光学 精密工程
OPTICS AND PRECISION ENGINEERING
核心期刊
主管：中国科学院
主办：中国科学院长春光学精密机械与物理研究所
中国仪器仪表学会
主编：曹健林

多模态特征融合与多任务学习的特种视频分类

吴晓雨，顾超男，王生进

引用本文:
吴晓雨, 顾超男, 王生进. 多模态特征融合与多任务学习的特种视频分类[J]. 光学 精密工程, 2020, 28(5): 1177-1186.
WU Xiao-yu, GU Chao-nan, WANG Sheng-jin. Special video classification based on multitask learning and multimodal feature fusion[J]. *Optics and Precision Engineering*, 2020, 28(5): 1177-1186.

在线阅读 View online: https://doi.org/10.3788/OPE.20202805.1177

文章编号 1004-924X(2020)05-1177-10

多模态特征融合与多任务学习的特种视频分类

吴晓雨[1*],顾超男[1],王生进[2]
(1.中国传媒大学 信息与通信工程学院,北京 100024;
2.清华大学 电子工程系,北京 100084)

摘要:特种视频(本文特指暴力视频)的智能分类技术有助于实现网络信息内容安全的智能监控。针对现有特种视频多模态特征融合时未考虑语义一致性等问题,本文提出了一种基于音视频多模态特征融合与多任务学习的特种视频识别方法。首先,提取特种视频的表观信息和运动信息随时空变化的视觉语义特征及音频信息语义特征;然后,构建具有语义保持的共享特征子空间,以实现音视频多种模态特征的融合;最后,提出基于音视频特征的语义一致性度量和特种视频分类的多任务学习特种视频分类理论框架,设计了对应的损失函数,实现了端到端的特种视频智能识别。实验结果表明,本文提出的算法在 Violent Flow 和 MediaEval VSD 2015 两个数据集上平均精度分别为 97.97%和39.76%,优于已有研究。结果证明了该算法的有效性,有助于提升特种视频监控的智能化水平。

关 键 词:特种视频识别;特征提取;多模态特征融合;语义一致性度量;多任务学习
中图分类号:TP391.4 **文献标识码**:A **doi**:10.3788/OPE.20202805.1177

Special video classification based on multitask learning and multimodal feature fusion

WU Xiao-yu[1*], GU Chao-nan[1], WANG Sheng-jin[2]

(1. *School of Information and Communication, University of China, Beijing* 100024, *China*;
2. *Department of Electronic Engineer, Tsiinghua University, Beijing* 100084, *China*)
* *Corresponding author, E-mail: wuxiaoyu@cuc.edu.cn*

Abstract: Classification of special videos is significant for intelligent surveillance of internet content. Existing algorithms that fuse multimodal features forclassification of special videoscannot measure multimodal audio-visual semantic correspondence. An algorithm for recognizing special videos based on multimodal audio-visual feature fusion was proposed herein over the framework of multitask learning. First, audio semantic features and spatial-temporal visual semantic cues, including appearance and motion, were extracted. A latent subspace to fuse audio and visual features whilst preserving their semantic information was learned and developed through jointly learning audio-visual semantic correspondence and special video classification. Subsequently, a multitask learning loss function was presented viacombination of the correspondence loss, obtained based on the measured audio-visual seman-

收稿日期:2019-11-29;**修订日期**:2020-01-08.
基金项目:国家自然科学基金资助项目(No. 61801441);北京信息科学与技术国家研究中心跨媒体智能专项资助(No. BNR2019TD01022);"北京市高精尖"学科建设项目(中国传媒大学互联网信息学科);中国传媒大学中央高校基本科研业务费专项资金资助项目(No. CUC2019B066, No. CUC18A002-2)

tic information, and the cross-entropy loss of special video classification. Finally, an end-to-end intelligent system for special video recognition was implemented. Experimental results demonstrate that the accuracy of the proposed algorithm is 97.97% with respect to the Violent Flow dataset, and the average accuracy is 39.76% with respect to the Media Eval VSD 2015 dataset, where by the algorithm outperforms the other existing methods. These results show that the proposed algorithm is effective for improving the intelligence of network content surveillance.

Key words: special video recognition; feature extraction; multimodal feature fusion; semantic correspondence measurement; multitask learning

1 引　言

随着移动智能手机和互联网技术的迅速发展，网络上的视频数据量也急剧增加，网络内容安全日渐成为一个重要问题[1]。单靠人工已无法实现对如此庞大的视频数据量进行审查，这使得色情、暴力等不良视频可能会直接暴露于用户面前，给用户带来视觉和心灵上的负面冲击。本文中的特种视频是指暴力视频。如何有效识别暴力视频以减少暴力内容等有害信息传播是一个亟需解决的问题。因此，本文以暴力视频检测为研究任务，深入探索了其中的关键技术和解决方案，旨在提升暴力视频的智能化检测性能，以净化网络环境。

“暴力”是一个具有高级语义的抽象概念，包括身体和心理暴力，本文只关注身体暴力视频识别，沿用文献[2]对暴力视频的定义如下：“不允许 8 岁以下的小孩观看的包含身体暴力的视频”。互联网的暴力场景视频画面上常伴有流血、打斗，声音上常伴有惊叫、爆炸和枪声等信息，故目前的暴力视频识别方法往往利用音视频信息。基于音视频信息融合的暴力视频识别技术主要涉及暴力音视频各模态特征提取和模态间信息有效融合的两方面问题[3-4]。

在暴力音视频特征提取方面：卷积神经网络(Convolutional Neural Network ，CNN)常被用来提取静态的图像特征，如文献[5]采用 RGB 帧作为输入，利用 ImagNet 数据集预训练的 CNN 初始化暴力视频分类的前 5 层网络，并对最后 3 个全连接层重新训练得到深度特征，实验结果证明与传统特征的分类效果相比，深度特征能帮助提升暴力视频系统识别性能。文献[6]采用深度学习特征和手工设计特征相结合的方法进一步提高暴力视频识别能力，并在分析比较了静态特征、运动特征、基于梅尔频率的倒谱系数 MFCC (Mel-Frequency Cepstral Coefficients)音频特征和基于深度学习的高级语义特征后发现，运动特征对暴力视频识别有较重要的影响。文献[7]借鉴了双流 CNN 网络结构[8]，以静态视频帧和光流图作为两路 CNN 的输入提取暴力视频的特征，并将 CNN 网络输出作为长短时记忆(Long Short-Term Memory，LSTM) 网络[9]的输入以分析长时间视频序列，同时提取并编码了多种手工设计特征，而后将手工设计的特征和深度学习得到的多种特征进行拼接，并训练了几个不同的 SVM 分类器，最后融合不同分类器的分数得到最终的决策结果。文献[10]将相邻视频帧的差分图作为神经网络的输入，利用了卷积 LSTM 网络提取暴力视频的帧间变化信息和场景语义信息。目前暴力视频特征提取方法多是粗暴地将经典特征描述算子和深度神经网络自动提取特征描述子进行简单地组合拼接，这无疑会制约暴力视频检测算法的计算效率，我们更应该从暴力场景的特点出发(如暴力场景有的以血腥场面为主，有的以打架场面为主，有的以爆炸着火场面为主)，采用有效的音视频特征提取方法来获得暴力场景的语义表征。

如何对提出的静态帧、运动和音频等多种特征进行有效地信息融合是暴力视频识别研究中的另一重要内容。在暴力音视频模态间信息融合方面：目前多路信息融合的技术方法主要有基于决策分数的后融合方法和基于特征层的前融合方法[11]。决策层的融合指将各模态的决策结果(如各模态的分类器给出的分数) 进行融合[12]。主

要的融合方法有基于规则的方法，如线性权重融合、平均融合、投票决策等。基于分类器学习的融合方法即将各模态分数作为特征通过训练学习得到一个判别函数，如基于 SVM(Support Vector Machine)、贝叶斯决策、logistic 回归和神经网络等方法。特征层的前融合是指将提取的各视角特征按照某种方法进行的融合，常见的特征融合方法有：(1)直接将特征拼接为一个长的特征向量，一般随后采用特征编码方法，如词包模型(Bag of Word，BOW)、Fisher 向量编码 (Fisher Vector，FV)方法或者 主成分分析(Principle Component Analysis，PCA)等方法，进行特征降维，最后利用 SVM 或者 Softmax 分类器得到分类的结果，这种特征融合方法虽实现简单，但是多模态数据间存在“语义鸿沟”的问题，故将不同含义异质的多种特征直接进行拼接后效果不稳定。(2)将多模态特征经过某些变换投影以得到潜在的、共享的特征子空间[13-14]，该方法在融合过程中往往会考虑模态间的关联性，故更为科学。目前已有的暴力音视频信息融合常采用决策层的融合技术，这主要是因为决策层的信息融合相当于对语义相近的、处在同一个特征空间的特征(即决策分数)进行融合，融合风险较小且实现也相对容易。但是，决策层融合方法对暴力视频识别性能的改善作用也是比较有限的，原因在于在进行决策层融合时可利用的只是各模态决策后的分数，融合信息很有限。较决策层融合相比，特征层的融合方法优势在于同时“看到”了更多的模态信息，能更好捕捉各模态的联系，好的特征融合方法能显著提高视频分类性能。但该方法难点也在于各特征含义不同、建立具有统一语义表示的特征子空间较难。总的来说，无论是决策层融合还是特征层融合方法，在融合音视频信息时均没有考虑音视频特征语义一致性的问题。多模态特征之间有时具有语义一致性(以暴力视频为例，语义一致性可以理解为暴力音视频特征同时具有暴力场面描述的特点，或同时不具有暴力场景描述的特点) 和信息互补，但有时多模态间信息是互相干扰的(如著名的“麦格克效应”-McGurk effect)，融合它们甚至会有相反的效果。因此要显式地思考什么时候进行哪些信息的融合，不加任何度量直接地进行模态间的特征融合有时不仅无法实现模态间信息互补，而且还会导致算法性能的下降[15]。

本文针对现有暴力音视频特征对暴力场景语义描述能力不足、融合音视频特征时未考虑语义一致性等问题，提出了一种基于音视频特征多任务学习的端到端暴力视频识别方法：提取具有时空相关性的音视频特征方法，构建具有语义保持的共享的特征子空间，提出了基于暴力音视频特征语义一致性度量和视频分类相结合多任务学习的暴力视频分类模型，实现了暴力音视频信息的有效融合与互补。在两个暴力视频公开数据集上的实验结果表明本文提出方法的有效性。同时该方法也将为类似任务的音视频特征融合提供了一定的理论参考。

2　基于语义一致性的暴力视频识别方法

本文整体技术路线如图 1 所示：首先，以 2～4 s 短视频数据为处理对象，以分析暴力场景视频的特点为出发点，基于 P3D+LSTM 网络提取适合暴力场景描述的、具有时空相关特性的视觉语义特征，基于 VGGish 网络提取暴力音频的语义特征；而后在多特征融合过程中，以暴力视频分类标签和音视频语义一致性信息为监督信号，自动学习并求取具有语义保持的特征映射的变换矩阵，实现基于暴力音视频特征语义一致性度量和视频分类相结合多任务学习的暴力视频分类。

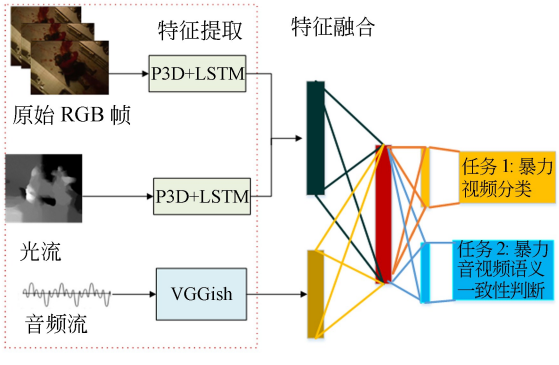


图 1　暴力视频分类算法框架图

Fig. 1　Framework of violent video classification

2.1 暴力音视频特征提取

暴力类视频从视觉信息上来讲，画面通常包括物体（枪支、刀、剑等）、场景（血液、死亡等场景）、动作或行为（如打斗、追逐、射击等）。在音频信息方面，暴力视频中经常会伴有尖叫、爆炸、枪声等，故本文利用深度学习算法提取表观特性和运动信息随时空变化的视频语义特征及音频语义特征，来表征血腥、打架和爆炸等暴力场景。

2.1.1 基于 P3D+LSTM 的暴力视频视觉语义特征提取

对于暴力视频，利用视频当前帧的前后多帧上下文的信息可以减少基于单帧信息引起的误判，有助于提高血腥场景检测的准确度。同时为充分考虑暴力视频在表观和运动上的特点，本文参考双流的框架，分别以原始视频 RGB 帧和光流作为输入，以伪 3D（Pseudo-3D，P3D）[16] 和长短时记忆 LSTM 网络[9] 为网络结构，提取暴力视频中表观特性和运动信息随时空变化的视觉语义特征。

（1）基于 P3D+LSTM 网络提取表观信息随时空变化的视频语义特征

对于血腥的暴力视频，提取基于原始 RGB 帧的表观语义特征是很有必要的。P3D 网络使用了“伪”3D 卷积的概念降低网络参数，即利用拆分的思想把原本 3×3×3 的卷积拆分成了 3×1×1 卷积与 1×3×3 卷积的结合，以 16 帧的连续图像作为网络输入单元，提取短时的视频时空连续性特征，显然 P3D 对于长视频的处理还存在一些不足，实际中往往将 P3D 最后一个平均池化层特征作为 LSTM 的输入以提取长序列视频的时空特征。因此，本文以视频暴力/非暴力标签信息作为监督信号，以暴力视频原始帧 RGB 信息作为输入，基于 P3D+LSTM 网络学习并提取表观信息随时空变化的 512 维视频语义特征 $\boldsymbol{fVa}$。

（2）基于 P3D+LSTM 网络提取运动信息随时空变化的视频语义特征

对于打斗暴力场景，运动特征对此具有较强的描述能力。目前运动特征的提取主要借助光流 Optical flow、改进稠密轨迹 iDT（improved Dense Trajectory）算子和帧间差分等方法，iDT 计算复杂度较高，帧间差分法虽然计算简单但是当目标运动较快时无法获取完整的运动目标。因此，本文选用光流法来表征视频的运动信息，以光流图像作为网络的输入，基于 P3D+LSTM 网络学习并提取运动信息随时空变化的 512 维视频语义特征 $\boldsymbol{fVm}$。

在视觉通道模型训练阶段，表观流和光流这两路 3D 网络模型的初始化参数来自于 Kinetics 400 数据集[17] 的预训练模型。参数的设置如下：P3D 训练模型初始学习率设为 0.000 01，且以 gamma=0.1 的幅度每 5 000 次对学习率进行一次调整；训练时 batch_size 设置为 4；最大迭代次数 max_iter=30 000；梯度影响因子 momentum 设置为 0.9；当 P3D 网络提取的最后一个平均池化层特征被发送到 LSTM 时，batch_size 被设置为 64，最大 epoch 设置为 55，初始学习率设置为 0.000 1。

2.1.2 基于 VGGish 网络的暴力视频音频语义特征提取

暴力视频在音频信息中经常会伴有尖叫、爆炸、枪声等，因此暴力视频智能化识别的研究不能仅考虑视觉方面的信息，音频信息同样也对暴力视频的识别提供指示性帮助。这里假定处理的视频存在音频流信息。本文首先提取音频 log-Mels 梅尔谱图，而后将梅尔谱图送入 VGGish 网络[18]，通过学习尖叫、爆炸、枪声的暴力音频数据使得网络学习到暴力音频的音效特征，获得 128 维暴力音频语义特征 $\boldsymbol{fA}$，以此辅助暴力视频的检测。这里不选择 P3D 网络进行暴力音频语义特征提取的原因在于，输入图像 log-mel 谱图虽然是一幅图，但是和自然图像空间位置信息的含义截然不同，因此并不适合提取音频语义特征。

训练采用的音频数据均是从原始暴力视频中利用 ffmpeg 工具分离出来的，而后将音频数据经过如下预处理：所有音频数据被重新采样到 16KHz 的单声道形式，对音频的分帧采用了窗口大小为 25 ms、窗口跳距为 10 ms 以及周期 Hann 窗口的短时傅里叶变换的幅度，而后映射得到稳定的 log-Mels 谱。然后这些特征被以 0.96 s 的时长组帧，且不会出现帧的重叠，其中每一帧都包含 64 个 Mel 频带，时长为 10 ms，即总共 96 frame。将提取 96×64×1 的音频数据送入 VGGish 网络提取暴力音频语义特征。

在音频特征提取模型训练过程中，网络是基于 VGGish 网络在 Audioset 数据集上预训练模型进行微调训练的。此外，在训练过程中对训练

音频数据进行扩充处理，每段音频再按照 1 s 的时间间隔截断扩充成 10 个，batch_size 设置为 16，epoch 设置为 60，初始学习率 0.000 01。

2.2　基于语义一致性的多特征融合与暴力视频识别

合理的特征融合方法相比决策层融合往往可以获得更高的性能提升。特征层融合常将多种特征投影变换到一个共享的特征子空间上，但是如何求取变换矩阵以构建合理的特征子空间是该方法的核心。在对多种特征进行融合时，只有将具有相同语义的特征进行融合处理才能充分利用各类特征之间的信息互补性。但现有的研究方法只是单纯地基于视频标签来对特征融合层进行训练[19]，没有考虑到各种特征之间可能存在语义不一致的情况，这导致在多特征融合过程中可能会出现特征信息相互“敌对”的问题，使得该类方法在本就数量有限的暴力视频训练数据集上会更容易出现过拟合现象，影响了暴力视频分类系统的泛化能力。

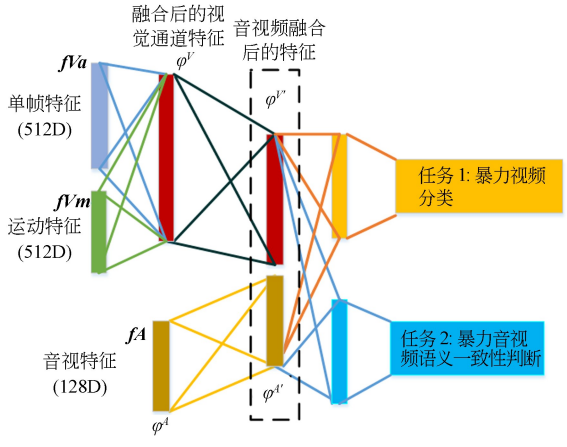


图 2　基于多任务学习的暴力音视频特征融合

Fig. 2　Violent audio-visual features fusion based on multitask learning

因此，本文提出了结合暴力音视频特征语义一致性度量的多模态融合方法，技术路线图如图 2 所示，实现了基于音视频特征多任务学习的暴力视频分类方法。在视觉通道上，我们通过构建并训练基于全连接的特征融合层的网络结构，将随时空变化的 512 维的表观语义特征 $\boldsymbol{fVa}$ 和 512 维运动语义特征 $\boldsymbol{fVm}$ 特征投影到 512 维的视觉特征融合空间，使得视觉特征从 1 024 维降为 512 维，这样不仅实现了表观和运动特征的融合，更为重要的是减少了视觉特征的维度，降低后续建立音视频共享特征子空间的技术难度。在音视频特征融合方面，构建 2 个全连接特征融合层，相当于分别求取视觉特征变换矩阵 $\boldsymbol{W}^V$ 和音频特征变换矩阵 $\boldsymbol{W}^A$，将视觉通道 512 维特征 $\boldsymbol{\varphi}^V$ 和音频通道 128 维特征 $\boldsymbol{\varphi}^A$，经过公式(1)各自矩阵变换得到音视频共享特征子空间，从而得到的融合后 512 维音视频特征 $\boldsymbol{\varphi}'=(\boldsymbol{\varphi V}',\boldsymbol{\varphi A}')$。其中融合后的特征维度通过反复实验选取，公式(1)中参数($\boldsymbol{W}^V$，$\boldsymbol{b}^V$，$\boldsymbol{W}^A$，$\boldsymbol{b}^A$)由模型训练得到。

$$\begin{aligned}\boldsymbol{\varphi}^{V'} &= \boldsymbol{W}^V\boldsymbol{\varphi}^V+\boldsymbol{b}^V\\ \boldsymbol{\varphi}^{A'} &= \boldsymbol{W}^A\boldsymbol{\varphi}^A+\boldsymbol{b}^A.\end{aligned}\tag{1}$$

本文提出的多模态特征融合方法创新在于：在学习特征融合层参数的模型训练阶段，不仅考虑了暴力视频的分类任务，还引入了音视频语义一致性任务进行协调反馈，两个任务并行学习训练且共享已学到的特征参数。因此暴力视频分类网络的损失函数由两部分组成：一是暴力分类的二值交叉熵损失函数，二是增添语义一致性分类的损失函数。具体的损失函数公式如(2)：

$$\begin{aligned}Loss &= L_{\text{classification}}+\lambda L_{\text{correspondence}}\\ L_{\text{classification}} &= -(y\times\log(\hat{y})+(1-y)\log(1-\hat{y})),\\ L_{\text{correspondence}} &= -(y'\times\log(\hat{y}')+\\ &\quad (1-y')\log(1-\hat{y}')),\end{aligned}\tag{2}$$

其中：整个深度神经网络的训练损失函数 $Loss$ 中，两种 $Loss$ 的权重比值 λ 取 2；暴力视频分类的二值交叉熵损失函数 $L_{\text{classification}}$ 中，$\hat{y}$ 表示暴力视频分类的预测值，y 表示暴力视频分类的真实值；音视频语义一致性度量损失函数 $L_{\text{correspondence}}$ 中，$\hat{y}'$ 表示语义一致性任务的预测值，y' 表示语义一致性任务的真实值，$y'=1$ 时表示音视频信息语义一致，即同时具备暴力场景的特点，或者同时都不具备暴力场景的特点。$y'=0$ 表示音频和视频语义不一致。

这里，增加语义一致性度量的交叉熵损失函数作用是在音视频特征映射到共享的特征子空间过程中，增加了一致性的约束条件，更好地保持音视频模态间及各模态内部特征数据的语义信息，引导网络学习到具有语义保持的音视频融合特征。相比于直接计算融合音视频特征的相似性距离，语义一致性任务的损失函数从语义保持为目标，较大程度实现了模态间“求同存异”，而相似性

距离过多强调了多模态特征相似性，弱化了其差异互补性，但过于相似的多特征则失去了互补性。从另一个角度来看，语义一致性度量损失函数相当于对暴力分类损失函数增加了正则项，在暴力视频数据集由于内容的敏感性构建过程比较困难的情况下，一定程度上降低了算法对暴力视频训练数据的要求，提升了暴力视频算法的泛化能力。

需要说明的是，本文仅在网络模型训练阶段，增加语义一致性度量的任务，采用基于音视频特征多任务学习方法方法训练得到更为有效的特征融合层参数。当训练结束暴力视频分类的整个网络模型参数固定后，在测试阶段，对测试的视频仅进行任务 1 即视频是否为暴力的判别。

3　实验结果及分析

3.1　Violent Flow 数据集实验结果及分析

3.1.1　数据集描述及评价指标

公开的暴力视频数据集 The Violent Flow 数据集[20]是一个群体暴力数据集，参与暴力事件的人数非常多。这个数据集中的大部分视频都是从足球比赛中发生的暴力事件中收集的。这个数据集中共有 246 个视频，其中暴力视频和非暴力视频各 123 个。

Violent Flow 库上的评测指标采用的准确率(Accuary)即：

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}\times 100\%, \tag{3}$$

其中：TP(True Positive-被正确分类的正例)TN(True Negative-被正确分类的负例)，FP(False positive -假正例)和 FN(False Negative-假负例)。

3.1.2　实验结果

The Violent Flow 数据集中视频的音频信息非原始音频信息，而多是后配上的没有意义的背景音乐。因此，本文只验证了以视频 RGB 帧和光流为网络输入，基于 P3D+LSTM 的视觉特征提取的有效性。从表 1 的实验结果可以看到基于 P3D+LSTM 提取表观和运动随时空变化的语义特征较其他方法更好地表述了暴力视频的特征，相比于已有方法取得了较好的实验结果。

表 1　在 Violent Flow 数据集上的实验结果比较

Tab. 1　Result comparison between other algorithms and our algorithm on Violent Flow dataset

方法	准确率
Bilinski *et al*.[21]	96.4%
MoIWLD[22]	(93.19±0.12)%
Swathikiran[10]	(94.57±2.34)%
本文仅利用 RGB 帧的 P3D+LSTM 网络模型	96.33%
本文仅利用光流的 P3D+LSTM 网络模型	90.63%
本文(视频 RGB 帧和光流 P3D+LSTM 网络模型)	97.97%

3.2　MediaEval VSD 2015 数据集实验结果及分析

3.2.1　数据集描述及评价指标

MediaEval VSD(Violent Scenes Detection) 2015[23]暴力视频公开数据集，是由欧洲 MediaEval 2015 暴力视频检测竞赛组织方提供的。该数据集来自于 199 部电影，由 10 900 个短视频组成，其中训练集 6 144 个短视频(暴力 272 个，非暴力 5 872 个)，测试集 4 756 个短视频(暴力 230 个，非暴力 4 526 个)。本文在此基础上对训练数据增加了语义一致性标签。

虽然 MediaEval VSD 2015 数据集由上万个视频数据，但是暴力视频所占的比例不足 5%。在暴力视频和非暴力视频样例比例严重不均衡的情况下，使用准确率作为评价指标将无法充分衡量暴力视频分类性能，因此 Media VSD 2015 官方采用了平均正确率 AP(Average Precision)，并提供了 AP 的计算工具。

3.2.2　实验结果

本文首先在 MediaEval VSD 2015 数据集上开展了基于单模态、双模态及多模态的对比实验，具体实验结果见表 2。由这些实验数据可以看出，本文提出的基于构建共享特征空间的前融合多特征融合方法分类准确率优于单特征通道和双通道融合的分类结果。

表 2　不同模态在 MediaEval VSD 2015 数据集结果比较

Tab. 2　Comparison based on different modalities on MediaEval VSD 2015 dataset

	模态	AP/%
单模态	仅基于 RGB 帧的 P3D+LSTM 网络模型	28.32
	仅基于光流的 P3D+LSTM 网络模型 22.29%	22.29
	仅基于音频的 VGGish 网络模型	14.16
双模态	RGB 帧＋光流的前融合结果	36.93
	光流＋音频的前融合结果	31.41
	RGB 帧＋音频的前融合结果	29.46
多模态(RGB＋光流＋音频)	基于后融合的暴力视频分类	38.12
	未加入语义一致性度量的前融合的暴力视频分类	38.65
	加入语义一致性度量的前融合的暴力视频分类	39.76

具体来说，(1)在基于单模态的暴力视频分类中：仅基于视觉通道特征(视频 RGB 帧)的暴力视频分类方法的 AP 值最高为 28.32%，而仅基于提取的音频特征的暴力视频分类方法的 AP 值最低为 14.16%。这说明在 MediaEval VSD 2015 暴力视频公开库中，对于暴力视频分类的任务来说，特征贡献率最大的是视觉通道的表观语义特征，其次是运动语义特征，最小贡献的是音频特征。这也是可以理解的，仅利用音频信息有时不足以做出是否暴力的判别，比如含有爆炸声和尖叫声的音频也可能是节日的欢庆，这时必须结合视觉信息或者附以情感分析才可能做出更准确的判断。(2)在基于双模态前融合的暴力视频分类方法中，基于 RGB 帧和运动光流两路特征前融合方法的 AP 值达到了 36.93%，基于光流和音频两路特征前融合方法的 AP 值达到了 31.41%，基于 RGB 和音频两路特征前融合方法的 AP 值达到了 29.46%。任何两路的融合结果都比单一特征分类结果要好，即使音频对暴力视频分类贡献最小，但加入音频特征仍然有助于提升暴力视频分类性能，这充分表明了表观、运动和音频三种特征，在暴力视频分类中具有彼此互补性。(3)在 RGB、光流和音频三种特征的多模态融合中，我们首先比较了决策层后融合和特征层前融合的实验结果：后融合的暴力视频分类方法是将 RGB 这路的 P3D ＋ LSTM 网络输出的分类分数、光流这路 P3D ＋ LSTM 网络输出的分类分数和音频这路 VGGish 网络输出的分类分数作为特征，送入高斯核 SVM 分类器学习分类器参数，该方法的 AP 值是 38.12%；而在未加入语义一致性下基于特征层的前融合方法分类准确率为38.65%，这进一步说明了后融合方法丢失了各特征之间的关系，结果不如特征层的前融合方法；最后，加入语义一致性度量的前融合的暴力视频分类 AP 值提升至 39.76%，这说明了增加音视频语义一致性度量约束的多任务特征前融合方法构建了较好的特征子空间，使得融合后的特征更为有效地实现暴力音视频信息互补性。

表 3　在 MediaEval VSD 2015 数据集不同方法实验结果比较

Tab. 3　Comparison based on different methods on MediaEval VSD 2015 dataset

方法	AP/%
Fudan-Huawei[7]	29.59
Esra *et al*. [24]	29.47
MIC-TJU [5]	28.48
本文方法	39.76

表 3 给出了在 MediaEval VSD 2015 暴力视频公开库上，已有公开方法和本文提出的方法的对比实验结果。从表 3 可以看出本文方法比其他方法 AP 值高了 10.17%，充分说明了本方法的

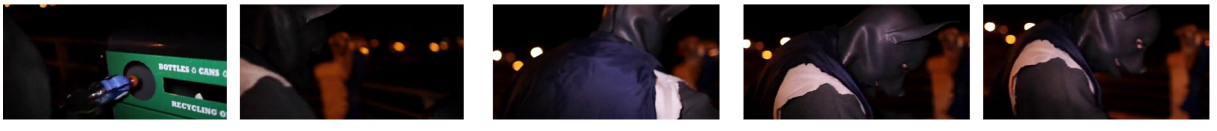
有效性。本文算法性能提升的原因主要得益于选取适合的深度学习方法构建了暴力视频多模态特征提取网络模型，更有效地提取了具有时空连续性的暴力视频的表观、运动和音频语义特征，获得了对暴力视频的有效表征。同时，本文提出了基于语义一致性度量和视频分类的多任务学习损失函数，构建了语义保持的多特征融合的特征共享子空间，进一步提升了暴力视频分类性能。

3.2.3 可视化实验结果

图 3 给出了 MediaEval VSD 2015 公开数据集中部分视频的序列帧。图 3(a)显示了真实标签为暴力的 ACCEDE02119 视频的 32，64，96，128 和 160 frame。该视频视觉通道上有明显打斗动作，音频通道含有痛苦的叫喊声，音视频均具有明显的暴力特征，算法经过多种特征提取和融合正确预测了该视频为暴力视频。图 3(b)显示了真实标签为暴力的 MEDIAEVAL00397 视频的 32，64，96，128 和 160 frame，该视频仅有流血场面较少，音频中有枪声信息。若利用未考虑语义一致性的前融合方法，该视频将被误判为非暴力，而采用提出的语义一致性度量的前融合方法可正确分类为暴力视频。图 3(c)显示了真实标签为非暴力的 ACCEDE09670 视频的 32，64，96，128 和 160 frame，该视频画面较昏暗，昏暗的灯光和流血有一定相似性，蒙面丢瓶的动作和打架出拳的动作有一定相似性，音频信息比较舒缓正常具有明显非暴力的特点。若利用未考虑语义一致性的前融合方法，该视频将被误判为暴力视频，而采用提出的语义一致性度量的前融合方法，可正确分类为非暴力视频。图 3(d)显示了真实标签为非暴力的 ACCEDE00591 视频的 32，64，96，128 和 160 frame，音视频均没有明显的暴力特征，算法经过多种特征提取和融合正确判别了该视频为非暴力视频。

(a) ACCEDE02119 #32 #64 #96 #128 #160

(b) MEDIAEVAL00397 #32 #64 #96 #128 #160



(c) ACCEDE09670 #32 #64 #96 #128 #160

(d) ACCEDE00591 #32 #64 #96 #128 #160

图 3 MediaEval VSD 2015 公开数据集中部分视频的序列帧

Fig. 3 Video sequences from MediaEval VSD 2015 dataset

4 结 论

针对暴力音视频特征融合时未考虑语义一致性的问题，本文提出了一种基于音视频特征多任务学习的端到端暴力视频分类方法。首先提取暴力视频在单帧图像、运动信息及音频方面的多种特征，即采用 P3D+LSTM 网络提取具

有时空特征的表观和运动的语义特征，基于VGGish网络获得暴力视频音频语义特征，而后在融合暴力音视频特征中，以构建具有语义保持的共享特征子空间为出发点，提出了基于语义一致性度量及多任务学习的特征融合方法，形成了以判断暴力视频分类和音视频语义一致性两种任务共同学习的暴力视频分类框架。最后，提出的算法在两个公开暴力视频数据集进行了测试，均取得较好的实验结果，其中在MediaEval VSD 2015数据集上平均正确率达到了39.76%，优于已有暴力视频判别算法。实验结果充分证明了本文提出的暴力视频多特征融合及分类算法的有效性。

目前的暴力视频分类主要依靠从有限的标注训练数据中获得的暴力视频特征，但是该方法学习到的特征和知识受限于训练数据规模和分布，下一步将考虑构建暴力视频的知识图谱，将知识图谱的外部先验信息嵌入到深度模型的网络结构中，探索外部知识和标注数据信息的有效融合，进一步提升暴力视频分类性能。

参考文献：

[1] 马晓晨，韦世奎，蒋翔，等. 基于相机溯源的潜在不良视频通话预警 [J]. 光学 精密工程，2018，26 (11)：2785-2794.
MA X CH，WEI SH K，JIANG X，*et al*.. Early warning of illegal video chats based on camera source identification [J]. *Opt. Precision Eng.*，2018，26 (11)：2785-2794. (in Chinese)

[2] CLAIRE H D，CEDRIC P，MOHAMMAD S，*et al*.. VSD，a public dataset for the detection of violent scenes in movies：design，annotation，analysis and evaluation [J]. *Multimedia Tools and Applications*，2014，74 (17)：7379-7404.

[3] MOREIRA D，AVILA S ，PEREZ M ，*et al*.. Multimodal data fusion for sensitive scene localization [J]. *Information Fusion*，2019 (45)：307-323.

[4] WANG H M，YANG L，WU X Y，*et al*.. A review of bloody violence in video classification [C]. *International Conference on the Frontiers & Advances in Data Science*，2017：86-91.

[5] YI Y，WANG H，ZHANG B，*et al*.. MIC-TJU at affective impact of movies task [C]. *MediaEval Workshop*，2015，7.

[6] LAM，LE S P，DO T，*et al*.. Computational optimization for violent scenes detection [C]. *International Conference on Computer，Control，Informatics and its Applications*，2016：141-146.

[7] DAI Q，ZHAO R，WU Z，*et al*.. Fudan-Huawei at mediaeval 2015：Detecting violent scenes and affective impact in movies with deep learning [C]. *MediaEval Workshop*，2015，5.

[8] SIMONYAN K，ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [C]. *NeurIPS*，2014：568-576.

[9] SUTSKEVER I，VINYALS O，LE Q V. Sequence to sequence learning with neural networks [C]. *NeurIPS*，2014：3104-3112.

[10] SWATHIKIRAN S，OSWALD L. Learning to detect violent videos using convolutional long short-term memory [C]. *IEEE International Conference on Advanced Video and Signal Based Surveillance*，2017：1-6.

[11] BALTRUSAITIS T，AHUJA C，MORENCY L P. Multimodal machine learning：a survey and taxonomy [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*，2019 41(2)：423-443.

[12] 崔鑫，彭宗举，陈芬. 联合多特征的未来视频快速编码 [J]. 光学 精密工程，2019，27 (4)：990-999.
CUI X，PENG Z J，CHEN F. Joint Multi-feature fast coding for future video coding [J]. *Opt. Precision Eng.*，2019，27 (4)：990-999. (in Chinese)

[13] WU Z ，JIANG Y G ，WANG X ，*et al*.. Multi-Stream multi-class fusion of deep networks for video classification [C]. *ACM International Conference on Multimedia*，2016：791-800.

[14] 潘仙张，张石清，郭文平. 多模深度卷积神经网络应用于视频表情识别 [J]. 光学 精密工程，2019，27 (4)：963-970.
PAN X ZH，ZHANG SH Q，GUO W P. Video-based facial expression recognition using multimodal deep convolutional neural networks [J]. *Opt. Precision Eng.*，2019，27 (4)：963-970. (in Chinese)

[15] ATREY P K，HOSSAIN M A，SADDIK A E，*et al*.. Multimodal fusion for multimedia analysis：a survey [J]. *Multimedia Systems*，2010，16(6)：345-379.

[16] QIU Z, YAO T, TAO M. Learning spatial-temporal representation with pseudo-3d residual networks [C]. *IEEE International Conference on Computer Vision*, 2017:5534-5542.

[17] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition A new model and the kinetics dataset [C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 6299-6308.

[18] HERSHEY S, CHAUDHURI S, ELLIS D P W, *et al*.. CNN architectures for large-scale audio classification [C]. *International Conference on Acoustics, Speech and Signal Processing*, 2017: 131-135.

[19] WU Z, JIANG Y G, WANG J, *et al*.. Exploring inter-feature and inter-class relationships with deep neural networks for video classification [C]. *ACM International Conference on Multimedia*, 2014: 167-176.

[20] HASSNER T, ITCHER Y, KLIPER C O. Violent flows: Real-time detection of violent crowd behavior [C]. *IEEE Conference on Computer Vision and Pattern Recognition*, 2012:1-6

[21] BILINSKI P, BREMOND F. Human violence recognition and detection in surveillance videos [C]. *IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2016: 30-36.

[22] 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.

[23] MATS S, YOANN B, HANLI W, *et al*.. The mediaeval 2015 affective impact of movies task [C]. *MediaEval Workshop*, 2015:1.

[24] ESRA A, FRANK H, SAHIN A. Breaking down violence detection: combining divide-et-impera and coarse-to-fine strategies [J]. *Neurocomputing*, 2016, 208: 225-237.

作者简介：

吴晓雨(1979－)，女，辽宁盘锦人，博士，副教授，2004 年于吉林大学获得硕士学位，2009 年于中科院自动化研究所获得博士学位，主要从事计算机视觉、视频分析与理解方面的研究。E-mail：wuxiaoyu@cuc. edu. cn

顾超男(1995－)，女，河北保定人，硕士研究生，2014 年于中国传媒大学获得学士学位，主要从事视频内容理解的算法研究。E-mail：gcn@cuc. edu. cn