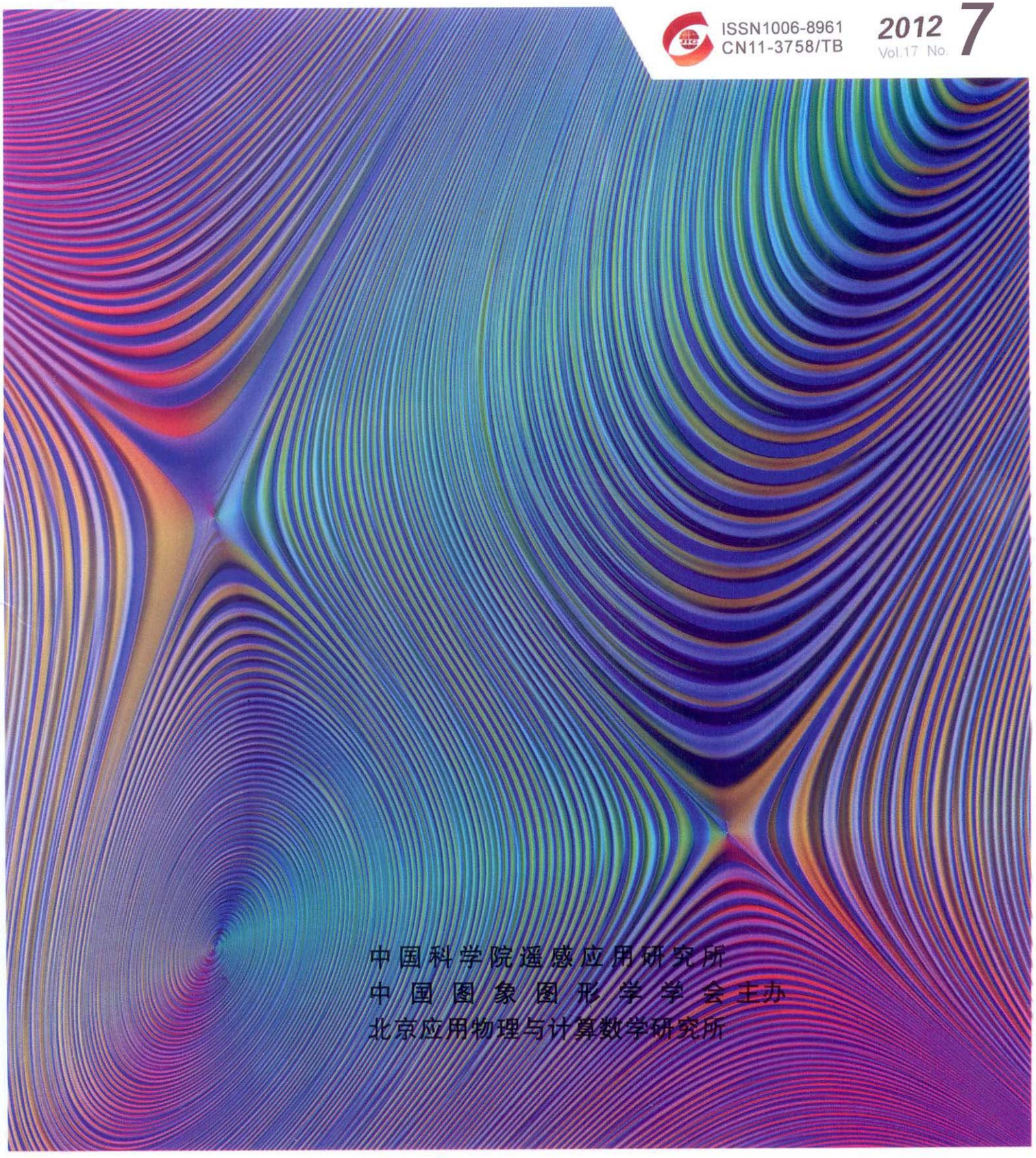
Journal of Image and Graphics

中国图象图形学报

ISSN1006-8961
CN11-3758/TB

2012 7
Vol.17 No.



中国科学院遥感应用研究所
中国图象图形学学会主办
北京应用物理与计算数学研究所

中国精品科技期刊

中国图象图形学报

Zhongguo Tuxiang Tuxing Xuebao

2012 年 7 月　第 17 卷　第 7 期(总第 195 期)

目　次

中国图象图形学报

刊名题字：宋　健　　月刊(1996 年创刊)　　第 17 卷　第 7 期　　2012 年 7 月 16 日出版

主管单位	中国科学院
主　　办	中国科学院遥感应用研究所 中国图象图形学学会 北京应用物理与计算数学研究所
主　　编	李小文
编辑出版	《中国图象图形学报》编辑出版委员会 北京 9718 信箱　邮编 100101 电子信箱:jig@ irsa. ac. cn 电话:010-64807995　010-82614429 网　址:www. cjig. cn
印刷装订	北京北林印刷厂
广告经营许可证	京朝工商广字第 0346 号
总 发 行	北京报刊发行局
订　　购	全国各地邮局
国外发行	中国国际图书贸易总公司 (中国国际书店) (北京 399 信箱　邮编 100044)

Superintended by	Chinese Academy of Sciences
Sponsored by	Institute of Remote Sensing Application, CAS China Society of Image and Graphics Institute of Applied Physics and Computational Mathematics
Chief editor	LI Xiaowen
Editor, Publisher	Editorial and Publishing Board of Journal of Image and Graphics (P. O. Box 9718, Beijing 100101, China) E-mail: jig@ irsa. ac. cn
Distributed by	Beijing Bureau for Distribution of Newspapers and Journals
Domestic	All Local Post Offices in China
Foreign	China International Book Trading Corporation (P. O. Box 399, Beijing 100044, China)
Printed by	Beijing Beilin Printing House

ISSN 1006-8961　CN11-3758/TB　CODE ZTTXFZ　国内邮发代号：82-831　国外发行代号：M1406　国内定价：45.00 元

Journal of Image and Graphics

(**Monthly, Started in 1996**)

Vol. 17 No. 7 July 2012

Contents

中图法分类号: TP391.41 **文献标志码**: A **文章编号**: 1006-8961(2012)07-0791-07
论文引用格式: 刘毅志,唐胜,王向东,林守勋, 张勇东. 融合音频单词与视觉特征的成人视频检测 [J]. 中国图象图形学报,2012,17(7):791-797

融合音频单词与视觉特征的成人视频检测

刘毅志[1,2,3], 唐胜[2], 王向东[2], 林守勋[2], 张勇东[2]
1. 湖南科技大学计算机科学与工程学院, 湘潭 411201; 2. 中国科学院计算技术研究所, 北京 100190;
3. 湖南省知识处理与网络化制造重点实验室, 湘潭 411201

摘 要: 基于多模态的检测方法是过滤成人视频的有效手段,然而现有方法中缺乏准确的音频语义表示方法。因此本文提出融合音频单词与视觉特征的成人视频检测方法。先提出基于周期性的能量包络单元(简称 EE)分割算法,将音频流准确地分割为 EE 的序列;再提出基于 EE 和 BoW(Bag-of-Words)的音频语义表示方法,将 EE 的特征描述为音频单词的出现概率;采用复合加权方法融合音频单词与视觉特征的检测结果;还提出基于周期性的成人视频判别算法,与基于周期性的 EE 分割算法前后配合,以充分利用周期性进行检测。实验结果表明,与基于视觉特征的方法相比,本文方法显著提高了检测性能。当误检率为 9.76% 时,检出率可达 94.44% 。

关键词: 成人视频检测;多模态融合;音频单词;视觉特征;能量包络单元

Fusing audio-words with visual features for adult video detection

Liu Yizhi[1,2,3], Tang Sheng[2], Wang Xiangdong[2], Lin Shouxun[2], Zhang Yongdong[2]
1. *Institute of Computer Science and Engineering, Hunan University of Science and Technology, Xiangtan* 411201, *China*;
2. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190, *China*;
3. *Key Laboratory of Knowledge Processing and Networked Manufacturing, College of Hunan Province, Xiangtan* 411201, *China*

Abstract: Multi-modality based adult video detection is an effective approach for filtering pornographic information. However, existing methods lack accurate representation methods of audio semantics. Therefore, a novel method is presented in this paper to fuse audio-words with visual features for adult video detection. First, we propose a periodicity-based segmentation algorithm of units of energy envelope (EE). Audio streams are divided into sequences of EE. Second, audio semantics representation method based on EE and BoW (Bag-of-Words) is presented to describe the features of the EE as the occurrence probabilities of audio-words. Integrated weighting methods are used to fuse the detection results of audio-words and visual features. Furthermore, we propose a periodicity-based decision algorithm to judge adult videos to cooperate with the preceding periodicity-based segmentation algorithm. Therefore, we make full use of the periodicity. Our experiments show that our approach remarkably improves the detection performance compared with the method based on visual features. The true positive rate achieves 94.44% while the false positive rate is 9.76%.

Key words: adult video detection; multi-modality fusion; audio-words; visual features; units of energy envelope

0 引 言

互联网上的成人信息已经泛滥成灾,其中,成人视频表达直观、内容丰富,对于青少年的危害不容忽视。目前,成人视频的过滤方法大多是基于视频关键帧(KF)的视觉特征。即:先提取视频关键帧及其视觉特征,再用分类器分析关键帧是否为成人图像,

收稿日期:2011-03-03;**修回日期**:2011-08-29
基金项目:国家重点基础研究发展计划(973)项目(2007CB311105);国家自然科学基金项目(60873165);北京市科技新星计划项目(2007B071);北京市教育委员会共建项目
第一作者简介:刘毅志(1973—),男,副教授,2011 年于中国科学院计算技术研究所获计算机应用技术专业博士学位,主要研究方向为多媒体内容分析与检索、服务计算。E-mail:liuyizhi928@gmail.com

再用阈值判决算法获得结果。阈值判别算法是指当视频中成人关键帧的数目超过设定阈值时,就判定此视频为成人视频。

基于视觉特征的检测方法实质上是一种静态图像检测方法。静态图像检测的代表性工作有 Forsyth 等人提出的基于人体轮廓的检测方法[1]、Jones 等人提出的基于肤色模型的检测方法[2]、以 WIPE 系统[3]和 Google 系统为代表的基于多特征融合的检测方法[4]。基于视觉单词的检测方法[5-6]已被用于降低肤色检测或全局特征带来的误检率。视觉单词具有一定的语义,但特征提取速度较慢。

随着互联网上低质视频的快速增长,单纯依靠视觉特征难以准确地检测成人视频。因此,融合视频中运动和音频等时域信息的检测方法逐渐成为人们关注的焦点。Endeshaw 等人[7]和 Jansohn 等人[8]较为深入地研究了利用成人视频中运动信息的周期性进行检测的方法,获得了较好效果。但是,文献[8]采用周期性检测来描述运动特征,其效果不如运动直方图。可见,运动信息的周期性没有充分发挥作用。另一方面,成人视频中运动信息的区分性不强。例如,鼓掌、俯卧撑和仰卧起坐等都难以与色情行为区分开来。

众所周知,成人视频中往往伴随着周期性的喘息、呻吟或尖叫。Rea 等人[9]通过音频流能量的自相关性分析表明,成人视频中音频流能量具有周期性。Zuo 等人[10]分别利用音频特征与视觉特征检测成人视频,分别通过最近中心分类器与高斯混合模型进行分类,再基于 Bayes 理论进行融合。其中,音频特征是采用音频帧的低层特征。

综上所述,现有的成人视频检测方法中,缺乏准确的音频语义表示方法。具体表现在 3 个方面:1)没有充分利用音频流能量的周期性;2)音频帧的尺度小,区分性较差;3)低层特征与高层概念之间存在语义鸿沟。

针对上述问题,提出融合音频单词(Audio-words)与视觉特征的成人视频检测方法。其中包括:1)提出用于音频流分割和成人视频判别的两种基于周期性的算法,充分利用音频流能量的周期性,将音频内容表示为能量包络单元(EE)的序列;2)提出基于 EE 和 BoW(Bag-of-words)的音频语义表示方法,将 EE 的特征描述为音频单词的出现概率;3)由于 EE 是一种不定长的、较大尺度的音频片段,并且音频单词是音频信息的一种中层表达,所以,音频单词直方图不仅区分性较强,而且能够有效缩小语义鸿沟。文献[11]的研究表明,大尺度音频片段的分类正确率要明显高于小尺度音频片段的分类正确率,并且这个趋势与分类器选择无关。

图 1 给出了本文方法的框架,包括基于音频单词的检测方法、基于视觉特征的检测方法和两者的融合方法 3 个部分。

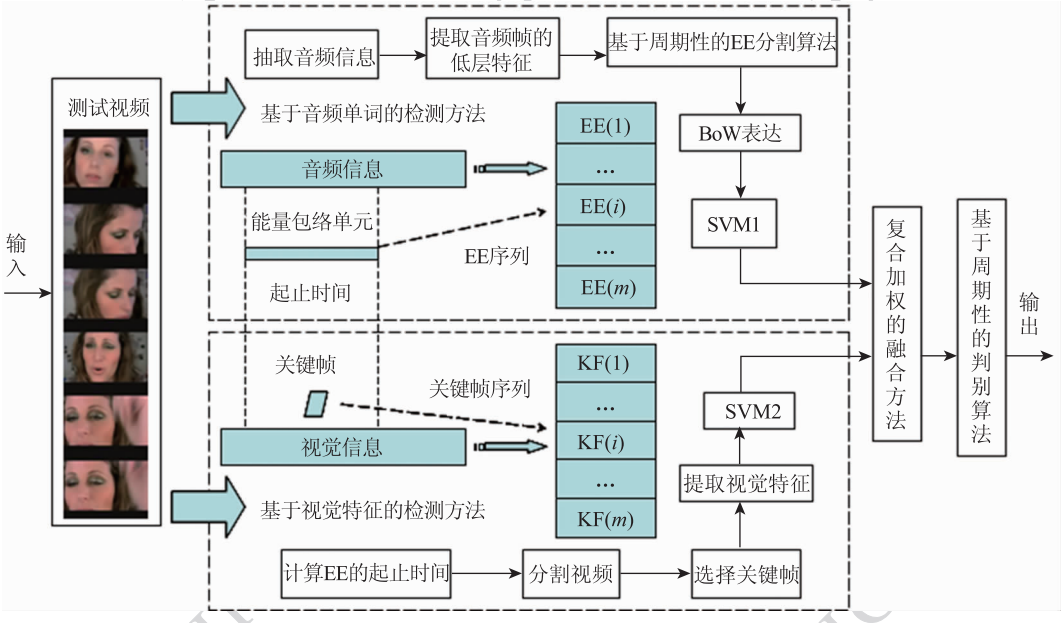


图 1 融合音频单词与视觉特征的成人视频检测方法的框架

Fig. 1 The framework of fusing audio words and visual features for adult video detection

1　基于音频单词的成人视频检测方法

本文首次将音频单词用于成人视频检测。因为音频单词在音频内容分析[12-13]、语义事件检测[14]和语义概念分类与标注[15-16]等领域表现出优良的性能。图 1 上半部分显示了基于音频单词的检测方法的主要过程:从视频中抽取音频信息,提取音频帧的低层特征,根据基于周期性的 EE 分割算法将音频流切分为 EE 的序列,采用 BoW 模型将 EE 的特征表示为音频单词的出现概率,再用支持向量机(SVM)分类。SVM 是当前性能最好的分类器之一[8]。下面详细叙述音频特征提取、音频单词生成、基于周期性的 EE 分割算法、以及基于 EE 和 BoW 的音频语义表示方法等。

1.1　音频特征提取和音频单词生成

将视频文件转换为音频文件后,无重叠地每隔 20 ms 提取一个音频帧,每帧包含 36 维音频低层特征。这些特征包括 13 维的美尔频率倒谱系数(MFCC)及其 13 维的 MFCC 差分系数、1 维过零率、1 维短时能量、4 维分带短时能量和 4 维分带短时能量比。

图 2 说明了音频单词的生成过程。在训练集的成人视频中,先粗略地剪切部分子视频,提取它们的音频信号;再从中剪辑典型的成人音频,提取其中音频帧的特征;聚类音频特征,获得音频词表 $\boldsymbol{V}_{\text{words}}$,其中的 M 个向量称为音频单词,即

$$\boldsymbol{V}_{\text{words}} = \{\boldsymbol{W}_1, \cdots, \boldsymbol{W}_M\} \tag{1}$$

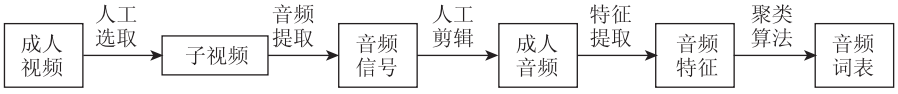


图 2　音频单词的生成过程

Fig. 2　The procedure of creating audio-words

1.2　基于周期性的 EE 分割算法

音频流分割有定长分割与变长分割两种方法。长度是指分割得到的音频片段中包含的音频帧数量。它在很大程度上影响着检测性能,所以,定长分割方法需要进行长度调节才能获得较好的性能。EE 分割是一种变长分割方法。现有方法根据短时能量将音频流分割为一个个完整的声学信号单元,速度快且较鲁棒[17]。

但是,在分割成人视频时,经常出现长度较大的 EE。因此,提出基于周期性的 EE 分割算法。具体如下:

输入:音频特征文件,音频片段的最大长度 $L_{\max}=100$,阈值 T_1 和 T_2;

输出:切分点的位置(即对应的音频帧号)。

1)采用中值滤波的方法对能量包络波形图进行平滑处理;

2)根据式(2)的检测函数,计算每一个音频帧 i 对应的检测函数值

$$d_i = \max_{j=1,\cdots,J}(E_{i+j}/E_i) \tag{2}$$

式中,E_i 表示第 i 帧对应的短时能量值,J 为求检测函数时选取的后续帧数(经验值为 10);

3)两个切分点之间的能量包络称为 EE。当 $d_i \leqslant T_1$ 时,当前点不是切分点;$d_i \geqslant T_2$ 时,当前点一定是切分点;$T_1 < d_i < T_2$ 时,$P(i)$ 为切分点,即

$$P(i) = \begin{cases} 1 & d_i \geqslant T_2 \\ \dfrac{d_i - T_1}{T_2 - T_1} & T_1 < d_i < T_2 \\ 0 & d_i \leqslant T_1 \end{cases} \tag{3}$$

4)如果两个切分点之间的音频片段长度(即音频帧数目)大于 $L_{\max}$,那么通过统计方法预测 EE 内部的周期 T,并按照 T 将其分割为长度较小的音频片段。周期 T 是指已经切分的 EE 的平均长度。

该算法利用成人视频中音频流能量的周期性,更准确地将音频流分割为 EE 的序列,为准确表达音频内容奠定了基础。

1.3　基于 EE 和 BoW 的音频语义表示方法

该方法将 EE 特征描述为音频单词出现概率。其目的有两个方面:其一,音频单词具有一定的语义,能够有效缩小语义鸿沟,其二,将不同长度的 EE 转化为相同特征维数的音频单词直方图,以便于 SVM 进行分类。

将训练集 $\boldsymbol{V}_{\text{train}}$ 分为 a 个正样本视频 $\{\boldsymbol{V}_{P1}, \cdots, \boldsymbol{V}_{Pa}\}$ 与 b 个负样本视频 $\{\boldsymbol{V}_{N1}, \cdots, \boldsymbol{V}_{Nb}\}$。提取它们的音频特征,切分为无重叠的 EE。设其中任意一个视频 $\boldsymbol{V}$,包含 p 个 EE(用 $\boldsymbol{U}$ 表示),则

$$\boldsymbol{V}=\{\boldsymbol{U}_1,\cdots,\boldsymbol{U}_p\} \tag{4}$$

设其中某个 $\boldsymbol{U}$ 包含 L 个音频帧,则 $\boldsymbol{U}$ 的长度为 L。

$$\boldsymbol{U}=\{\boldsymbol{F}_1,\cdots,\boldsymbol{F}_L\} \tag{5}$$

将 $\boldsymbol{U}$ 的每一个音频帧和音频词表 $\boldsymbol{V}_{\text{words}}$ 的每一个音频单词相匹配,得到结果 $\boldsymbol{S}$。即

$$\begin{aligned}&\boldsymbol{S}=Match(\boldsymbol{U},\boldsymbol{V}_{\text{words}})=\\&\{Match(\boldsymbol{F}_1,\boldsymbol{V}_{\text{words}}),\cdots,Match(\boldsymbol{F}_L,\boldsymbol{V}_{\text{words}})\}\end{aligned} \tag{6}$$

对于任意一个音频帧 $\boldsymbol{F}_j$,根据式(1)有

$$\begin{aligned}Match(\boldsymbol{F}_j,\boldsymbol{V}_{\text{words}})&=Match(\boldsymbol{F}_j,\{\boldsymbol{W}_1,\cdots,\boldsymbol{W}_M\})=\\&\{D_{j,1},\cdots,D_{j,M}\}\end{aligned} \tag{7}$$

如果音频帧 $\boldsymbol{F}_j$ 与音频单词的欧氏距离中,第 k 个($k=\{1,\cdots,10\}$)最小,表明音频帧 $\boldsymbol{F}_j$ 与第 k 个音频单词最相似,则该音频帧的 M 维特征中,第 k 维记作 1,其他维记作 0。即

$$D_{j,k}=\begin{cases}1 & k=\underset{k}{\operatorname{argmin}}E(\boldsymbol{F}_j,\{\boldsymbol{W}_1,\cdots,\boldsymbol{W}_k,\cdots,\boldsymbol{W}_M\})\\0 & \text{其他}\end{cases} \tag{8}$$

式中,$E(A,B)$ 表示 A 与 B 的欧氏距离。这样,音频帧 $\boldsymbol{F}_j$ 就可以表示为一个 M 维的特征序列,每一维由 0 或 1 组成,表示对应的音频单词是否出现。即

$$\begin{aligned}\boldsymbol{D}_j&=\{D_{j,1},\cdots,D_{j,k-1},D_{j,k},D_{j,k+1},D_{j,M}\}=\\&\{0,\cdots,0,1,0,\cdots,0\}\end{aligned} \tag{9}$$

为了避免长度的影响,将不定长的 EE 的特征或定长的音频片段的特征都表示为各个音频单词在其中的出现概率。即将 $\boldsymbol{U}$ 中 $\boldsymbol{L}$ 个音频帧的特征按列叠加,其和除以长度 $\boldsymbol{L}$。

$$\boldsymbol{U}=\left\{\frac{\sum_{j=1}^{L}D_{j,1}}{\boldsymbol{L}},\cdots,\frac{\sum_{j=1}^{L}D_{j,M}}{\boldsymbol{L}}\right\} \tag{10}$$

训练集 $\boldsymbol{V}_{\text{train}}$ 中,正样本视频的所有 EE 特征前标注为 1,负样本视频的所有 EE 特征前标注为 -1,将它们的顺序打乱后进行 SVM 训练以获取 EE 的 SVM 分类模型。

2 音频单词与视觉特征的融合方法

图 1 下半部分说明了基于视觉特征的检测方法。其主要过程是:根据 EE 对应的起止时间,选择离其中心最近的关键帧(KF),提取视觉特征后,用 SVM 分类。视觉特征是指图像的各种特征,包括全局特征和视觉单词直方图等。为了不失一般性,选择颜色矩与边缘纹理的前融合作为视觉特征。图 1 右部采用复合加权的融合方法,还提出基于周期性的成人视频判别算法。

2.1 复合加权的融合方法

在实验中发现:基于音频单词的方法查准率较高,查全率偏低;基于视觉特征的方法查全率较高,查准率偏低。因此,先将两种检测方法的 SVM 结果进行加权平均,再采用最大值融合方法。本文称之为复合加权的融合方法。

如果 $\text{Method}_{\text{EE}}$ 表示基于音频单词的方法,$\text{Method}_{\text{KF}}$ 表示基于视觉特征的方法,$\boldsymbol{R}(\boldsymbol{V})$ 表示视频 $\boldsymbol{V}$ 的检测结果,W 表示权值,那么

$$\boldsymbol{R}_{\text{Method}_{\text{EE}}}(\boldsymbol{V})=(v_1^{(1)},\cdots,v_m^{(1)}) \tag{11}$$

$$\boldsymbol{R}_{\text{Method}_{\text{KF}}}(\boldsymbol{V})=(v_1^{(2)},\cdots,v_m^{(2)}) \tag{12}$$

式中,$(v_1^{(1)},\cdots,v_m^{(1)})$ 分别为 SVM1 分类器的 m 个预测值,$(v_1^{(2)},\cdots,v_m^{(2)})$ 分别为 SVM2 分类器的 m 个预测值。加权融合过程为

$$\begin{aligned}\boldsymbol{R}_W(\boldsymbol{V})&=W\times\boldsymbol{R}_{\text{Method}_{\text{EE}}}(\boldsymbol{V})+(1-W)\times\\&\boldsymbol{R}_{\text{Method}_{\text{KF}}}(\boldsymbol{V})=W\times(v_1^{(1)},\cdots,v_m^{(1)})+\\&(1-W)\times(v_1^{(2)},\cdots,v_m^{(2)})\end{aligned} \tag{13}$$

最终融合结果 $\boldsymbol{R}(\boldsymbol{V})$ 是式(11)(13)的最大值。为了获得更好的性能,结合下面的成人视频判别算法调整权值。

$$\boldsymbol{R}(\boldsymbol{V})=MAX(\boldsymbol{R}_{\text{Method}_{\text{EE}}}(\boldsymbol{V}),\boldsymbol{R}_W(\boldsymbol{V})) \tag{14}$$

2.2 基于周期性的成人视频判别算法

为了配合基于周期性的 EE 分割算法,更充分地利用音频流能量的周期性获取更好的性能,提出基于周期性的成人视频判别算法。因为含有成人内容的音频流能量在一段时间内总是连续出现,所以,当含有成人内容的 EE 连续出现 $N(1\leqslant N\leqslant 10)$ 次时,则认为它们所在的视频就是成人视频;否则,为非成人视频。调节连续次数 N,可以减少个别音频片段误判带来的干扰。

3 实验与分析

从互联网上收集视频建立训练集和测试集。训练集有 48 个成人视频和 300 个非成人视频,用于训练 SVM 分类模型。测试集有 50 个成人视频和 150 个非成人视频。实验使用的机器配置是 1.86 G Hz 的双核 CPU 和 2 G 内存,在 visual studio 2003 环境下实现。性能评价指标选用 ROC(R)曲线,其横坐标是误检率(FPR),纵坐

标是检出率(TPR)。

分类器都采用 RBF 核的 SVM。

3.1　基于音频单词的检测方法

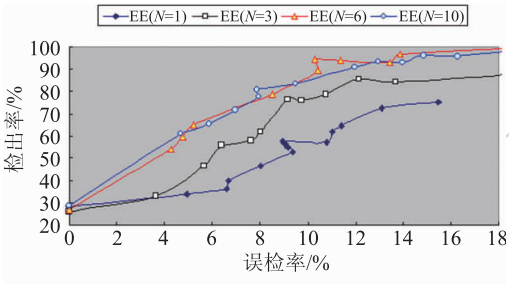
将视频转换为音频,所有音频数据均为采样率为 44 100/s、量化位数为 16 的混和单声道 WAV 格式。如表 1 所示,采用基于周期性的 EE 分割算法将训练集与测试集分别切分为 60 941 个和 69 211 个 EE。

表 1　本文数据库的相关统计

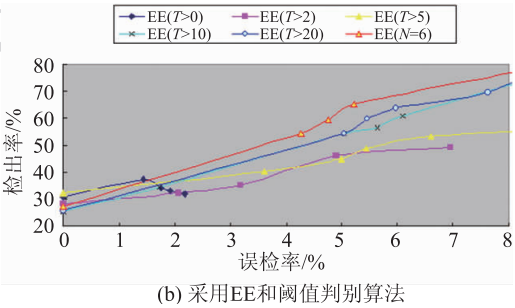
Table 1　The statistics of our dataset

	成人视频数量/个	非成人视频数量/个	作　用	EE 数量/个
训练集	20	0	构建音频单词	0
	28	300	训练 SVM 分类模型	61 941
测试集	50	150	测试检测性能	69 211

首先,评估基于音频单词的检测方法,其 ROC 曲线如图 3 所示。图 3(a)给出了基于周期性的判别算法中调节连续次数 $N(1\leqslant N\leqslant 10)$ 获得的性能,而图 3(b)显示了传统的阈值判别算法在阈值 T 从 0 调整到 30 的性能。其中,EE($N=6$)效果较好。不难看出,与传统的阈值判别算法相比,基于周期性的判别算法能够较大地提高检测性能。



(a) 采用EE和基于周期性的判别算法

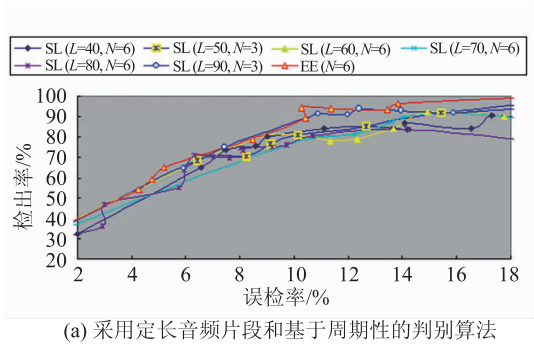


(b) 采用EE和阈值判别算法

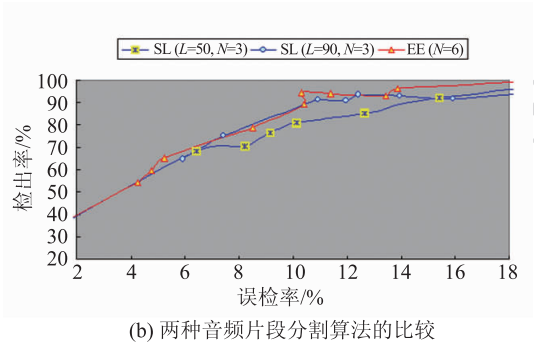
图 3　基于音频单词的检测方法中两种成人视频判别算法的比较

Fig. 3　Comparisons of decision algorithms in adult video detection based on audio-words

其次,评估两种音频片段的分割算法。如图 4(a)所示,定长的音频片段分割算法中,分别选取切分长度 $L=10,20,\cdots,100$ 和连续次数 $N=1,3,6,10$ 进行实验,其 ROC 曲线表示为 SL($\boldsymbol{L}$,N)。图 4(b)显示了两种分割算法的性能对比,基于周期性的 EE 分割算法优于定长的音频片段分割算法,无需调节切分长度即可获得更好性能。



(a) 采用定长音频片段和基于周期性的判别算法



(b) 两种音频片段分割算法的比较

图 4　基于音频单词的检测方法中两种分割算法的性能比较

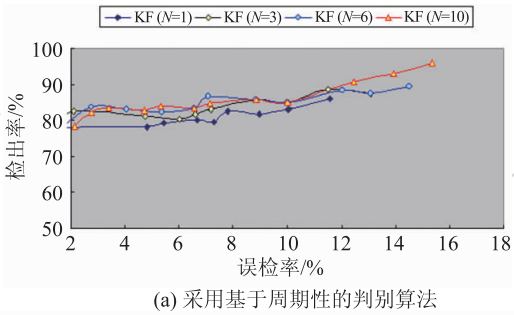
Fig. 4　Comparisons of segmentation algorithms in adult video detection based on audio-words

但是,由于每个音频帧要和音频词表内的每个音频单词相匹配,这会增加计算时间。

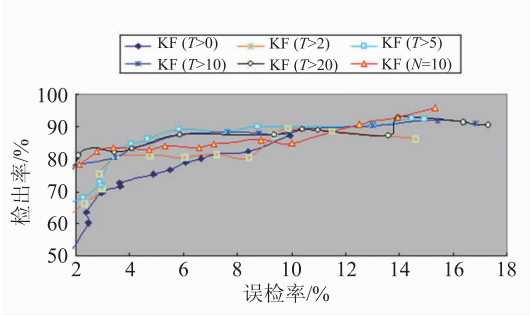
3.2　基于视觉特征的检测方法

从训练集视频中抽取关键帧,经人工标注后得到 8 083 幅成人图像与 31 616 幅非成人图像,将它们的颜色矩与边缘纹理前融合,再用 SVM 训练分类模型。

如图 5(a)所示,在基于视觉特征的检测方法中采用基于周期性的判别算法,分别选取连续次数 $N=1,3,6,10$ 进行实验,其中,KF($N=10$)效果较好。图 5(b)采用阈值判别算法,阈值 T 由 0 到 20 变化,随着阈值的增加,开始时性能提高较快,阈值达到某个值时性能不再提高,并有可能下降。这里,KF($T>5$)效果较好。



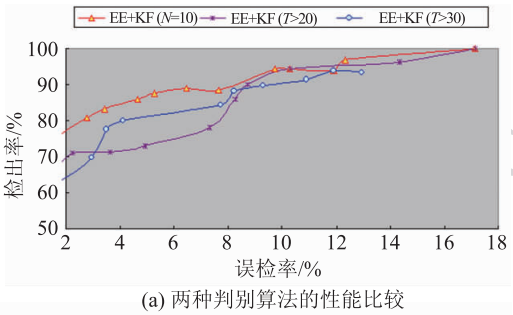
(a) 采用基于周期性的判别算法



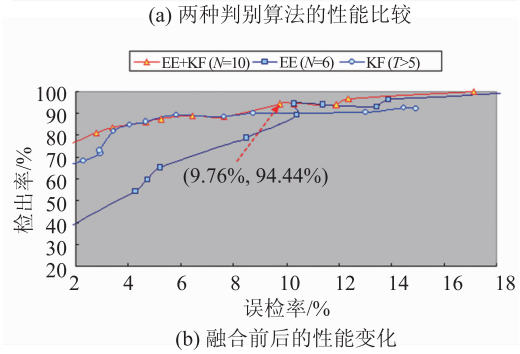
(b) 采用阈值判别算法

图 5 基于视觉特征的成人视频检测方法

Fig. 5 Adult video detection based on global features of keyframes



(a) 两种判别算法的性能比较



(b) 融合前后的性能变化

图 6 融合音频单词和视觉特征的成人视频检测方法

Fig. 6 Fusing audio-words and global features for adult video detection

3.3 融合音频单词与视觉特征的检测方法

利用 2.1 节方法融合音频单词和视觉特征的 SVM 分类结果,再用 2.2 节方法判别成人视频。其 ROC 曲线表示为 EE + KF。

由图 6(a)可知,当权值 $W = 0.39$ 且连续次数 $N = 10$ 时,EE + KF($N = 10$)性能优良。

从图 6(b)看出,基于音频单词的检测方法能够获得较高的检出率,但随着误检率的变化波动很大;基于视觉特征的检测方法的检出率偏低,但性能较稳定。

融合两种方法之后,性能显著提高,当误检率为 9.76% 时,检出率达到了 94.44% 。

4 结 论

由于现有的成人视频检测方法缺乏准确的音频语义表示方法,造成检测性能仍然偏低。因此,本文提出融合音频单词与视觉特征的成人视频检测方法,主要创新点有基于 EE 和 BoW 的音频语义表示方法、基于周期性的 EE 分割算法与成人视频判别算法。

利用 EE 特征中音频单词的出现概率表示音频内容,再和视觉特征进行复合加权融合。实验结果表明,与基于视觉特征的检测方法相比,该方法显著提高了成人视频的检测性能。当误检率为 9.76% 时,检出率可达 94.44% 。

下一步工作是,充分利用视频中的运动信息,将它与音频、视觉等模态信息进行融合,进一步提高检测性能。

参考文献(References)

[1] Forsyth D A, Fleck M M. Automatic detection of human nudes [J]. International Journal of Computer Vision, 1999, 32 (1): 63-77.

[2] Jones M J, Rehg J M. Statistical color models with application to skin detection [J]. International Journal of Computer Vision, 2002, 46 (1): 81-96.

[3] Wang J Z, Li J, Wiederhold G, et al. System for screening objectionable images[J]. Computation Communication, 1998, 21: 1355-1360.

[4] Rowley H A, Yushi J, Baluja S. Large scale image-based adult-content filtering[C] // Proceedings of the 1st International Conference on Computer Vision Theory and Applications. New York City: Springer, 2006: 290-296.

[5] Wang Y S, Li Y N, Gao W. Detecting pornographic images with visual words[J]. Transactions of Beijing Institute of Technology, 2008, 28 (5): 410-413.[王宇石, 李远宁, 高文. 基于局部视觉单词分布的敏感图像检测[J]. 北京理工大学学报, 2008, 28 (5): 410-413.]

[6] Deselaers T, Pimenidis L, Ney H. Bag-of-visual-words models for adult image classification and filtering[C] // Proceedings of the 19th International Conference on Pattern Recognition. Tampa, USA: University of South Florida, 2008: 1-4.

[7] Endeshaw T, Garcia J, Jakobsson A. Classification of indecent video by low complexity repetitive motion detection[C] // Proceedings of 37th Applied Imagery Pattern Recognition Workshop. Washington DC, USA: IEEE Computer Society Press, 2008.

[8] Jansohn C, Ulges A, Breuel T M. Detecting pornographic video contents by combining image features with motion information [C]// Proceedings of the 17th ACM International Conference on Multimedia. New York City, USA: ACM Press, 2009: 601-604.

[9] Rea N, Lacey G, Lambe C, et al. Multimodal periodicity analysis for illicit content detection in videos[C]// Proceedings of 3rd European Conference on Visual Media Production. Washington D. C., USA: IEEE Computer Society Press, 2006: 106-114.

[10] Zuo H Q, Wu O, Hu W M, et al. Recognition of blue movies by fusion of audio and video[C] // Proceedings of IEEE International Conference on Multimedia & Expo. Washington DC, USA: IEEE Computer Society Press, 2008: 37-40.

[11] Zhang Y B, Zhou J, Bian Z Q, et al. A two-stage content-based audio segmentation algorithm[J]. Chinese Journal of Computers, 2006, 29(3): 457-465.[张一彬, 周杰, 边肇祺, 等. 一种基于内容的音频流二级分割方法[J]. 计算机学报, 2006, 29(3): 457-465.]

[12] Lu L, Hanjalic A. Towards optimal audio "keywords" detection for audio content analysis and discovery[C] // Proceedings of the 14th ACM International Conference on Multimedia. New York City, USA: ACM Press, 2006: 825-834.

[13] Lu L, Hanjalic A. Audio keywords discovery for text-like audio content analysis and retrieval[J]. IEEE Transactions on Multimedia, 2008, 10 (1): 74-85.

[14] Xu M, Xu C, Duan L Y, et al. Audio keywords generation for sports video analysis[J]. ACM Transactions on Multimedia Computing, Communications and Applications, 2008, 4 (2): Article 11.

[15] Peng Y X, Lu Z W, Xiao J G. Semantic concept annotation based on audio PLSA model[C] // Proceedings of the 17th ACM International Conference on Multimedia. New York City, USA: ACM Press, 2009: 841-844.

[16] Zeng Z, Zhang S W, Li H P, et al. A novel approach to musical genre classification using probabilistic latent semantic analysis model [C] // Proceedings of IEEE International Conference on Multimedia & Expo. Washington DC, USA: IEEE Computer Society Press, 2009: 486-489.

[17] Zhao D, Wang X D, Qian Y L, et al. Fast commercial detection based on audio retrieval [C] // Proceedings of IEEE International Conference on Multimedia & Expo. Washington DC, USA: IEEE Computer Society Press, 2008: 1185-1188.