基于多种上下文结构与线性融合的特定敏感视频识别

王方石[1],王辰龙[1],李 兵[2],丁昕苗[3],胡卫明[2]
(1.北京交通大学软件学院,北京 100044; 2.中国科学院自动化研究所模式识别国家重点实验室,北京 100190;
3.山东工商学院,山东烟台 264005)

摘 要: 本文中特定敏感视频是指恐怖和暴力视频,现有的特定敏感视频识别算法或是忽略了视频的多种上下文结构信息;或是忽略了各种特征间潜在的依赖关系.因此,本文提出了一种基于多种上下文结构与线性融合的特定敏感视频识别方法,首先针对某种视频提取多种有效特征,并获取镜头间的上下文结构信息;然后,在每一个特征空间中利用上下文结构训练一个SVM分类器;最后,获取不同特征间的依赖关系,采用线性依赖模型融合多个分类器的结果,提高视频的识别率.在特定敏感视频库上的实验结果验证了该方法比现有的其它算法有更好的性能和稳定性.

关键词: 特定敏感视频; 多种上下文结构; 分类融合; 线性依赖模型
中图分类号: TP37 **文献标识码:** A **文章编号:** 0372-2112 (2015)04-0675-09
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.04.008

Specified Sensitive Video Recognition Based on Multi-Context Construction and Linear Fusion

WANG Fang-shi[1], WANG Chen-long[1], LI Bing[2], DING Xin-miao[3], HU Wei-ming[2]
(1. *School of Software Engineering, Beijing Jiaotong University, Beijing* 100044, *China*;
2. *National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing* 100190, *China*;
3. *Shandong Institute of Business and Technology, Yantai, Shandong* 264005, *China*)

Abstract: Along with the ever-growing Web, specified sensitive videos including horror videos and violent videos are disseminated over Internet and have threatened children's psychological health. It is necessary to effectively recognize and filter out these videos. So far, the existing recognition methods for specified sensitive videos either ignore the multiple contextual information among the shots, or ignore the dependent relationship among the multiple features. This paper presents a novel recognition method for specified sensitive video based on multi-context construction and linear fusion. First, multiple effective features are extracted and the multiple contextual structure graphs are constructed for the shots in one video clip. Then, a SVM classifier is trained using contextual information in each feature space. Finally, a linear dependency model is learnt to fuse multiple classifiers obtained in different feature spaces. Experiments on a specified sensitive video dataset demonstrate that the performance of our method is superior to the other existing algorithms.

Key words: specified sensitive video; multi-context construction; classification fusion; linear dependency model

1 引言

互联网的开放性及网上信息内容的参差不齐,使得各种不良信息也随之泛滥,特别是反动、色情、暴力、恐怖等敏感信息极大地危害着青少年的身心健康和社会的稳定.

现有的研究更多地关注反动、色情等信息的过滤.实际上,心理学和生理学研究表明,恐怖和暴力信息对青少年身心健康的危害绝不亚于色情信息的危害.

恐怖视频可唤起观众消极、恐惧情绪的反应,主要分为:连环杀人、鬼魂、怪兽、吸血鬼、杀人动物和反宗教等.暴力视频则表现在使用武力对人、对己产生伤害或疼痛,造成恐慌情绪,主要分为:打斗、枪击、爆炸和自残等.

收稿日期:2014-09-01;修回日期:2014-11-24;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61370038,No.61303086)

有效、自动地过滤恐怖和暴力视频是一个亟待解决的新课题,该课题的研究具有广泛的应用前景和重要的现实意义.

2 相关工作

与文字、图片类恐怖和暴力信息相比,特定敏感视频的视觉冲击力更强、危害更大、识别难度也更高,是当前网络特定敏感信息过滤的难点之一.

2.1 恐怖视频识别的相关工作

早期的研究工作主要是根据情感对视频进行分类[1~3],其中一个类别为"恐怖",并没有专门识别恐怖视频的方法.例如,Rasheed 等[1]提出基于视觉特征的电影分类框架,把电影分为动作、喜剧、恐怖、戏剧等不同风格.Kang[2]等先提取视频的底层特征,然后采用隐马尔科夫模型将视频场景分为三种情感类别:高兴、恐惧和悲伤.Xu 等[3]提出了基于隐马尔科夫模型的方法,用于检测电影中大笑和尖叫等情感事件,以此来区分喜剧和恐怖电影.

近年来,已有研究者开始关注专门的恐怖视频识别方法.Wang 等[4]从音频流中提取音频特征,从视频关键帧中提取视觉特征和颜色情感特征,然后将各种特征进行融合,并以支持向量机为分类器来识别恐怖视频.Wang[5]等发现:恐怖视频场景中包含多个镜头,其中至少一个是恐怖镜头,而非恐怖场景中不包含任何恐怖镜头.该特点恰好符合多示例学习要解决的问题,于是将视频场景作为多示例学习中的"包"、镜头作为"包"中的"示例",采用多示例学习方法来识别恐怖视频.

在传统多示例学习框架中,假定示例彼此相互独立,这种假设只能反映恐怖视频帧之间的一个特性.实际上,恐怖视频同一场景的镜头间存在着上下文语义关联,某些镜头组合在一起才能更好地识别其中的语义信息.因此,Ding 等[6]引入代价敏感的稀疏编码模型模拟视频帧之间、音视频之间的上下文结构.随后,Ding[7]又引入多视角融合稀疏表示模型,该模型分别从集合、上下文以及统计特性三个不同视角来描述一个视频片段,并利用联合稀疏表示框架将三个不同视角融合到一个分类框架中,用以识别恐怖视频.

2.2 暴力视频识别的相关工作

早期暴力视频识别方法多采用基于音、视频信息的多模态分类策略.Datta 等[8]利用加速运动矢量来检测电影中打斗的暴力场景.Giannakopoulos 等[9]提出基于频域和时域的七种音频特征的暴力视频检测方法.Nam 等[10]结合音、视频特征来检测暴力场景中的火焰、血液等画面和声音.Lin 等[11]先采用音频特征检测暴力,然后采用视频特征检测发生火焰、爆炸和流血的暴力场景,最后对多种分类器进行协同训练,用以识别暴力视频.Giannakopoulos 等[12]采用多模态特征,将音频、视频、评价文本特征拼接为一个 10 维的特征向量,用以在视频共享网站中识别暴力视频.Wang 等[13]提出一个新方法检测视觉暴力,通过学习稠密运动轨迹得到慢特征函数,然后进行差异性的慢特征分析(Discriminative Slow Feature Analysis),最后采用 SVM 分类器区分暴力与非暴力视频.

在上述工作中,大部分恐怖视频识别方法及全部暴力视频识别方法均假定镜头间相互独立,忽略了视频内部的上下文结构信息.实际上,视频各镜头间存在时间上的连续性[14].有研究工作[15]表明:发现视频中固有的时间上下文结构有助于理解视频内容.另外,虽然文献[6,7]考虑了视频镜头间的关联性,但只是对多种不同特征做了简单的拼接,而忽略了各种特征间潜在的依赖关系及其在分类中作用的差别.通过对大量恐怖和暴力视频的观察和研究,我们发现简单拼接有时会弱化这些特征的刻画性能,而且各种特征间并非毫无关联,不同种类的特征对于分类结果所起的作用也不同.

3 基于多种上下文结构分类融合

为解决上述问题,我们提出了一种基于多种上下文结构与线性融合的特定敏感视频识别方法,简记为 MCS-LDF(Multiple Context Structure & Linear Dependency Fusion).本文贡献有两点:

第一,该方法不但充分考虑了多个特征空间中镜头间的上下文相关性,而且兼顾了不同特征间的依赖关系及各种特征在分类中的不同作用.

第二,将该方法分别应用于恐怖和暴力视频数据集上,实验结果及数据分析对以后特定敏感视频识别具有指导意义.

3.1 整体结构框架

MCS-LDF 方法的整体结构框架如图 1 所示.首先,对视频提取 M 种不同特征,并在 M 个特征空间中分别分析镜头间的关联性,构造镜头间上下文结构图 ε-Graph;然后,针对每种特征训练一个基于上下文结构的 SVM 分类器;最后通过学习获取不同特征间的依赖关系,即各种特征在分类中作用的权重,创建一个基于线性依赖的多分类器融合模型,进行特定敏感视频的识别.

3.2 基于多种上下文结构的分类器

上下文结构是指每段视频中镜头与镜头之间的一种结构关系.假设给定 N 个视频片段 $V_1, V_2, \cdots, V_N$ 及其标注 $y_1, y_2, \cdots, y_N$($y_i \in \{-1, +1\}$),其中每个视频片段 V_i 被划分为 n_i 个镜头 $S_{i,1}, S_{i,2}, \cdots, S_{i,ni}$,从每个镜头中提取中间帧作为关键帧,则视频 V_i 由一组关键帧 $\{F_{i,1},$

$F_{i,2},\cdots,F_{i,ni}\}$表示.针对每个关键帧 $F_{i,j}$,提取 M 种特征,其中第 m 个特征记为 $x_{i,j}^m(i=1,\cdots,N;\ j=1,\cdots,n_i;\ m=1,\cdots,M)$,则 V_i 在第 m 个特征空间中的特征向量记为 $\boldsymbol{X}_i^m=(x_{i,1}^m,\cdots,x_{i,j}^m,\cdots,x_{i,n_i}^m)$.

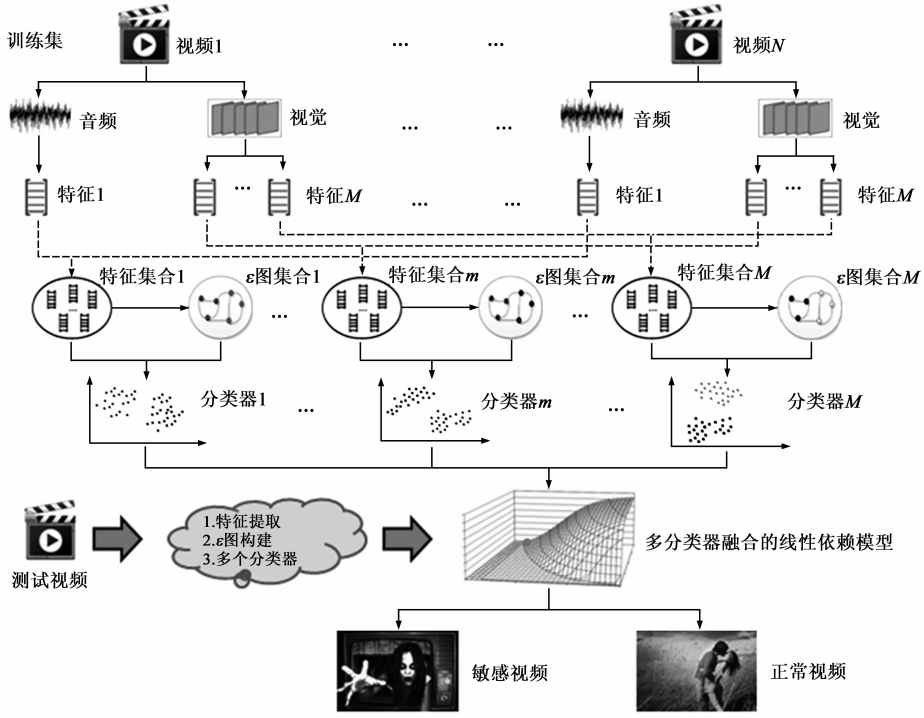

图1　本方法整体结构框架

ε-Graph 是一种描述上下文关系的比较有效的方法[15],在第 m 个特征空间中为视频片段 V_i 建立镜头间的结构图 G_i^m,以 n_i 个镜头为图中结点,其相似性权重矩阵记为 $\boldsymbol{W}_i^m\in\boldsymbol{R}^{n_i\times n_i}(m=1,\cdots,M;i=1,\cdots,N)$.计算 V_i 中任意两个镜头特征 $x_{i,j}^m$ 和 $x_{i,k}^m(j,k=1,\cdots,n_i)$间的距离,若该距离小于预设阈值 ε,则在 G_i^m 中建立这两个镜头结点之间的边,并在矩阵 $\boldsymbol{W}_i^m$ 中将该边的权重 $w_{i,j,k}^m$ 置为 1,否则置为 0.每个视频 V_i 在 M 个特征空间中共有 M 个上下文结构图$\{G_i^1,\cdots,G_i^m,\cdots,G_i^M\}$.

我们在 M 个特征空间中得到了 N 个训练样本的描述,其中在第 m 个特征空间中,N 个训练样本可表示为$\{(X_1^m,G_1^m,y_1),\cdots,(X_i^m,G_i^m,y_i),\cdots,(X_N^m,G_N^m,y_N)\}$.由于上下文结构图 G_i^m 无法直接在特征空间进行分类,于是,借助映射函数 $\varphi:G\to\boldsymbol{R}^d$,将镜头特征投影到一个高维空间,即:$G\to\varphi(G)$.得到高维空间特征 $\boldsymbol{F}^m=[\varphi(G_1^m),\cdots,\varphi(G_k^m),\cdots,\varphi(G_N^m)]$,$\varphi(G_k^m)\in\boldsymbol{R}^d$.

为在此空间进行样本分类,建立 SVM 分类器,参考文献[15],径向基核函数定义如下:

$$K_{\text{graph}}(G_i^m,G_j^m)=[\varphi(G_i^m)]^{\mathrm{T}}\varphi(G_j^m)=\frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j}\omega_{ia}^m\omega_{jb}^m K_I(x_{i,a}^m,x_{i,b}^m)}{\sqrt{\sum_{a=1}^{n_i}\omega_{ia}^m}\sqrt{\sum_{b=1}^{n_j}\omega_{jb}^m}}$$

$$K_I(x_{i,a}^m,x_{i,b}^m)=\exp(-\gamma\parallel x_{i,a}^m-x_{i,b}^m\parallel^2)\tag{1}$$

其中,$\omega_{ia}^m=1\Big/\sum_{u=1}^{n_i}w_{i,a,u}^m$,$\omega_{jb}^m=1\Big/\sum_{u=1}^{n_j}w_{j,b,u}^m$.

至此,在每个特征空间得到一个基于上下文结构的 SVM 分类器,其输出结果为视频 V_i 在第 m 个特征空间中属于 c_l 类的概率,记为 $\Pr(c_l\mid\boldsymbol{X}_i^m)$,$m=1,\cdots,M$,$l=1,\cdots,L$,$L$ 为类别数.

3.3　基于线性依赖分类器融合模型

由于不同特征空间训练出的分类器不同,针对同一个视频,会产生不同的预测结果,因此需要进行分类结果融合.

为简化问题,通常假设这些分类器的结果是条件独立的,根据贝叶斯理论有:

$$\Pr(c_l\mid X_i^1,\cdots,X_i^M)=\frac{\Pr(c_l)\Pr(X_i^1,\cdots,X_i^M\mid c_l)}{\Pr(X_i^1,\cdots,X_i^M)}=\frac{P_0}{\Pr(c_l)^{M-1}}\prod_{m=1}^{M}\Pr(c_l\mid\boldsymbol{X}_i^m)\tag{2}$$

其中　$P_0=\dfrac{\prod_{m=1}^{M}\Pr(\boldsymbol{X}_i^m)}{\Pr(X_i^1,\cdots,X_i^M)}$

文献[16]实验结果显示:采用加法法则融合多个独立分类器的效果最好.假设后验概率与先验概率间的

偏差不大,令 $\delta_l^m(\delta_l^m \ll 1)$表示两者间的偏离,则后验概率可表示为:

$$\Pr(c_l \mid \boldsymbol{X}_i^m) = \Pr(c_l)(1+\delta_l^m) \tag{3}$$

将式(3)带入式(2)有:

$$\begin{aligned}\Pr(c_l \mid x_i^1,\cdots,x_i^M) &= \frac{P_0}{\Pr(c_l)^{M-1}}\Big\{\prod_{m=1}^{M}\Pr(c_l)(1+\delta_l^m)\Big\}\\ &= P_0 * \Pr(c_l) * \prod_{m=1}^{M}(1+\delta_l^m)\end{aligned}$$

由于 δ_l^m 值较小,故可忽略 $\prod_{m=1}^{M}(1+\delta_l^m)$中的二阶项和更高阶项,则得到基于独立假设的多分类器融合模型如下:

$$\begin{aligned}&\Pr(c_l \mid X_i^1,\cdots,X_i^M)\\ &= P_0\Big[(1-M)\Pr(c_l) + \sum_{m=1}^{M}\Pr(c_l \mid \boldsymbol{X}_i^m)\Big]\end{aligned} \tag{4}$$

事实上,在一些实际应用中独立假设并不成立[17],下面给出去除独立假设的分类融合方法.尽管 Ma 等人[18]指出特征级融合的效果比分类器级融合性能略好,但又指出前者耗时远长于后者.考虑到此模型将应用于网络特定敏感视频的识别,运行速度是一个重要指标,因此,经过综合权衡,决定采用分类器级融合方法.文献[18]给出式(5)中的线性依赖模型,用以融合多个分类器.

$$\begin{aligned}&\Pr(c_l \mid x_i^1,\cdots,x_i^M)\\ &= P_0 * \Big(\sum_{m=1}^{M}(1+\alpha_l^m)[\Pr(c_l \mid x_i^m) - \Pr(c_l)] + \Pr(c_l)\Big)\end{aligned} \tag{5}$$

其中 $\alpha_l^m(0 \leqslant \alpha_l^m \leqslant 1, l = 1,\cdots,L,\ m = 1,\cdots,M)$ 是样本 V_i 在第 m 个特征空间中属于 c_l 类的权值.

我们对式(5)进行如下变换:

$$\begin{aligned}&\Pr(c_l \mid x_i^1,\cdots,x_i^M)\\ &= P_0 * \Big(\sum_{m=1}^{M}(1+\alpha_l^m)\Pr(c_l \mid x_i^m)\\ &\quad - \sum_{m=1}^{M}(1+\alpha_l^m)\Pr(c_l) + \Pr(c_l)\Big)\\ &= P_0 * \Big(\sum_{m=1}^{M}\Pr(c_l \mid x_i^m) + \sum_{m=1}^{M}\alpha_l^m\Pr(c_l \mid x_i^m)\\ &\quad - M * \Pr(c_l) - \sum_{m=1}^{M}\alpha_l^m\Pr(c_l) + \Pr(c_l)\Big)\\ &= P_0 * \Big((1-M)\Pr(c_l) + \sum_{m=1}^{M}\Pr(c_l \mid x_i^m)\\ &\quad + \sum_{m=1}^{M}\alpha_l^m[\Pr(c_l \mid x_i^m) - \Pr(c_l)]\Big)\end{aligned}$$

则得到基于线性依赖的多分类器融合模型如下:

$$\begin{aligned}&\Pr(c_l \mid X_i^1,\cdots,X_i^M)\\ &= P_0\Big[(1-M)\Pr(c_l) + \sum_{m=1}^{M}\Pr(c_l \mid \boldsymbol{X}_i^m) + D_i^l\Big]\end{aligned} \tag{6}$$

其中 D_i^l 表示为:

$$D_i^l = \sum_{m=1}^{M}\alpha_l^m[\Pr(c_l \mid \boldsymbol{X}_i^m) - \Pr(c_l)] \tag{7}$$

比较基于独立假设的融合模型式(4)和基于线性依赖的融合模型式(6),可见两者之间仅相差 D_i^l 项.当所有分类器相互独立时,即 $\alpha_l^m = 0$,则 $D_i^l = 0$,依赖模型退化为独立模型.因此可用式(6)统一表示多分类器融合模型.

3.4 优化求解线性依赖模型

训练样本集中共有 N 个样本、L 个类别和 M 个特征空间.设训练样本视频片段 V_i 中包含 n_i 个镜头,V_i 在第 m 个特征空间中的特征向量及其相应的标签表示为 $\boldsymbol{X}_i^m = (x_{i,1}^m,\cdots,x_{i,j}^m,\cdots,x_{i,n_i}^m)$ 和 y_i.

若所有训练样本的分类标签均正确,则真正的分类结果 $\Pr(y_i \mid \boldsymbol{X}_i)$ 和错误预测结果 $\max_{c_l \neq y_j}\Pr(c_l \mid \boldsymbol{X}_i)$ 之间的差值应该最大.

可采用 LPBoost 方法求解依赖权值 $\alpha_l^m(l = 1,\cdots,L, m = 1,\cdots,M)$.文献[19]指出:在同一特征空间中赋予所有类别相同权值时,即 $\alpha_l^m = \alpha^m$,分类性能最好,故需求解的依赖权值向量可简化为 $\boldsymbol{\alpha} = \{\alpha^1,\cdots,\alpha^m,\cdots,\alpha^M\}$.因此,我们采用 LP-$\beta$ 算法[19]求解 $\boldsymbol{\alpha}$,其目标函数为:

$$\begin{aligned}&\min_{\alpha,\rho,\xi} -\rho + \frac{1}{vN}\sum_{i=1}^{N}\xi_i\\ &\text{s.t.}\\ &\text{(i)}\sum_{m=1}^{M}\alpha^m\Pr(y_i \mid X_i^m)\\ &\quad - \arg\max_{c_l \neq y_i}\sum_{m=1}^{M}\alpha^m\Pr(c_l \mid X_i^m) \geq \rho - \xi_i, \forall\, i, c_l \neq y_i\\ &\text{(ii)}\Pr(c_l \mid X_i^m) \geq 0, \forall\, i, c_l\\ &\text{(iii)}\xi_i \geq 0, \forall\, i, c_l \neq y_i\\ &\text{(iv)}\sum_{m=1}^{M}\alpha^m = 1, 0 \leqslant \alpha^m \leqslant 1\end{aligned} \tag{8}$$

其中 ρ 是分类间隔,ξ_i 是非负的松弛变量,$v \in (0,1)$ 是常数参数.

由于在特定敏感视频识别应用中类别数 $L = 2$,当 $c_l \neq y_i$ 时,则有 $c_l = -y_i$,我们对式(8)中第一个约束条件做如下变换:

$$\begin{aligned}&\sum_{m=1}^{M}\alpha^m\Pr(y_i \mid \boldsymbol{X}_i^m)\\ &\quad - \arg\max_{c_l \neq y_i}\sum_{m=1}^{M}\alpha^m\Pr(c_l \mid \boldsymbol{X}_i^m), \forall\, i, c_l \neq y_i\\ &= \sum_{m=1}^{M}\alpha^m\Pr(y_i \mid \boldsymbol{X}_i^m) - \sum_{m=1}^{M}\alpha^m\Pr(c_l \mid \boldsymbol{X}_i^m)\\ &= \sum_{m=1}^{M}\alpha^m[\Pr(y_i \mid \boldsymbol{X}_i^m) - \Pr(c_l \mid \boldsymbol{X}_i^m)]\end{aligned}$$

则式(8)可改写为:

$$\min_{\alpha,\rho,\xi} -\rho+\frac{1}{vN}\sum_{i=1}^{N}\xi_i$$

s.t.

$$\text{(i)}\sum_{m=1}^{M}\alpha^m[\Pr(y_i\mid \boldsymbol{X}_i^m)-\Pr(c_l\mid \boldsymbol{X}_i^m)]\geq\rho-\xi_i,\forall i,c_l\neq y_i$$

$$\text{(ii)}\Pr(c_l\mid \boldsymbol{X}_i^m)\geq 0,\forall i,c_l$$

$$\text{(iii)}\xi_i\geq 0,\forall i,c_l\neq y_i$$

$$\text{(iv)}\sum_{m=1}^{M}\alpha^m=1,0\leqslant\alpha^m\leqslant 1 \tag{9}$$

求解上述线性规划问题,便可得到依赖权值向量 $\boldsymbol{\alpha}=\{\alpha^1,\cdots,\alpha^m,\cdots,\alpha^M\}$.

3.5 分类

提取测试视频 V_t 的 M 种不同特征及其上下文结构$\{(X_t^1,G_t^1),\cdots,(X_t^m,G_t^m),\cdots,(X_t^M,G_t^M)\}$,分别输入到不同特征空间所对应的分类器中,得到 $M*L$ 种结果,记为 $\Pr(c_l\mid X_t^m)(l=1,\cdots,L;m=1,\cdots,M)$,表示视频 V_t 在第 m 个空间特征被判定为类别 c_l 的概率,则 V_t 的最终类别 y_t 为:

$$y_t=\arg\max_{c_l}\sum_{l=1}^{L}\Pr(c_l\mid X_t^1,\cdots,X_t^M);l=1,\cdots,L$$

在用式(6)计算 $\Pr(c_l\mid X_t^1,\cdots,X_t^M)$时,发现很难精确计算出 P_0中的联合概率 $\Pr(X_t^1,\cdots,X_t^M)$,而且当分类器数很多时,则需要大量数据才能精准计算前述联合分布[18].事实上,对于两个不同类别 c_l 和 c_k,只需比较它们后验概率大小即可.令

$$diff=\Pr(c_l\mid X_t^1,\cdots,X_t^M)-\Pr(c_k\mid X_t^1,\cdots,X_t^M) \tag{10}$$

若 $diff\geq 0$,则 V_t 属于 c_l 类,否则 V_t 属于 c_k 类.将式(6)带入式(10),则有:

$$diff=P_0\left(\begin{array}{l}(1-2M)(\Pr(c_l)-\Pr(c_k))\\+\sum_{m=1}^{M}(1+\alpha^m)(\Pr(c_l\mid X_t^m)-\Pr(c_k\mid X_t^m))\end{array}\right)$$

令

$$\Delta=(1-2M)(\Pr(c_l)-\Pr(c_k))+\sum_{m=1}^{M}(1+\alpha^m)(\Pr(c_l\mid X_t^m)-\Pr(c_k\mid X_t^m)) \tag{11}$$

由于 $P_0>0$,则视频 V_t的标签 y_t 取值:

$$y_t=\begin{cases}c_l, & \Delta\geq 0\\ c_k, & \Delta<0\end{cases} \tag{12}$$

若先验概率相同,即 $\Pr(c_l)=1/L(l=1,2,\cdots,L)$,则 Δ 可简化为:

$$\Delta=\sum_{m=1}^{M}(1+\alpha^m)(\Pr(c_l\mid X_t^m)-\Pr(c_k\mid X_t^m)) \tag{13}$$

若采用基于独立假设的分类融合方法,则 Δ 又可简化为:

$$\Delta_{in}=\sum_{m=1}^{M}(\Pr(c_l\mid X_t^m)-\Pr(c_k\mid X_t^m)) \tag{14}$$

4 特定敏感视频识别实验

恐怖视频和暴力视频有一定差别,恐怖视频主要与情感相关、镜头画面的色调较暗、往往伴随令人毛骨悚然的背景音乐,而暴力视频的镜头画面较血腥、运动节奏较快、常常出现打斗时发出的嘶喊声,正是基于这种区别,我们在描述恐怖视频时,采用了视觉、颜色情感强度、颜色和谐度及音频特征;在描述暴力视频时,除了前述 4 种特征外,还采用了运动模板特征.

本节将我们提出的 MCS-LDF 方法分别用于识别恐怖视频和暴力视频,通过在两个视频库上的对比实验验证本文算法的有效性.

4.1 恐怖视频识别实验

4.1.1 数据集及评价标准

本文采用与文献[5,7]相同的恐怖视频数据集,图 2 给出部分视频的海报示例[7].

(*a*) 恐怖视频海报

(*b*) 非恐怖视频海报

图2 恐怖数据集中视频海报示例[7]

首先对视频进行结构化分析,提取镜头的中间帧

作为关键帧;然后采用与文献[5]相同的方法提取视觉、颜色情感(Color Emotion-CE)及音频三种特征.需要说明的是:文献[5]中的颜色情感特征(CE)是将颜色情感强度(Color Emotion Intensity-CEI)和颜色和谐度(Color Harmony-CH)简单拼接而成.通过分析,发现颜色情感强度和颜色和谐度相对独立,两者在分类中起的作用应有所不同,我们预测如此简单拼接反而会降低分类性能.因此,本文将颜色情感特征拆分为颜色情感强度和颜色和谐度两个独立的特征,即提取视觉、颜色情感强度、颜色和谐度及音频四种特征来表示恐怖视频.

采用视频分类中广泛使用的查准率 P (Precision),查全率 R (Recall)及 F_1 指标(F-measure)来评价各种算法的性能.假设数据集中恐怖视频集合为 HS,算法识别得到的恐怖视频集合为 ES,则 P,R 和 F_1 的计算如下:

$$P=\frac{|HS\bigcap ES|}{|ES|},R=\frac{|HS\bigcap ES|}{|HS|},F_1=\frac{2\times P\times R}{P+R} \tag{15}$$

4.1.2 实验结果及分析

为了验证算法的性能,我们设计 11 个实验进行对比,如表 1 所示.

在构造上下文结构图 ε-Graph 时,参数 ε 的取值范围为{0.05,0.1,…,0.95},式(1)中 RBF 核的参数 γ 取值范围为{0.5,0.1,…,9.5}.式(9)中参数 v 的取值范围为{0.05,0.1,…,0.95}.通过在训练集上进行三重交叉验证,获得每次实验参数的最优值.对上述每种方法都采用 10 次 10 重交叉验证,各种算法的实验结果如表 2 所示,其中 ± 号后的数字表示十次重复实验的标准差.

表 1 11 个实验说明

实验编号	实验名称	特征	是否考虑上下文结构	分类器	分类器融合方法
1	CE[5]	CEI + CH 简单拼接	否	多示例学习方法	无
2	VF	VF	是	SVM	无
3	CEI	CEI	是	SVM	无
4	CH	CH	是	SVM	无
5	AF	AF	是	SVM	无
6	NCS-IF	VF,EI,CH,AF	否	SVM	基于独立假设的多分类器融合
7	MCS-IF	VF,EI,CH,AF	是	SVM	基于独立假设的多分类器融合
8	NCS-LDF	VF,EI,CH,AF	否	SVM	基于线性依赖的多分类器融合
9	本文的 MCS-LDF	VF,EI,CH,AF	是	SVM	基于线性依赖的多分类器融合
10	MI-SVM[5]	VF,EI,CH,AF 简单拼接	否	多示例学习方法	无
11	Multi-view-sparse[7]	VF,EI,CH,AF 简单拼接	融合三种视角的镜头间上下文结构	基于稀疏表示的分类方法	无

表 2 恐怖视频库上的实验结果(%)

实验编号	实验名称	Precision(P)	Recall(R)	F-measure(F_1)
1	CE = CEI + CH[5]	69.42	68.11	68.76
2	VF	71.90 ± 0.66	73.67 ± 0.69	72.77 ± 0.54
3	CEI	69.75 ± 1.27	70.65 ± 0.71	69.69 ± 0.95
4	CH	79.50 ± 0.47	77.64 ± 0.36	78.56 ± 0.33
5	AF	82.20 ± 0.63	84.01 ± 0.42	83.09 ± 0.43
6	NCS-IF	80.60 ± 1.02	77.35 ± 0.63	78.94 ± 0.56
7	MCS-IF	82.25 ± 0.48	79.28 ± 0.36	80.74 ± 0.40
8	NCS-LDF	86.15 ± 0.88	86.29 ± 0.56	86.22 ± 0.46
9	MCS-LDF [本文]	88.15 ± 0.47	86.49 ± 0.32	87.31 ± 0.17
10	MI-SVM[5]	79.78	78.92	79.35
11	Multi-view-sparse[7]	84.8 ± 0.49	84.31 ± 0.38	84.55 ± 0.33

由于本文所用的恐怖视频数据集与文献[5,7]中的均相同,所以实验 1 和实验 10 的数据直接摘自文献[5],实验 11 的数据直接摘自文献[7].从表 2 中可以看出:

(1)四种不同特征(VF,CEI,CH,AF)的分类性能差异很大,说明它们在分类中所起的作用确实不同,其中 CEI 的分类效果和稳定性均最差.

(2)CE 是由 CEI 和 CH 简单拼接而成,其性能较之于最差的 CEI 仍略逊一筹,更是远不如 CH,降低了 9%,这恰恰验证了我们预测的正确性,即特征的简单拼接反而会降低分类的性能.

(3)NCS-IF 的结果好于 VF,CEI 和 CH,但不如 AF,说明不考虑镜头间上下文结构、基于独立假设的分类融合有时甚至不如某单个强势特征的效果好.

(4)MCS-IF 优于 NCS-IF,以及 MCS-LDF 优于 NCS-LDF,均说明视频镜头间上下文结构信息有助于提高分类性能,识别率可提高 1% ~ 2%.

(5)NCS-LDF、MCS-LDF 均远远优于 NCS-IF、MCS-IF,说明基于线性依赖的分类融合可大幅度提高分类性能,识别率的改善可高达 7% 左右.

(6)与文献[5,7]的数据相比,本文的 MCS-LDF 方法的识别性能远高于 MI-SVM 和 Multi-view-sparse,且 MCS-LDF 方法的稳定性比文献[7]也略有提高.

(7)在 11 种方法中,本文提出的 MCS-LDF 的准确率、查全率和 F_1 指标全部高于其它方法,证明本方法的有效性;而且较小的标准差表明 MCS-LDF 方法具有更好的稳定性.

(8)本文中,考虑上下文结构和线性融合分类是提高识别性能的两方面策略,MCS-IF 较之于 NCS-IF、MCS-LDF 较之于 NCS-LDF,识别率可提高 1% ~ 2%.而 NCS-LDF 较之于 NCS-IF、MCS-LDF 较之于 MCS-IF,识别率提高达 7% 左右,说明线性融合策略比上下文结构更有效.

4.2　暴力视频识别实验

由于目前还没有专门用于暴力视频识别测试的数据集,我们从互联网上下载了分别由中国、美国,韩国及泰国制作的 100 部暴力电影和 100 部非暴力电影(包括喜剧、动作、感情剧和动画片),图 3 给出其中部分电影的海报示例.我们从这些电影中截取了 400 个暴力场景和 400 个非暴力场景.这些场景被分成 A 和 B 两个集合,A 和 B 分别包含 200 个暴力场景和 200 个非暴力场景,为了消除相关性,来自同一部电影的场景被分到同一个集合中.在实验中,A 作为训练集则 B 作为测试集,反之,B 作为训练集则 A 作为测试集.

(*a*) 暴力视频海报

(*b*) 非暴力视频海报

图3　暴力数据集中视频海报示例

暴力视频与恐怖视频相比,主要区别之一是目标的运动速度快.我们预测:快速运动信息可更有效地表达暴力的语义信息,提高暴力视频识别率.因此,采用视觉(VF)、颜色情感强度(CEI)、颜色和谐度(CH)、音频(AF)及运动模板特征(MTF)五种特征来表示暴力视频.

仍采用文献[5]中的方法提取前四个特征,参考文献[20,21]提取运动模板特征.

为了验证算法的性能,本文设计了 9 个实验,分别为 VF、CEI、CH、AF、MTF、NCS-IF、MCS-IF、NCS-LDF 与本文所提出的 MCS-LDF.其中 MTF 方法是:仅提取运动模板特征及镜头间上下文结构,采用 SVM 分类器识别暴力视频,其余 8 个方法与恐怖视频识别方法相同.评价标准、参数选择方法亦同于 4.1 节.各种算法的实验结果如表 3 所示.

表 3　暴力视频库上的实验结果(%)

实验编号	实验名称	Precision(P)	Recall(R)	F-measure(F_1)
1	VF	63.05 ± 0.34	67.03 ± 0.37	64.96 ± 0.32
2	CEI	61.50 ± 1.13	67.04 ± 0.45	63.75 ± 0.32
3	CH	65.05 ± 0.80	68.81 ± 0.81	66.87 ± 0.40
4	AF	73.00 ± 0.63	74.87 ± 0.65	73.92 ± 0.83
5	MTF	81.50 ± 0.83	76.52 ± 0.26	78.93 ± 0.50
6	NCS-IF	72.00 ± 0.86	80.90 ± 0.68	76.19 ± 0.71
7	MCS-IF	78.00 ± 0.39	82.98 ± 0.41	80.41 ± 0.56
8	NCS-LDF	83.00 ± 0.53	85.57 ± 0.68	84.26 ± 0.65
9	MCS-LDF [本文]	85.50 ± 0.31	86.80 ± 0.32	86.15 ± 0.31

从表 3 可以看出:

(1)五种不同特征中 MTF 的性能最好,这验证了我们预测的正确性,说明运动信息确实对暴力视频的表征能力最强.

(2)NCS-IF 的结果好于 VF,CEI,CH 和 AF,但略逊于 MTF,说明不考虑镜头间上下文结构、基于独立假设的分类融合有时甚至不如某个单独的强势特征效果好.

(3)MCS-IF 优于 NCS-IF,以及 MCS-LDF 优于 NCS-LDF,均说明镜头间上下文结构有助于提高分类性能,

识别率可提高 1% ~4%.

(4)NCS-LDF、MCS-LDF 均远远优于 NCS-IF、MCS-IF,说明基于线性依赖的分类融合可大幅度提高分类性能,识别率可提高约 6% ~8%.

(5)在 9 种方法中,MCS-LDF 的性能远远高于其它方法,证明本文方法的有效性和较强的稳定性.

(6)本文中,考虑上下文结构和线性融合分类是提高识别性能的两方面策略,MCS-IF 较之于 NCS-IF、MCS-LDF 较之于 NCS-LDF,识别率可提高 1% ~4%.而 NCS-LDF 较之于 NCS-IF、MCS-LDF 较之于 MCS-IF,识别率提高达 6% ~8%左右,在暴力视频识别实验中再次证明线性融合策略比上下文结构更有效.

5 结论

现有的大部分特定敏感视频识别方法或是假定镜头间相互独立,忽略了视频内部的结构信息,或是对多种不同特征做简单拼接,忽略了各种特征之间潜在的依赖关系及其在分类中作用的差别.本文提出了一种基于多种上下文结构与线性融合的特定敏感视频识别方法,该方法不但充分考虑了多个特征空间中镜头间的上下文结构,而且充分考虑了不同特征间的依赖关系及各种特征在分类中的不同作用.本方法有效地实现了特定敏感视频的识别,兼顾了对敏感视频多样性的适应能力和识别效率,而且具有更好的稳定性,相比于传统策略,识别性能得到了明显改善.

参考文献

[1] Z Rasheed, Y Sheikh, M Shah. On the use of computable features for film classification[J]. IEEE Transactions CSVT, 2005, 15(1):52 – 64.

[2] H B Kang. Affective content detection using HMMs[A]. Proceedings of the Eleventh ACM International Conference on Multimedia[C]. New York: ACM, 2003. 259 – 262.

[3] M Xu, L T Chia, J Jin. Affective content analysis in comedy and horror videos by audio emotional event detection[A]. Proceedings of IEEE International Conference on Multimedia and Expo[C]. Torino: IEEE, 2005. 622 – 625.

[4] J C Wang, B Li, W M Hu, et al. Horror movie scene recognition based on emotional perception[A]. Proceedings of IEEE International Conference on Image Processing[C]. Paris: IEEE, 2010. 1489 – 1492.

[5] J C Wang, B Li, W M Hu, et al. Horror video scene recognition via multiple-instance learning[A]. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing[C]. Prague: IEEE, 2011. 1325 – 1328.

[6] X M Ding, B Li, W M Hu, et al. Context-aware horror video scene recognition via cost-sensitive sparse coding[A]. Proceedings of IEEE International Conference on Pattern Recognition[C]. Tsukuba: IEEE, 2012. 1904 – 1907.

[7] 丁昕苗,李兵,胡卫明,等.基于多视角融合稀疏表示的恐怖视频识别[J].电子学报,2014,42(2):301 – 305.
X M Ding, B Li, W M Hu, et al. Horror video scene recognition based on multi-view joint sparse coding[J]. Acta Electronica Sinica, 2014, 42(2):301 – 305. (in Chinese)

[8] Datta A, Shah M, Lobo NDV. Person-on-person violence detection in video data[A]. Proceedings of IEEE International Conference on Pattern Recognition[C]. Quebec: IEEE, 2002. 433 – 438.

[9] T Giannakopoulos, D Kosmopoulos, A Aristidou, et al. Violence content classification using audio features[A]. Proceedings of the 4th Helenic Conference on AI[C]. Heraklion: Springer Berlin Heidelberg, 2006. 502 – 507.

[10] J Nam, M Alghoniemy, A H Tewfik. Audio-visual content-based violent scene characterization[A]. Proceedings of IEEE International Conference on Image Processing[C]. Chicago: IEEE, 1998. 353 – 357.

[11] J Lin, W Wang. Weakly-supervised violence detection in movies with audio and video based co-training[A]. Proceedings of the 10th Pacific Rim Conference on Multimedia[C]. Bangkok: Springer Berlin Heidelberg, 2009. 930 – 935.

[12] T Giannakopoulos, A Pikrakis, S Theodoridis. A multimodal approach to violence detection in video sharing sites[A]. Proceedings of IEEE International Conference on Pattern Recognition[C]. Istanbul: IEEE, 2010. 3244 – 3247.

[13] K Wang, Z Zhang, L Wang. Violence video detection by discriminative slow feature analysis[A]. Proceedings of Chinese Conference on Pattern Recognition[C]. Beijing: Springer Berlin Heidelberg, 2012. 137 – 144.

[14] L Wen, Z Cai, Z Lei, et al. Robust online learned spatio-temporal context model for visual tracking[J]. IEEE Transactions IP, 2014, 23(2):785 – 796.

[15] Z Zhou, Y Sun, Y Li. Multi-instance learning by treating instances as non-i.i.d. samples[A]. Proceedings of the 26th Annual International Conference on Machine Learning[C]. New York: ACM, 2009. 1249 – 1256.

[16] J Kittler, M Hatef, R P W Duin, et al. On combining classifiers[J]. IEEE Transactions PAMI, 1998, 20(3):226 – 239.

[17] A Jain, K Nandakumar, A Ross. Score normalization in multimodal biometric systems[J]. Pattern Recognition, 2005, 38(12):2270 – 2285.

[18] A J Ma, P C Yuen, J H Lai. Linear dependency modeling for classifier fusion and feature combination[J]. IEEE Transactions PAMI, 2013, 35(5):1135 – 1148.

[19] P Gehler, S Nowozin. On feature combination for multiclass object classification[A]. Proceedings of IEEE 12th International Conference on Computer Vision[C]. Kyoto: IEEE,

2009.221 – 228.

[20] U Mahbuba, H Imtiaza, T Roya, et al. A template matching approach of one-shot-learning gesture recognition[J]. Pattern Recognition Letters, 2013, 34(15): 1780 – 1788.

[21] M A R Ahad, J Tan, H Kim, et al. Action recognition by employing combined directional motion history and energy images[A]. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops[C]. San Francisco: IEEE Computer Society, 2010.73 – 78.

作者简介

王方石　女, 1969 年 2 月出生, 吉林长春人. 1990 年、1993 年分别在吉林大学获理学学士、硕士学位, 2007 年在北京交通大学获工学博士学位, 现为北京交通大学软件学院教授, 主要研究方向为网络多媒体信息理解与识别.
E-mail: fshwang@bjtu.edu.cn

王辰龙　男, 1988 年 12 月 11 日出生, 吉林长春人. 2011 年在北京交通大学获工程学士学位, 现为北京交通大学软件学院硕士研究生, 研究方向为网络多媒体信息理解与识别.
E-mail: morndragon@126.com