴力内容检测，解决了现有分类算法要求特征尺寸一致的问题；

(2) 针对高维且大量的卷积特征，将池化运算应用于相邻特征可以快速降低特征的维度，同时保证特征的准确性，构造出了一种新的特征描述子 Conv-CoHOG3D；

(3) 在 2 个公共测试数据集[32]上验证了新提出的 Conv-CoHOG3D 特征对视频时空信息的刻画能力和在视频暴力内容检测问题上的有效性.

3 算法基本框架

3.1 HOG 特征

HOG 是由 Dalal 等人[33]提出的一种梯度统计算子，对光照和形变有较强的鲁棒性. HOG 以梯度或边缘方向的统计描述物体的边缘，进而描述物体的轮廓. 特征提取的前期需要对图像进行分割，被分割的区域又进一步划分为更小的块(Block)，Block 继续分割为 Cell，用每个区域提取出的方向梯度直方图表征本区域，整合图像中所有直方图完成对整图的表征，过程如下：

(1) 图像灰度化；

HOG 特征主要计算相邻像素的梯度，以梯度对图像进行表征，关注的是像素值间的变化，所以图像的颜色信息、光照强度等可以忽略. 灰度化的图像可以使颜色信息和光照的影响得到削弱，同时图像由 3 层矩阵转换成 1 层，特征提取加快. 原图像与灰度图的对比如图 1 所示.

图 1 原图与灰度图

(2) Gamma 矫正法压缩处理灰度图像；

像素值间的细微差异在一定程度上可影响梯度的计算，进而影响特征的表征，这种影响可以通过压缩处理得到弱化. 压缩处理时，通常取 Gamma 的值为 1/2，其中 $I(x,y)$ 表示 (x,y) 处的像素值.

$$I(x,y)=I(x,y)^{\text{gamma}} \tag{1}$$

(3) 计算图像梯度；

经常使用的算子有 Sobel、Laplacian 等，梯度方向和幅值可由这些算子得到. 早期的研究中，Dalal 等人提到简单的 $[-1,0,1]$ 及 $[1,0,-1]^{\mathrm{T}}$ 算子(即式(2))效果反而比复杂梯度算子好，像素点的梯度大小和方向利用式(3)、(4)计算得到. Gamma 压缩后计算得到的梯度图会更清晰，与未使用 Gamma 压缩的对比如图 2 所示.

图 2 使用 Gamma 与未用 Gamma 的梯度图对比

$$\begin{aligned} G_x(x,y)&=I(x+1,y)-I(x-1,y),\\ G_y(x,y)&=I(x,y+1)-I(x,y-1) \end{aligned} \tag{2}$$

$$G(x,y)=\sqrt{G_x(x,y)^2+G_y(x,y)^2} \tag{3}$$

$$\alpha(x,y)=\tan^{-1}\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \tag{4}$$

(4) 图像分隔为块 Cell；

块 Cell 的划分中，其大小是任意的，不同的场景中，使用的大小不一致，一般为 6～8 像素. 图 3 显示了块重叠与非重叠的差异.

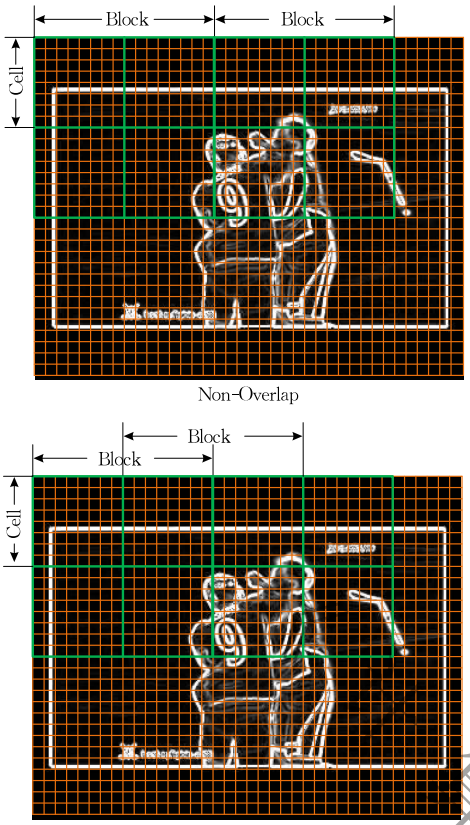

图 3　Cell 的划分

(5) Cell 中计算方向梯度直方图；

一个 Cell 对应一个梯度直方图，梯度的方向分为：0°～360°、0°～180°. 如图 4 所示.

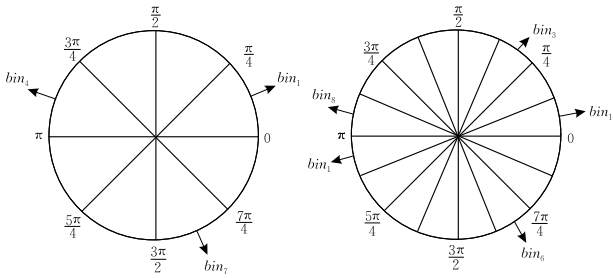

图 4　梯度方向的划分

(6) 相邻多个 Cell 构建 Block 特征；

Block 的 HOG 特征由所有 Cell 的特征向量串联得到. Block 的形状可以是矩形或星形.

(7) Block 特征归一化；

不同的场景，由于光照等因素的变化，即使同一物体计算得到的梯度也有差异，为减弱这种影响，需要对直方图进行归一化. 特征归一化后如图 5 所示. 常用的 4 种归一化方法如式(5)所示：

$$\text{L2-norm},\ V=V/\sqrt{\|V\|^2+\varepsilon^2},$$

$$\text{L2-Hys},\ V=(V/\sqrt{\|V\|^2+\varepsilon^2})/\sqrt{\|(V/\sqrt{\|V\|^2+\varepsilon^2})\|^2+\varepsilon^2},$$

(先进行 L2-norm，限制 V 的最大值到 0.2，再重新标准化)

$$\text{L1-norm},\ V=V/(\|V\|^1+\varepsilon),$$

$$\text{L1-sqrt},\ V=\sqrt{V/(\|V\|^1+\varepsilon)} \tag{5}$$

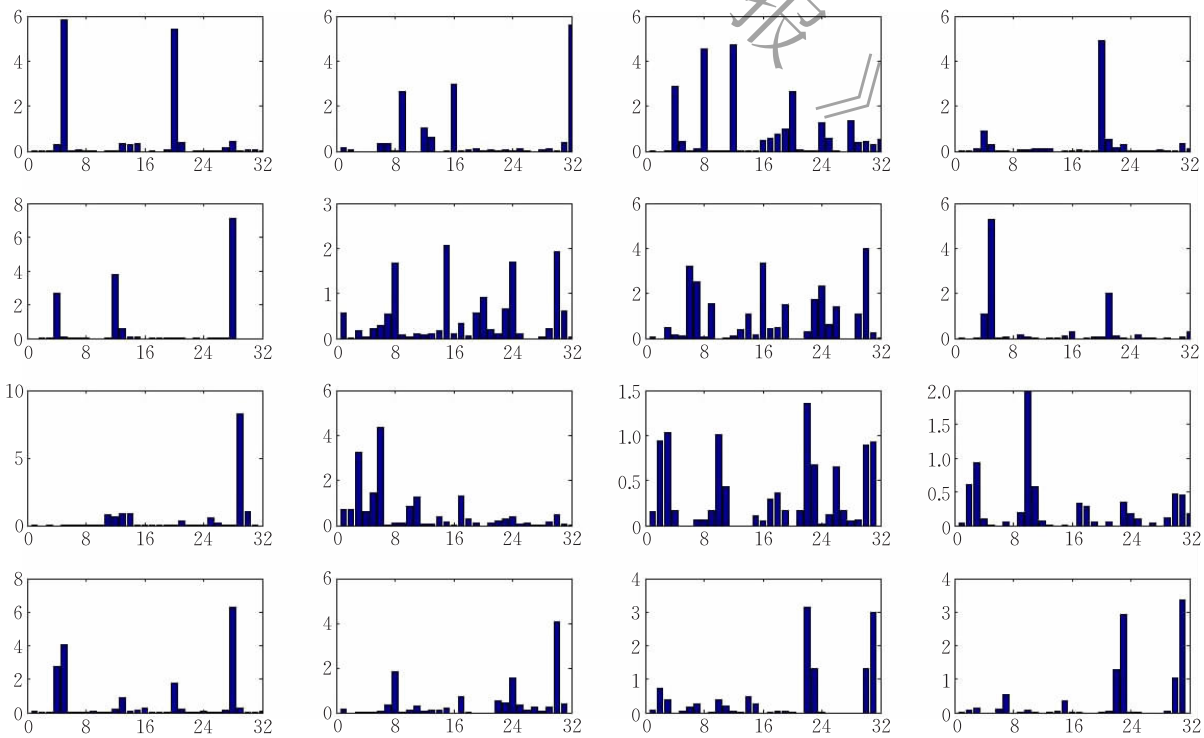

图 5　归一化的 HOG 特征

V 由 Block 中的直方图组合而成. $\|\boldsymbol{v}\|^k$ 表示向量 $\boldsymbol{v}$ 的 k 阶范数,k 取 1 或 2,ε 是很小的常数.

(8) 整幅图像的表征.

所有 Block 计算得到的 HOG 特征组合为一个特征向量,表征整幅图像. 自上而下排列所有的 HOG 特征,其横向显示如图 6.

图 6 归一化的 HOG 特征

3.2 算法简介

Co-HOG 特征是 HOG 特征在梯度对上的扩展形式,相关研究表明 Co-HOG 对特征的描述力更准确. 鉴于 Co-HOG 特征在局部信息描述方面的优势,针对视频的时空结构,将 Co-HOG 特征扩展到三维空间,利用卷积运算提取视频动作信息的 Co-HOG3D 特征,之后采用池化降低特征维度,得到新的特征 Conv-CoHOG3D,最后用 MIL 构造数据模型,检测视频中的暴力内容. 新特征 Conv-CoHOG3D 相比于 CoHOG 特征具有更高的维度,不仅提取了视频帧平面上的动作信息,也能提取到时间维度上相邻帧所包含的动作变化信息;相比于 HOG3D 特征,新特征不仅考虑了单像素的梯度信息,而且计算了相邻梯度对的方向与强度关系,即在帧平面上计算梯度对关系,同时在时间维度上计算梯度对关系. 利用新特征进行视频暴力内容的检测,检测结果均高于 CoHOG 特征和 HOG3D 特征的结果. 其检测过程如图 7 所示.

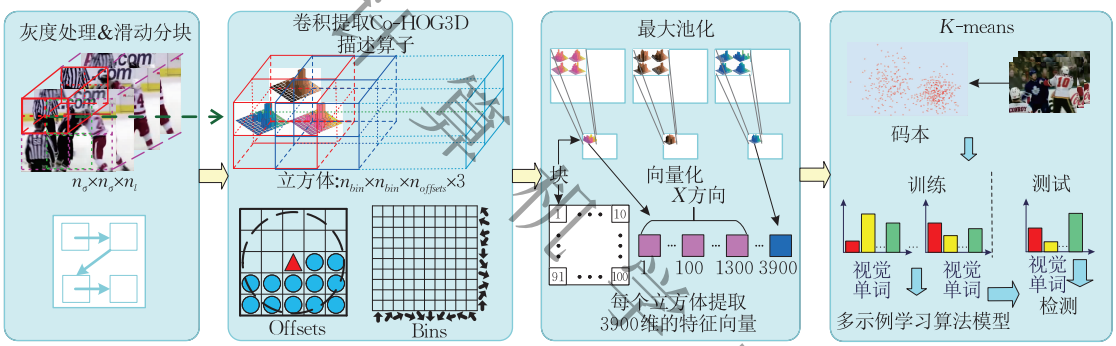

图 7 暴力内容检测流程

步骤 1. 视频 V 划分为重叠的视频片段 $V_1, V_2, \cdots, V_M$:

$$V = V_1 \| V_2 \| \cdots \| V_M \tag{6}$$

其中,M 由视频的时长和帧率决定,重叠率为 0.5,在时间效率上重叠方式的特征提取时间是非重叠方式的 2 倍;

步骤 2. 在帧平面上 V_i 细分成相互重叠的立方体 $C_{1,1}, C_{1,2}, \cdots, C_{m,n}$(重叠率为 0.5),在时间效率上重叠方式的时间是非重叠方式的 4 倍;

$$V_i = C_{1,1} \| C_{1,2} \| \cdots \| C_{m,n} \tag{7}$$

步骤 3. 立方体中提取三维梯度;

$$Cube_{\text{Gra3d}} = gradient(C_{a,b}),\quad a \in \{1,2,\cdots,m\}, b \in \{1,2,\cdots,n\} \tag{8}$$

步骤 4. 利用卷积操作对三维梯度构造 Co-HOG3D 特征;

$$Cube_{\text{Cohog3d}} = \sum_{p=1}^{3} \sum_{q=1}^{13} Conv(Cube_{\text{Gra3d}}) \tag{9}$$

步骤 5. 最大池化处理 Co-HOG3D 特征,减少特征的数目,降低特征维度;

$$ConvCoHOG3D = max\_pool(Cube_{\text{Cohog3d}}(i+x, j+y)) \tag{10}$$

其中,(i,j) 表示块的起始位置,$x, y \in \{1,2,3,4\}$.

步骤 6. 使用 K-means 算法,对部分视频的 Conv-CoHOG3D 特征聚类,得到包含 m 个类中心 $W_1, W_2, \cdots, W_m$ 的视觉词典 D;

$$Dic\{W_1, W_2, \cdots, W_m\} = K\text{-means}(ConvCoHOG3D) \tag{11}$$

步骤 7. 每个视频片段 V_i 中参照视觉词典 D 统计所有特征 $X_{m,n}(m=1,\cdots,M; n=1,\cdots,N)$ 的词频向量 $\boldsymbol{F}_i$. 根据多示例模型与视频结构的对应关系,词频向量 $\boldsymbol{F}_i$ 表示示例,视频 $V_1, V_2, \cdots, V_N$ 的词频向量 $\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_N$ 共同表征 MIL 的一个包;

$$Bag\{F_1, F_2, \cdots, F_N\} = hist_x\left(\sum_{m=1}^{M} \sum_{n=1}^{N} X_{m,n}\right),\quad x \in \{1,2,\cdots,N\} \tag{12}$$

步骤 8. 训练多示例学习分类器.

$$Model=MIL(\{F_{x,1},F_{x,2},\cdots,F_{x,N}\}),\quad x\in\{1,2,3,\cdots,S\} \tag{13}$$

S 为用于 MIL 模型训练的视频数量.

3.3　特征提取单元的构建

视频是复杂的数据载体，其信息不仅仅存储于单帧中，更多的语义内容在连续的帧中得以体现，传统基于关键帧提取特征的算法，无法完整表征视频的内容，且关键帧是否准确将对检测结果产生直接影响，基于此，本文结合视频时空结构特性，有效解决了单帧图像提取动作特征实现视频内容检测中出现的信息丢失问题. 在时间轴上，首先将视频分割成相互重叠的视频片段 $V_1,V_2,\cdots,V_M$，每个视频片段 V_i 的长度是 8 帧，视频片段间重叠的长度为 4 帧，在帧平面上将每个视频片段 V_i 进一步划分为相互重叠的 8×8 像素的立方体 $C_{1,1},C_{1,2},\cdots,C_{m,n}$，立方体间重叠的距离是 4 像素. 使用此重叠的方式，有效降低了动作被立方体分隔的风险. 在构建特征提取单元之前，已将视频各帧转换成矩阵存储，所以获取单个特征提取立方体的时间复杂度是 $O(1)$.

视频 $V=VF_1,VF_2,\cdots,VF_N$，VF_i 为视频的单帧，$V_i=VF_{4(i-1)+1},VF_{4(i-1)+2},\cdots,VF_{4(i-1)+7},VF_{4(i-1)+8}$. 若视频帧的分辨率为 $M\times N$，单帧定义如下：

$$VF_i=\begin{bmatrix}P_{(1,1)} & \cdots & P_{(1,N)}\\ \cdots & \cdots & \cdots\\ P_{(M,1)} & \cdots & P_{(M,N)}\end{bmatrix} \tag{14}$$

取第 i 帧中 x 行到 y 行，m 列到 n 列的区域 $VF_i(x:y,m:n)$：

$$VF_i(x:y,m:n)=\begin{bmatrix}P_{(x,m)} & \cdots & P_{(x,n)}\\ \cdots & \cdots & \cdots\\ P_{(y,m)} & \cdots & P_{(y,n)}\end{bmatrix} \tag{15}$$

对于视频片段 V_i，a 行 b 列立方体的定义如下：

$$C_{a,b}=VF_{4(i-1)+1}(4a-3:4a+4,4b-3:4b+4),\cdots,VF_{4(i-1)+8}(4a-3:4a+4,4b-3:4b+4) \tag{16}$$

3.4　三维梯度的量化

HOG3D 是 HOG 特征在三维空间的扩展，在提取 HOG3D 特征时，原 HOG 特征中块和胞元概念转换成三维空间中立方体的形式（如图 7 中左上角长方体）. 以立方体 $C_{a,b}(a=1,\cdots,M/4;b=1,\cdots,N/4)$ 为单位在 3 个方向提取三维梯度，$C_{a,b}$ 中对梯度方向进行量化，量化时常以正四、六、八、十二、二十面体五种正面体为量化标准. 通常坐标原点为正面体外接球的球心，球心到各个面中心的向量作为梯度方向量化的标准方向向量. Kläser 等人在文献[34]中使用 HOG3D 描述子描述视频的动作信息，并且提出了梯度强度以及梯度方向在三维空间量化的方法，文章使用 SVM 作为分类器实现视频内容的识别，取得了较高的识别准确率. 使用正二十面体对三维梯度进行方向及强度量化时，以正二十面体外接球的球心为坐标原点，使正二十面体每个面中心点的坐标为

$$(\pm1,\pm1,\pm1),\left(0,\pm\frac{1}{\varphi},\pm\varphi\right),\left(\pm\frac{1}{\varphi},\pm\varphi,0\right),\left(\pm\varphi,0,\pm\frac{1}{\varphi}\right) \tag{17}$$

其中 $\varphi=\dfrac{1+\sqrt{5}}{2}$ 为黄金比例，原点到各面中心点的向量为单位方向向量，利用单位方向向量将得到的三维梯度量化到二十个方向上，梯度的幅值作为量化时的长度投影到各个方向，按照投影长度和阈值决定梯度在对应量化方向的取舍，并做进一步的数值计算. 给定向量 $\bar{\boldsymbol{g}}_b$，量化过程如下：

(1) 记 $\boldsymbol{P}=(\boldsymbol{p}_1,\cdots,\boldsymbol{p}_{20})^{\mathrm{T}}$，其中 $\boldsymbol{p}_i(i=1,2,\cdots,20)$ 为梯度量化时参照的单位方向向量（列向量），值由式(17)给出；

(2) 依式(18)计算 $\bar{\boldsymbol{g}}_b$ 到各单位方向向量的投影：

$$\hat{\boldsymbol{q}}_b=\frac{\boldsymbol{P}\cdot\bar{\boldsymbol{g}}_b}{\|\bar{\boldsymbol{g}}_b\|_2} \tag{18}$$

(3) 对 $\hat{\boldsymbol{q}}_b$ 进行阈值处理，阈值 $t=\boldsymbol{p}_i^{\mathrm{T}}\cdot\boldsymbol{p}_j$，$\boldsymbol{p}_i,\boldsymbol{p}_j$ 为相邻向量，正二十面体中 $t\approx1.29107$，$\hat{\boldsymbol{q}}'_b$ 为处理后的向量，则

$$\hat{\boldsymbol{q}}'_{bi}=\begin{cases}\hat{\boldsymbol{q}}_{bi}-t, & \text{若 } \hat{\boldsymbol{q}}_{bi}-t>0\\ 0, & \text{其他}\end{cases} \tag{19}$$

(4) $\bar{\boldsymbol{g}}_b$ 在各个方向上的权重 q_b 为

$$q_b=\frac{\|\bar{\boldsymbol{g}}_b\|_2\cdot\hat{\boldsymbol{q}}'_b}{\|\hat{\boldsymbol{q}}'_b\|_2} \tag{20}$$

特征提取单元中量化三维梯度得到三维梯度方向直方图的时间复杂度是 $O(n)$.

3.5　Co-HOG3D 的求解

文献[17]中提及 3D 共生时空描述子，并对描述子的定义、求解过程、特点做了详细的介绍，同时，使用了多种描述算子提取视频中的暴力动作信息. 本文依据量化后的三维梯度，分别从 x、y、t 三个方向平面利用卷积运算按照不同的偏置计算每 2 个梯度之间的共生关系，并且根据梯度在量化方向上的投影长度将方向对量化成表示共生关系强弱的数值

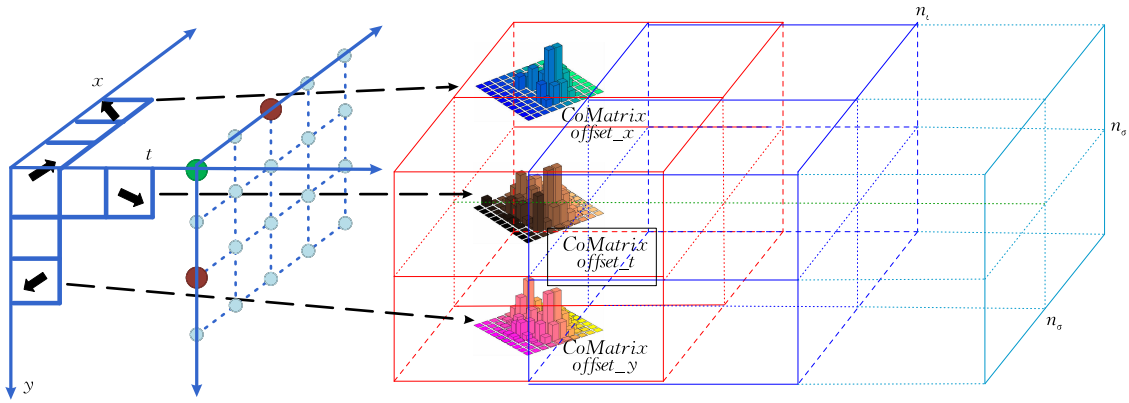

图 8 3D Co-HOG 特征的提取

(2 个梯度的投影长度直接求和),以此区分不同梯度所产生的梯度对差异,特征提取过程如图 8 所示. 文献[35]表明当使用 5×5 大小的 Offsets 时检测效果最佳,其偏移距离为 13 种,如图 7 中所示. Co-HOG3D 特征中既有单个像素的信息,且包含了梯度对的相关信息,与 HOG3D 特征相比,其维度更高,对视频动作信息的描述更详细. 特征提取时,在立方体 $C_{a,b}$ 的某个方向上首先将不同偏移距离的梯度共生图依次连接,之后,依次级联 t、x、y 三个方向的特征,最后得到的特征向量即为整个立方体 $C_{a,b}$ 的 Co-HOG3D 特征. 此过程的时间复杂度是 $3900O(n)$.

3.6 Conv-CoHOG3D 的求解

文献[31,35]首先提取 CoHOG 特征,之后使用池化操作,相邻特征块的主要信息由池化提取,池化使得总的特征数目减少,聚类、模型计算的难度降低. 如图 9 所示,本文为降低 Co-HOG3D 特征的维度,减少特征的数目,降低模型训练的难度,将池化操作应用在 Co-HOG3D 特征之后,池化通过无重叠的滑动方式进行,利用 Max Pooling 4×4 的窗口将视频特征降维为原来的 1/16. 在视频片段的帧平面上进行池化的时间复杂度是 $1/4O(n)$. 提取过程以伪代码的形式在算法 1 中列出.

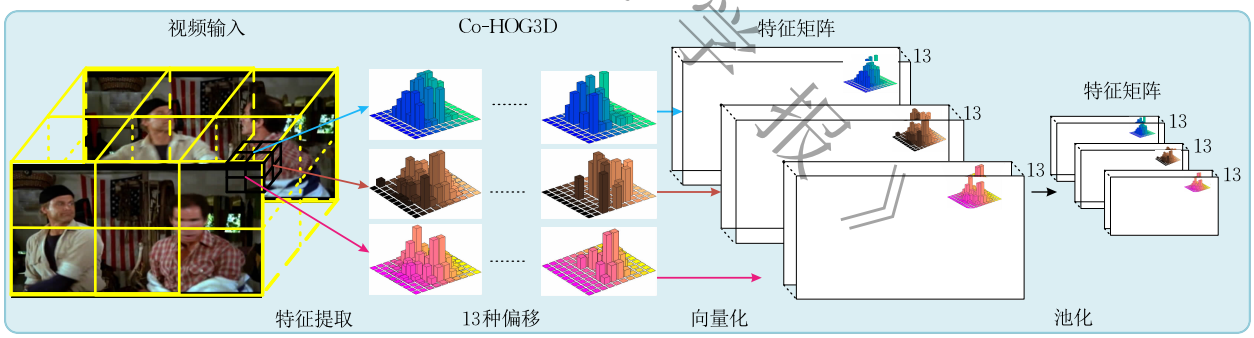

图 9 特征提取及卷积过程

算法 1. 提取 Conv-CoHOG3D 特征.

输入:梯度图 M,池化窗口长度 W,步长 S,池化方式 Max

输出:新特征 Conv-CoHOG3D

初始化:Co-HOG3D 特征图 $M_{i,m,n}$,其中 i、m、n 初始为 1,i、m、n 分别表示视频的片段号、帧平面中被分割视频的行号、列号.

```
FOR 视频片段 V_i DO
  FOR 帧平面中的一行(m=1) DO
    FOR 帧平面中的一列(n=1) DO
      提取与 M_{i,m,n} 相邻的 M_{i,m~(m+W-1),n~(n+W-1)},即 16 个卷积块.
      最大池化.
      Conv-CoHOG3D 特征存储.
      n=n+S.
    END
    m=m+S.
  END
END
```

3.7 多示例学习算法模型的训练

Dietterich 等人[36]在研究药物活性预测问题时最早提出了多示例学习模型,算法中包与示例的概念使得对包属性的判定更容易,特别是对正包的判定,只需存在一个正示例即可将此包标记为正包,而负包则要求所有示例均为负. 本文在计算时,每个视

频 V 对应于多示例学习中的一个包，每个视频片段 V_i 对应于模型中的一个示例. 在同一视频中，视频帧的分辨率相同，V_i 中立方体的数目一致，即 Conv-CoHOG3D 特征 $X_{m,n}(m=1,\cdots,M;\ n=1,\cdots,N)$ 的数目一致. 将部分视频的 Conv-CoHOG3D 特征用于 K-means 聚类，构造视觉词典 $Dic\{W_1,W_2,\cdots,W_m\}$，词典的大小可以通过 K-means 聚类算法的输入参数人为控制. 在每个视频片段 V_i 中，Conv-CoHOG3D 特征按照视觉字典 $Dic\{W_1,W_2,\cdots,W_m\}$ 计算距离最近的单词，将特征标记为与其最近的单词，之后统计视频 V_i 中所有单词类别出现的频数，以词频向量表征本视频片段 V_i. 多示例学习算法的最大优点是在特征分类前无需为得到归一化的特征而进行额外的计算，避免了可能由此造成的特征表征弱化. 本文 Citation-kNN[37] 算法采用 Hausdorff 距离，定义如下：

设 $A=\{a_1,a_2,\cdots,a_n\}$ 和 $B=\{b_1,b_2,\cdots,b_n\}$ 为两个包，$a_i,b_j\in R^d(i=1,2,\cdots,n;\ j=1,2,\cdots,m)$ 分别为包 A 和包 B 中的示例，包 A 和包 B 间距离 $minHD(A,B)$ 定义为

$$minHD(A,B)=\max\{h(A,B),h(B,A)\},$$

其中，$h(A,B)=\min_{a\in A}\min_{b\in B}\|a-b\|$.

训练视觉词典的算法复杂度取决于 K-means 聚类算法，在视觉词典确定的情况下，统计词频向量的算法复杂度是 $O(n)$. 多示例学习模型的训练过程如算法 2 所示.

算法 2.　训练多示例学习算法模型.

输入：Conv-CoHOG3D 特征，每个词典的单词数目 $N_1,N_2,\cdots,N_n$，近邻的参照数目 num

输出：视觉词典 $Dic\{W_1,W_2,\cdots,W_n\}$，视频特征检测模型 $model$

初始化：交叉检验时子集数目 $num\_Cross\text{-}validation=5$，$num=5$.

视频集中随机抽取部分视频的全部 Conv-CoHOG3D 特征.

FOR $N_1,N_2,\cdots,N_n$ DO

　训练并存储视觉词典 $Dic_1,Dic_2,\cdots,Dic_n$

END

FOR $Dic_1,Dic_2,\cdots,Dic_n$ DO

　FOR 视频片段 V_i DO

　　FOR 每个视频特征 DO

　　计算并标记词典中最近的单词 d_i.

　　$num\_of\_d_i=num\_of\_d_i+1$.

　　END

　　由 $num\_of\_d_i$ 计算词频特征 F_i，标记示例 F_i 的正负.

　END

　构造“包”特征 V_{bag}.

END

对视频的“包”特征 V_{bag} 进行训练.

保存模型 $Model_1,Model_2,\cdots,Model_n$.

4　检测结果分析

4.1　测试数据集

实验使用公共的 Hockey 数据集[12,32]、Movies 数据集[12,32] 作为基准测试数据. Hockey 数据集包含 1000 段视频，场景为冰球比赛，小视频的时间长度基本为 2 秒，共含 41 帧，视频帧的分辨率为 360×288. 所有视频被标记为包含暴力内容或正常，两种视频各有 500 段. 图 10 是 Hockey 数据集视频内容示例，第 1 行是包含暴力内容的视频截图，第 2 行是不包含暴力内容的视频截图.

Movies 数据集是由动作和非动作电影的片段组成，共 200 段，一半是含暴力内容的打斗视频，另一半是正常的，所有视频均做了相应的标记. 大部分暴力视频的分辨率为 720×576 像素，视频长度为 42 帧，少量视频的分辨率为 720×480，长度为 60 帧. 正常动作电影片段的分辨率为 720×480，帧数从 10 帧到 35 帧各异，每秒 25 帧. 图 11 为 Movies 数据集部分视频截图，第 1 行为打架场景截图，第 2 行为非打架场景截图.

图 10　Hockey 数据集内容示例

图 11 Movies 数据集内容示例

4.2 参数的设定和评价标准

Dalal 等人[33]对 Bins 数目进行了详细的实验分析，当梯度方向映射为 360°时，选取 18 Bins 或者 12 Bins 得到较低的假阳率；Kläser 等人[34]将三维梯度映射到正立方体，Bins 数目为 10 时得到较高的检测准确率，所以本文在量化三维梯度时，设置 Bins 的值为 10；按照文献[35]的表述，Offsets 选择 5×5 大小的区域效果最佳，所以本文在提取 CoHOG3D 特征时，同样使用 5×5 大小的偏移区域，即一共包含 13 种不同的偏移距离，如图 7 中所示. 池化窗口的设置主要考虑后期特征聚类时对计算机内存的要求，因为计算机内存的限制所以将窗口大小由常用的 2×2 增大为 4×4. 每种偏移对应一个 10×10 的二维矩阵；以非重叠的 4×4 窗口进行池化，16 个特征缩小为 1 个；提取 12 种视觉词典，分别含有 5、10、20、30、50、100、200、300、400、500、700、1000 个视觉单词.

本文采用公开数据库的标准测试准则，即准确率来评估算法效果，具体公式如式(21)：

$$准确率=\frac{真阳性样本}{测试总样本}\times 100\% \tag{21}$$

构造视觉词典时未使用视频集的全部特征，在 Hockey 数据集上，随机抽取暴力及非暴力视频各 100 个；在 Movies 数据集上，随机抽取暴力及非暴力视频各 20 个. 利用随机抽取的视频特征完成特征聚类进而构建视觉词典. 使用 MIL 作为分类器时，在 Hockey、Movies 数据集上的检测结果如图 12 所示.

4.3 实验结果分析

由图 12 可见，算法在 Movies 数据集上具有较高的检测准确率. 在 Hockey 数据集中，准确率随着词典单词数目的增加而上升，当单词数目少于 50 时，准确率增长迅速；当单词数目超过 50，准确率变化缓慢，但仍保持上涨趋势. 当单词数目较少时，视觉词典不足以充分表征视频的动作特征，当单词数目变大时，词典能够实现视频特征的准确区分，故随着单词数目的增加，准确率上升. 实验结果显示无需使用全部特征来构造视觉词典，即利用数据集中 20%的视频构造词典足以较准确的描述视频特征，进而获得较高的检测准确率. 同时，曲线的变化表明在使用词包模型时，训练能够充分表征视频特征的视觉词典有利于提高最终的检测结果.

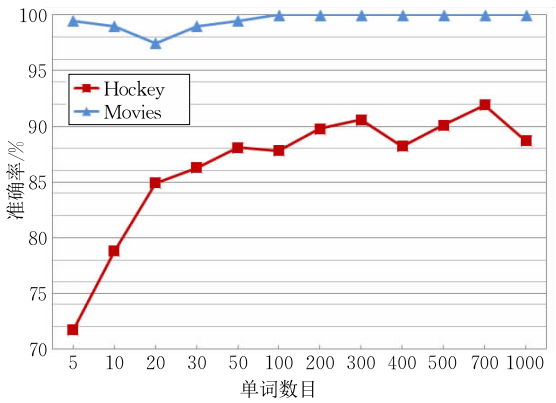

图 12 Hockey 与 Movies 数据集的检测结果

以 Hockey 数据集作为测试数据，MIL 作为分类器，识别效果如图 13 所示，KNN 的实现方式(Citation KNN)获得最好的识别效果，而 DD 方式(Diverse Density)的准确率在 50%上下波动，效果

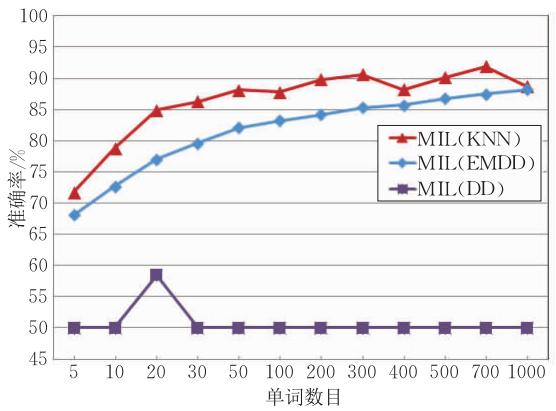

图 13 基于 Conv-CoHOG3D 特征与 MIL 的识别结果

最差，EMDD 的准确率非常接近 KNN，且随单词数目增加而稳定上升. 从 3 种方法的检测结果曲线可知，同一算法理论下，即使是使用相同的特征，因为分类器的实现方式不同会呈现不同的检测效果，所以在进行最终的算法框架设计时，需要多次调试，寻找最优的特征与算法实现方式的组合.

为进一步检验 Conv-CoHOG3D 特征的有效性，在 Hockey 数据集上，对特征进行预处理以使得视频特征维度一致，分别以 SVM(Support Vector Machine)、ELM(Extreme Learning Machine)为分类器进行视频暴力内容的检测.

基于 SVM 分类器的结果如图 14 所示，SVM 分类器一共使用了 5 种不同的核，即 5 种不同的实现方式，分别是：直方图交叉核(Histogram Intersection Kernel，HIK)、线性核(Linear Kernel，LINE)、径向基核(Radial Basis Function Kernel，RBF)、多项式核(Polynomial Kernel，POLY)以及 Sigmoid 核(SIG). 基于 LINE 核与 POLY 核的准确率变化稳定，随单词数目的增加而增长，其结果与最好的 HIK 核基本一致，三种核的准确率均达到 93%，HIK 核的分类器获得最好的检测准确率，当视觉词典的单词数为 700 时，准确率达到 93.98%. 基于 RBF 核与 SIG 核的识别效果较差，SIG 核的准确率一直低于 60%，RBF 核的准确率在单词数目少于 200 时一直低于 50%.

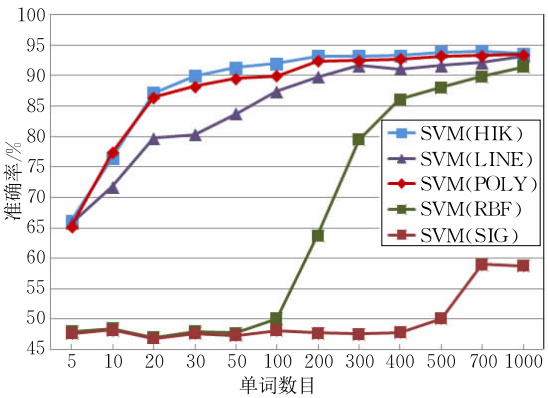

图 14　基于 Conv-CoHOG3D 特征与 SVM 的识别结果

图 15 显示基于极限学习机(Extreme Learning Machine，ELM)的检测结果，同样使用 5 种不同的计算内核实现极限学习机算法. 使用 LINE 核与 POLY 核实现 ELM 算法时，准确率较高，使用 5 到 100 的单词量时，达到 100%；大于 100 时，准确率随着词典变大而下降，且幅度明显；但是当单词数目超过 200 时，准确率又再次符合随单词数目的增加准确率上涨的一般性特点. 通过折线的变化可知，存在一个词典单词数目的临界值，当小于临界值时，基于 LINE 核与 POLY 核的 ELM 能够较好的处理视频特征，从而得到 100%的检测准确率；当单词数目进一步增加，此时单词间距变小，对每个特征计算所属单词造成了干扰，进而影响了词频向量的计算，导致最终识别准确率的下降；当单词数目继续增加时，单词间距进一步变小，视觉词典对单个特征的描述更加准确，逐渐弥补了之前由于单词间距变小造成的特征分类干扰，从而呈现出随单词数目进一步增加而准确率上涨的趋势. HIK 核 ELM 的检测结果较稳定，优于其它的实现方式，识别准确率达到了 94.24%(1000 词时). RBF 核的识别准确率最差，大部分情况下的结果低于其它实现方式. SIG 核的结果略低于 HIK 核. 从多种实现方式的对比中可知，当视觉词典单词数目小于 100 时，使用 LINE 核与 POLY 核可以得到最好的检测效果；当单词数目大于 100 时，使用 HIK 核较合适.

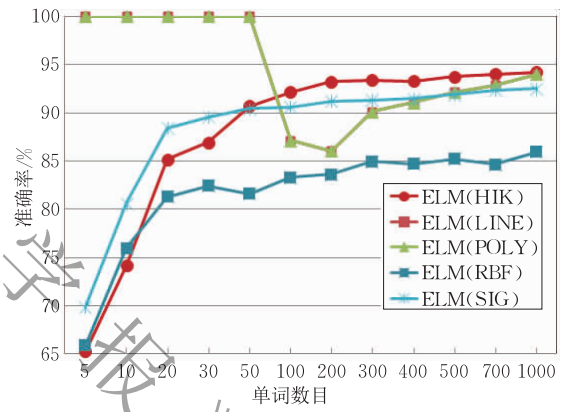

图 15　基于 Conv-CoHOG3D 特征与 ELM 的识别结果

图 16 显示了 MIL、SVM 和 ELM 三种分类器的最优实验结果，基于 SVM 与 ELM 两种分类器的准确率走势一致，识别效果较好一方面表明这两种组合能够较好的适应视频暴力内容的检测，另外，新提出的 Conv-CoHOG3D 特征具有较强的表征能

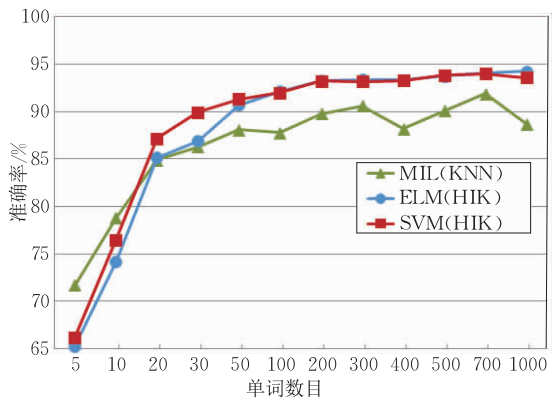

图 16　基于 Conv-CoHOG3D 特征的 3 种分类器识别结果比较

力,检测准确率与分类器弱相关,不会因分类器的差异导致检测结果的较大波动.MIL 分类器的实现方式低于另外 2 种分类器,但是识别准确率的变化趋势与 ELM、SVM 一致,也达到了较高的识别准确率.

将基于 Conv-CoHOG3D 特征的检测结果与基于 Conv-CoHOG 特征的检测结果进行对比,如图 17~图 19 所示,三幅图分别对应 MIL、SVM 和 ELM 分类器的结果,图中识别结果较高的是 Conv-CoHOG3D,说明相比于 Conv-CoHOG 特征,Conv-CoHOG3D 特征的表征能力更强,更适合视频动作特征的描述.

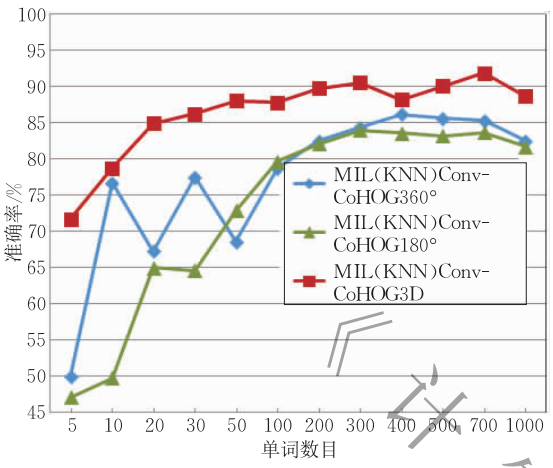

图 17 基于 MIL 分类器的识别结果比较

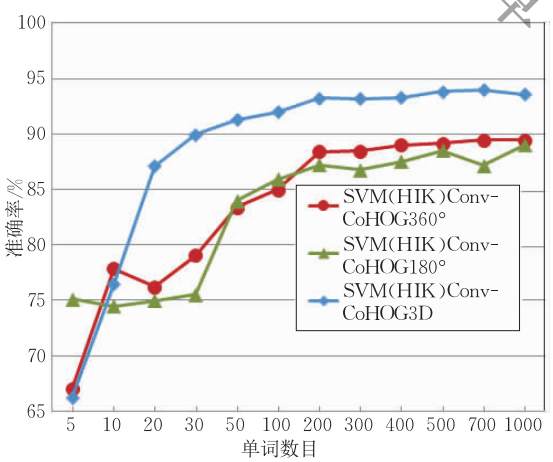

图 18 基于 SVM 分类器的识别结果比较

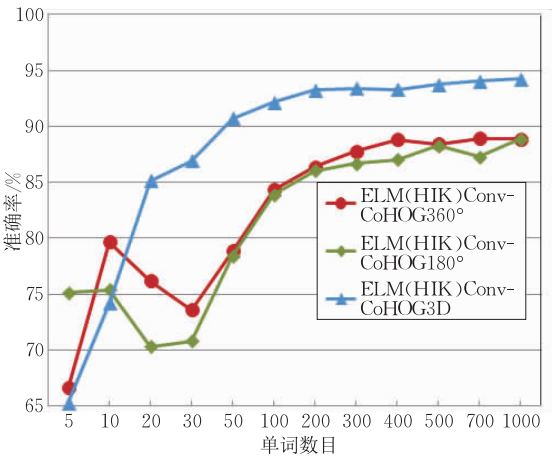

图 19 基于 ELM 分类器的识别结果比较

图 20 是 9 种特征与分类器组合的识别准确率,Conv-CoHOG3D 特征的识别结果大部分位于折线图的上端,即基于 MIL、ELM 和 SVM3 种分类器的实现方式均优于 Conv-CoHOG 特征.

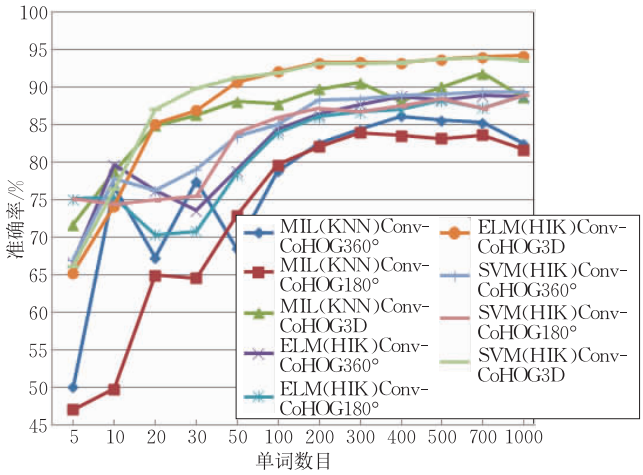

图 20 多种组合的识别结果比较

将本文算法和现有算法进行对比,在 Hockey、Movies 数据集上的检测结果如图 21、图 22 所示.Hockey 数据集上,Nievas 等人[12]使用 HOG 特征和直方图交叉核的 SVM,最终的准确率为 91.7%;Ding 等人[25]在视频暴力内容检测中使用 3D 卷积神经网络,准确率为 91%;Lin 等人[7]基于动作的加速度表示视频特征,检测准确率为 90.1%;Xu 等人[15]利用了 MoSIFT 特征、KDE 方法和稀疏编码,获得 94.3%的准确率.Senst 等人[38]将拉格朗日场

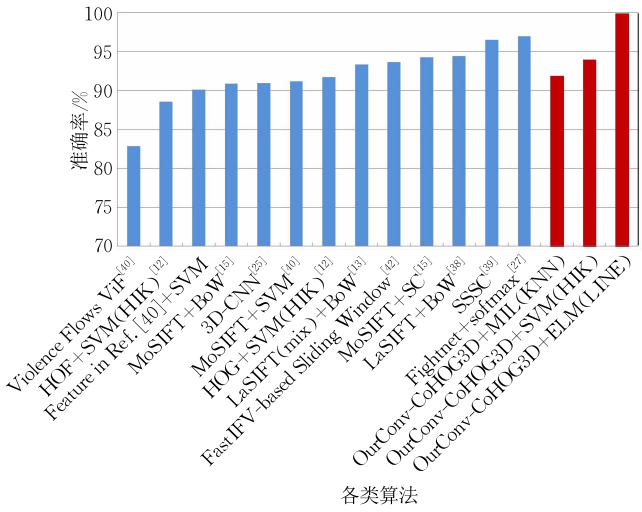

图 21 Hockey 数据集的检测准确率

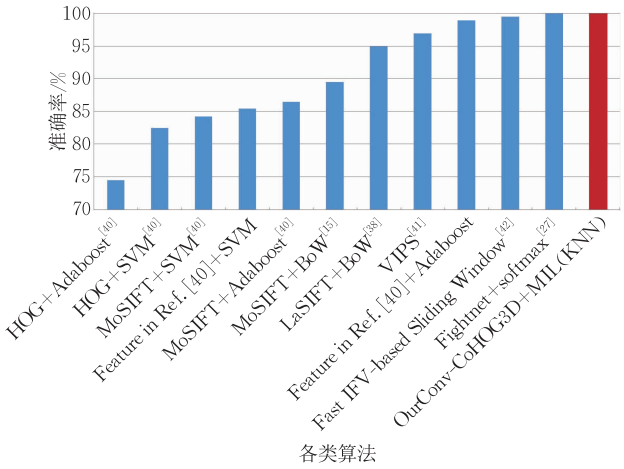

图 22　Movies 数据集的检测准确率

用于描述视频中的运动场景，找到判别效果和计算复杂度的最佳均衡，并且结合了包含尺度敏感特征编码机制的词袋方法和后融合机制来进行暴恐检测，获得 94.42%的准确率. Zhang 等人[39]提出了一种半监督字典学习算法，在原先的字典学习算法中加入了特征约束项和系数非一致项，在 Hockey 数据集上得到 96.5%的准确率. Zhou 等人[27]使用 Wang 等人[26]提出的时域分段网络(TSN)构建了一种用于暴恐视频检测的网络框架 FightNet，以原图像帧、光流场和加速场作为输入进行视频内容识别，获得 97%的检测准确率.

在单词数目少于 100 时，本文基于线性核的 ELM 分类器取得接近 100%的准确率；基于直方图交叉核的 SVM 分类器在单词数目为 700 时，取得 93.98%的准确率；基于 K 近邻 MIL 分类器在单词数目为 700 时取得 91.9%的准确率.

Movies 数据集上，Senst 等人[38]的 LaSIFT+BoW 算法获得 94.95%的检测准确率，Deniz 等人[40]和 Mohammadi[41]等人提出的 VIPS 算法获得 96.91%的检测准确率，文献[42]获得 99.5%的准确率(基于视频序列运动模糊加速度描述特征和 Adaboost 分类器)，Zhou 等人[27]获得 100%的准确率. 本文基于 Conv-CoHOG3D 特征和多示例学习的算法取得 100%的准确率.

在 2 个数据集上，基于 Conv-CoHOG3D 特征的所有分类器均取得较高的识别准确率，达到或者超过当前最优水平，说明本文所提出的 Conv-CoHOG3D 特征对视频动作信息具有较优的表征能力，同时验证词袋模型和多示例学习算法在视频暴力内容检测方面的有效性.

5　结　论

针对现有视频内容检测算法要求视频特征维度一致与视频多尺度之间的矛盾，本文充分考虑了视频的特殊结构，利用多示例学习算法词包概念与视频概念的对应关系，将视频暴力内容检测问题转化为多示例学习问题，使得在处理视频特征时无需做进一步的特征归一化；为了提取视频在三维空间的动作信息，同时降低特征的维度，本文首先利用卷积运算提取 Co-HOG3D 特征，之后对多个相邻的 Co-HOG3D 特征进行池化处理，提出了新的 Conv-CoHOG3D 特征. 在 Hockey 数据集上，检测准确率与文献[15]相比，提高了 5.7%；与文献[27]相比，提高了 3%. 在 Movie 数据集上，准确率相比文献[41]提高了 0.5%，与文献[27]获得相同的检测准确率，算法在该公有测试数据集上取得了最高的检测准确率. 实验结果表明新的特征描述算子可有效刻画视频的动作信息，验证了新算法在视频暴力内容检测问题上的有效性.

参　考　文　献

[1] Yang C, Yuan J, Liu J. Abnormal event detection in crowded scenes using sparse representation. Pattern Recognition, 2013, 46(7): 1851-1864

[2] Wang F S, Wang C L, Bing L, et al. Specified sensitive video recognition based on multi-context construction and linear fusion. Acta Electronica Sinica, 2015, 43(4): 675-683

[3] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA, 2014: 1725-1732

[4] Giannakopoulos T, Kosmopoulos D, Aristidou A, et al. Violence content classification using audio features. Advances in Artificial Intelligence, 2006, 3955: 502-507

[5] Clarin C T, Dionisio M, Echavez M T, et al. DOVE: Detection of movie violence using motion intensity analysis on skin and blood//Proceedings of the 6th Philippine Computing Science Congress (PCSC 2005). Cebu City, Philippines, 2005: 150-156

[6] Nam J H, Alghoniemy M, Tewfik A H. Audio-visual content-based violent scene characterization//Proceedings of the International Conference on Image Processing. Chicago, USA, 1998: 353-357

[7] Lin J, Wang W. Weakly-supervised violence detection in movies with audio and video based co-training. Lecture Notes in Computer Science, 2009, 5879: 930-935

[8] Datta A, Shah M, Vitoria Lobo N D. Person-on-person violence detection in video data//Proceedings of the International Conference on Pattern Recognition. Quebec City, Canada, 2002: 433-438

[9] Wang K, Zhang Z, Wang L. Violence video detection by discriminative slow feature analysis. Communications in Computer & Information Science, 2012, 321: 137-144

[10] Chen D, Wactlar H, Chen M Y, et al. Recognition of aggressive human behavior using binary local motion descriptors. Engineering in Medicine and Biology Society, 2008, (1): 5238-5241

[11] Giannakopoulos T, Makris A, Kosmopoulos D, et al. Audio-visual fusion for detecting violent scenes in videos//Proceedings of the 6th Hellenic Conference on Artificial Intelligence: Theories, MODELS and Applications. Athens, Greece, 2010: 91-100

[12] Nievas E B, Suarez O D, García G B, et al. Violence detection in video using computer vision techniques//Proceedings of the 14th International Conference on Computer Analysis of Images and Patterns. Seville, Spain, 2011: 332-339

[13] Chen M Y, Hauptmann A. MoSIFT: Recognizing human actions in surveillance videos. Annals of Pharmacotherapy, 2009, 39(1): 150-152

[14] Kataoka H, Hashimoto K, Iwata K, et al. Extended co-occurrence HOG with dense trajectories for fine-grained activity recognition//Proceedings of the 12th Asian Conference on Computer Vision (ACCV 2014). Singapore, 2014: 336-349

[15] Xu L, Gong C, Yang J, et al. Violent video detection based on MoSIFT feature and sparse coding//Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Florence, Italy, 2014: 3538-3542

[16] Peng X, Qiao Y, Peng Q, et al. Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition//Proceedings of the British Machine Vision Conference. Bristol, UK, 2013: 1-11

[17] Peng X, Qiao Y, Peng Q. Motion boundary based sampling and 3D co-occurrence descriptors for action recognition. Image & Vision Computing, 2014, 32(9): 616-628

[18] Mousavi H, Mohammadi S, Perina A, et al. Analyzing tracklets for the detection of abnormal crowd behavior//Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision. Big Island, USA, 2015: 148-155

[19] Senst T, Eiselein V, Sikora T. A local feature based on Lagrangian measures for violent video classification//Proceedings of the International Conference on Imaging for Crime Prevention and Detection. London, UK, 2015: 1-6

[20] Yeffet L, Wolf L. Local trinary patterns for human action recognition//Proceedings of the International Conference on Computer Vision. Kyoto, Japan, 2009: 492-497

[21] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies//Proceedings of the Conference on Computer Vision and Pattern Recognition. Alaska, USA, 2008: 1-8

[22] Hinton G E, Salakhutdinov R R. Reducing the dimensionality of data with neural networks. Science, 2006, 313(5786): 504-507

[23] Chen X W, Lin X. Big data deep learning: Challenges and perspectives. IEEE Access, 2014, 2: 514-525

[24] Volodymyr M, Koray K, David S, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518(7540): 529-533

[25] Ding C, Fan S, Zhu M, et al. Violence detection in video by using 3D convolutional neural networks//Proceedings of the 10th International Symposium on Visual Computing (ISVC 2014). Las Vegas, USA, 2014: 551-558

[26] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition. ACM Transactions on Information Systems, 2016, 22(1): 20-36

[27] Zhou P, Ding Q, Luo H, et al. Violent interaction detection in video based on deep learning//Proceedings of the 6th Conference on Advances in Optoelectronics and Micro/Nano-Optics. Nanjing, China, 2017: 012044

[28] Feichtenhofer C, Pinz A, Zisserman A. Convolutional two-stream network fusion for video action recognition//Proceedings of the Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1933-1941

[29] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 4489-4497

[30] Watanabe T, Ito S, Yokoi K. Co-occurrence histograms of oriented gradients for human detection. IEEE Transactions on Computer Vision & Applications, 2010, 2: 39-47

[31] Su B, Lu S, Tian S, et al. Character recognition in natural scenes using convolutional co-occurrence HOG//Proceedings of the International Conference on Pattern Recognition. Stockholm, Sweden, 2014: 2926-2931

[32] Ren Dong, Song Wei, Yu Jing, et al. A survey on special video content detection algorithms. Netinfo Security, 2016, (9): 184-191(in Chinese)
(任栋，宋伟，于京等. 特殊视频内容检测算法研究综述. 信息网络安全，2016，(9)：184-191)

[33] Dalal N, Triggs B, Triggs B. Histograms of oriented gradients for human detection//Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA, 2005: 886-893

[34] Kläser A, Marszalek M, Schmid C. A spatio-temporal descriptor based on 3D-gradients//Proceedings of the British Machine Vision Conference. the University of Leeds, UK, 2008: 995-1004

[35] Tian S, Bhattacharya U, Lu S, et al. Multilingual scene character recognition with co-occurrence of histogram of oriented gradients. Pattern Recognition, 2015, 51(C): 125-134

[36] Dietterich T G, Lathrop R H, Lozano-Perez T. Solving the multiple instance learning with axis-parallel rectangles. Artificial Intelligence, 1997, 89(96): 31-71

[37] Wang J, Zucker J D. Solving the multiple-instance problem: A lazy learning approach//Proceedings of the 17th International Conference on Machine Learning (ICML 2000). Stanford, USA, 2000: 1119-1126

[38] Senst T, Eiselein V, Kuhn A, et al. Crowd violence detection using global motion-compensated Lagrangian features and scale-sensitive video-level representation. IEEE Transactions on Information Forensics & Security, 2017, PP(99): 1

[39] Zhang T, Jia W, Gong C, et al. Semi-supervised dictionary learning via local sparse constraints for violence detection. Pattern Recognition Letters, 2017, 107: 98-104

[40] Deniz O, Serrano I, Bueno G, et al. Fast violence detection in video//Proceedings of the 2014 International Conference on Computer Vision Theory and Applications. Lisbon, Portugal, 2014: 478-485

[41] Mohammadi S, Perina A, Kiani H, et al. Angry crowds: Detecting violent events in videos//Proceedings of the European Conference on Computer Vision. Amsterdam, The Netherlands, 2016: 3-18

[42] Bilinski P, Bremond F. Human violence recognition and detection in surveillance video//Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance. Colorado, USA, 2016: 30-36

SONG Wei, born in 1983, Ph. D., associate professor. His research interests include image processing, video content recognition and multimedia network.

REN Dong, born in 1990, M. S. His research interests include pattern recognition and video content detection.

YU Jing, born in 1971, Ph. D., professor. His research interest is image processing.

QI Zhen-Guo, born in 1989, Ph. D. His research interests include machine learning and video recognition.

Background

Special video detection is an important research direction in machine learning, which can help people surveil the violence motion to handle emergency situations.

Existing special video content detection algorithm based on gradient direction information focused on the features in the video frames, ignored the correlation of the video frames on the time dimension. This paper summarized the achievements of previous research, classified the findings of the research. All algorithms are divided into 3 categories, based on multi-modal features of audio and video and fused color feature, based on fusion of different action features, and the content detection algorithm based on neural network and unsupervised feature extraction. The accuracy of other researches are around 88 percent to 94 percent. This paper considers the potential symbiotic relationship between local gradient direction features, which is beneficial to describe the video action information and to improve the algorithm accuracy. Constructed 3D Conv-CoHOG feature by using Pooling operation and Convolution operation in video frames and neighboring features. The new feature extracts action information from adjacent features, and reduces feature dimension, solved the problems of reduction recognition accuracy because of ignoring the inter-frame information and the difficulty in training high-dimensional features. At the same time, using MIL algorithm mapped video features to instances and bags, which need not extra feature processing. With the above advantages, the algorithms achieved the highest detection accuracy on hockey and movie detection sets, comparing the different detection results on different features, analyzing the causes of the difference results, illustrating the availability of the proposed new algorithm on the special video detection.

This research group mainly study video content detection, pattern recognition, and image processing. And publishing a number of papers on special video content detection, such as horror video recognition based on fuzzy comprehensive evolution, a novel video based face detection algorithm, violent video detection based on visual semantic concept, a review of special video content detection algorithms, and Research on Horror Video Recognition Algorithms.

This work is partly supported by the National Natural Science Foundation of China (Grant Nos. 61503424, 61331013), the China Scholar Council, and the First Class University and First Class Discipline of Minzu University of China ("Image Engineering"), and the Promotion Plan for Young Teachers' Scientific Research ability of Minzu University of China.