高技术通讯 2009 年 第 19 卷 第 10 期:1041 ~ 1047

doi:10.3772/j.issn.1002-0470.2009.10.009

基于视觉单词和语义映射的色情图像检测算法①

王宇石②* 高 文* **

(* 哈尔滨工业大学计算机科学与技术学院 哈尔滨 150001)

(** 北京大学信息科学技术学院 北京 100871)

摘 要 针对传统类型的色情图像检测方法误检率高的问题,提出了一种基于多层视觉单词的检测方法。该方法首先对色情场景的各种视觉元素建立视觉单词,然后通过这些视觉单词建立更高层的编码,包括视觉词组和兴趣区域类别,从而实现对图像不同形态级别的描述与分析。图像的识别特征由相应的编码直方图组成,并将特征映射到一个低维空间中,使图像间的语义距离与空间距离相协调。该方法在各种图像测试中都表现出出色的性能,例如在人物类图像测试中,误检率比传统方法降低了 40%。实验结果证明,多层单词体系能够更高效地分析色情图像等复杂场景。

关键词 图像识别, 色情图像, 视觉单词, 视觉词组, 多层描述, 降维

0 引 言

自动分析图像内容的方法能够帮助我们主动、高效地在互联网中扫描色情图像、甄别成人网站。国内外学者对此进行了大量的研究。基本的策略是先对图像进行肤色检测,然后提取适当的特征进行识别。传统方法所提取的特征主要有两类:全局的低层特征和肤色区域特征。前者主要是指颜色、纹理等方面的特征。例如,Jeong 等人[1]从 MPEG-7 的可伸缩颜色描述子和颜色结构描述子发展出自己的颜色特征,并依据颜色对比和肤色概率获得图像的关键区域,从中提取特征。Shih 等人[2]除了提取颜色直方图外,还提取了图像不同分区的边缘直方图。在肤色区域特征方面,研究者使用了各种描述区域的特征,如矩和致密度,以及面积、密度、位置等[3-5]。Hu 等人首先计算人体的轮廓,从中提取了关于非肤色的区域及像素的分布特征,然后使用最近邻分类器进行检测[6,7]。进而,上述两方面的特征被加以结合,例如曾炜等人开发的系统中融合了颜色、形状、纹理多方面的知识[8]。Arentz 等人也结合了颜色、纹理特征对分割的区域进行描述,然后利用遗传算法获取典型的区域描述[9]。

传统方法在检测到多数的色情图像的同时会产生大量的误检,主要的原因是所用的颜色、纹理等类型的全局特征区分能力有限。此外,传统方法严重依赖肤色检测的结果,除了人物类图片会引起混淆外,包含大量类似肤色内容(如沙漠、晚霞、某些动物等)的图片也会被误检。事实上,色情图像与其它图像关键性的区别在于各种与人体有关的局部形态(如敏感器官)。这些局部形态与肤色分布信息结合在一起,构成了人们对色情图像的认知。为此,本文从局部特征着手,从全新的角度捕获色情图像更高层次的语义信息。这里所说的局部特征,是对局部形态(视觉元素)的描述特征,并通过聚类,量子化为视觉单词,每个视觉单词对应一定的局部形态。依据视觉单词(下文中简称为"单词")的出现规律,可以分析图像的语义[10,11]。

本文的贡献主要有以下三个方面。第一,在图像识别中引入了一种描述单词局部共现关系的"词组"模型。Zheng 等人提出的词组模型仅仅利用了单词对的相邻共现关系[11];文献[12]则考察了单词对之间的距离关系;文献[13]的方法将单词对之间的距离、角度也予以描述,具有较高的描述复杂度。与之相比,本文的词组同时考察所有单词的局部共现关系,不但具有更强的描述能力,且适于处理局部形变,此外描述的复杂度也较低。第二,基于前面所提取的单词、词组,建立了多层的图像描述,所涉及的

① 863 计划(2003AA142140)和国家自然科学基金(60702035)资助项目。

② 男,1978 年生,博士;研究方向:图像的分析与理解;联系人,E-mail: yswang@jdl.ac.cn

(收稿日期:2008-09-25)

视觉元素包括局部形态、纹理、皮肤。其中，通过分割得到了相关信息密集分布的“兴趣区域”(region of interest, ROI)，然后在各 ROI 中进一步分析单词共现规律。而传统的 ROI 提取策略仍局限于分析肤色区域的形状或颜色相关性，例如文献[1,2]。第三，基于图像的语义，对高维的图像特征向量进行了有效的降维处理，不但提高了识别效率，而且使图像的语义距离在新的特征空间中得到更充分的体现，特征空间的局部结构得到更充分的净化，从而使识别性能得到进一步提高。最终，本文的方法从多个层次和角度，全面地分析了色情图像中单词的分布规律，从而显著减低了系统对肤色检测的依赖。相比于传统方法，系统在带有类似肤色的图像中误判大为减少。

1 基于视觉单词的色情图像检测

图 1 展示了检测算法的框架。皮肤信息仍然有其特定的作用，本文选择了文献[6]提出的皮肤检测算法，其利用期望最大化算法求得肤色概率的高斯混合模型。系统将把肤色比例低于 5% 的图像直接判为非色情的。在提取了图像三个级别(普通、词组、ROI 话题)的视觉单词之后，通过分析与色情场景高度相关的单词分布规律，提取了一组识别特征，称为敏感度特征。用这些单词的直方图和敏感度特征共同组成图像的特征向量。进而该特征向量被降维映射到更符合语义区分性的空间中，并使用支持向量机(support vector machine, SVM)作为图像的分类器。

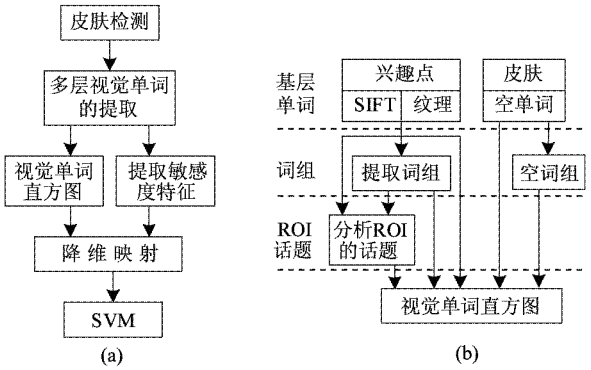

图 1 色情图像检测算法的框架(a)及对(a)中“多层视觉单词的提取”部分的描述(b)

1.1 提取普通视觉单词

色情图像的视觉元素(视觉单词)主要包括两大类型：皮肤，以及人体中局部发生突出变化的位置，即兴趣点，例如五官、敏感器官、毛发、手。本文选用了适于检测块状物的高斯差分(difference of Gaussian, DoG)兴趣点检测子[14]。DoG 在尺度空间中搜寻图像的局部突出点。

该检测子需要计算各局部位置特定尺度的高斯均值，具有较高的计算代价，为此进行了相应的简化[15]。然后在兴趣点邻域内建立局部描述。人体主要包含两类局部形态：结构上有一定规律的和纹理类型的(如毛发)。为此，对应实现了两种局部形态描述子：(1)尺度不变特征变换(scale invariant feature transform, SIFT)算法[14]，这是一种 128 维的局部梯度的方向-分布描述子；(2)一种 65 维纹理描述子，描述了图像的局部二值模式(local binary patterns[16])的分布以及梯度的强度分布和方向分布。

随后，对从训练集中提取的 SIFT 及纹理描述向量集合，分别进行 K-均值聚类。产生的每个聚类作为 1 个视觉单词，所有获取的聚类组成单词表。为了提高在单词表中查词的速度，采用了双层聚类。在第一层聚类中设 K 为 10，然后在得到的每个聚类 C_i($i = 1 \sim 10$) 中，进行第二层聚类，得到真正的单词。其中在各 C_i 中，聚类的数目相等。用树结构来组织单词表，在显著提高查表速度的同时(约为原来的 10 倍)，可以尽量减少对单词表准确性的损害。

除了上述兴趣点类型的局部视觉元素外，对于皮肤块也建立了多尺度描述体系。首先将图像分成 8×8 和 12×12 的网格 MAT_8 和 MAT_{12}，如果一格所对应的区域中肤色像素比例大于 0.5，并且其中兴趣点的个数小于 T_e，则称此格为 1 个“空单词”(共两种空单词)。换句话说，空单词对应一个基本平滑的肤色区域。这里的 T_e 得自对实际图像的观察，本文设为 4.5(在 MAT_8 中)和 2.5(在 MAT_{12}中)。MAT_8 和 MAT_{12}两种网格分别在不同的尺度上考察了图像的肤色分布，前者涵盖了身体的主要躯干，而后者则描述了远景的情况或局部较为细致的身体部位。

1.2 构造高层视觉单词

普通的视觉单词仅仅利用了图像中一些突出点的局部形态信息，有明显的局限性。(1)对于较为复杂的局部形态(如敏感部位)缺乏描述能力，例如 SIFT 将局部区域分成 4×4 的分区，如果对一个形态复杂的区域使用单个 SIFT 向量来描述，其每一分区覆盖的面积过大，描述能力显得不足。(2)仅利用了图像各处的局部信息，而图像的视觉内容是按多个层次表达的；每次观察一个兴趣点和每次观察图像

的某一区域,分别会给观察者带来不同的信息。为此,构建了两层更高层级的视觉单词来提高对色情图像的识别能力。第一层是“词组”(phrase)[11],表达典型的相邻单词的组合关系。第二层是层级更高的“ROI 类别”,即在较大的区域中描述视觉单词的共现关系。在建立较高层级单词的描述时,均利用了较低级别的单词。最终,单词、词组、ROI 类别联合构成了一个多层的图像描述体系。三级单词分别以不同的粒度对图像进行了描述,低层单词侧重于刻画局部形态,较高级别的“单词”则偏重于描述局部形态的上下文关系。

利用高层单词来考察普通单词的局部共现关系,除了有助于描述较为复杂的局部形态[17]外,还能降低单词的歧义性[18]。例如,不同的事物可能会具有相似的结构(人眼和某些树叶)[10],如果被归类为同一单词,则分析其邻域背景内其它单词的分布(是五官,还是大量其他树叶),有助于提高识别的精确性。此外,这些高层“单词”并不刻画局部形态精确的几何关系,这反而有利于系统灵活地面对检测目标所发生的形变、遮挡等各种情况。

1.2.1 词组描述

首先介绍一组紧凑的视觉词组描述算法。该算法将以每一种单词为中心,构造一组典型词组。其中一个词组主要由如下的规则所定义:(1)以哪个单词为中心;(2)有哪些单词出现在中心单词的邻域中,而且出现的频率为多少。这些规则将通过决策树路径的形式加以组织,下面给出详细的解释。以单词 w 为例,对于图像中 w 的某个实例 P,在 P 周围 1.5 倍于 P 的尺度的区域(N_P)内,提取所有与 P 尺度相近的单词实例,产生一个局部单词直方图(包括 SIFT 单词和纹理单词)。由于色情图像中单词分布并不均匀,为考察邻域内单词分布的密度定义了一种“空邻居”。在 N_P 的每个 45°等分扇形内,若没有与 P 尺度相近的单词实例出现,则该扇形被称为一个空邻居。空邻居作为一种独特的“单词”也统计入上述单词直方图。

设 M_w 是以 w 为中心的局部单词直方图训练集,通过对 M_w 进行聚类,获取以 w 为中心的典型词组。聚类时采用自底向上的合并聚类,每次将距离最近的节点合并为一个节点。指定一定的剪枝参数,即可得到一株聚类树,其每个叶子节点对应一个聚类(即一个典型的词组),所有的叶子组成一个词组子集 P_w。只要指定了总体的剪枝参数,就不需要预先规定在 M_w 中聚类时应该产生多少个聚类。最终,总的词组表是 $\bigcup_w P_w$。在聚类树的每个节点处,为了降低查表的复杂度,只选取两个子节点中取值分布差别最大的前 10 维,作为查表依据。仅以此 10 个有区分性的单词来计算距离,向距离最近子节点前进。此外,为了降低词组构造的复杂度,在构造词组时各兴趣点的单词标号实际上来自于一个小型单词表,即构造单词时选取了较粗的粒度。

基于 1.1 节所提出的空单词,进一步构造了对应的“空”词组,用于描述 MAT 中空单词的旋转不变的相邻关系。一个空词组的编号 $LBP^{\mathrm{rot}}_{\mathrm{empty}}(e_c)$ 按式(2)得到,总共有 36 种编号[16]。其中 e_c 表示当前某个空单词,e_j 表示 MAT 中与 e_c 相邻的格。当 e 是空单词时 $Empty(e)$ 值为 1,否则为 0。$ROT(x,l)$ 表示对一个二进制数 x 进行循环右移 l 位的操作。

$$LBP_{\mathrm{empty}}(e_c) = \sum_{j=1}^{8} Empty(e_j)\cdot 2^{j-1} \tag{1}$$

$$LBP^{\mathrm{rot}}_{\mathrm{empty}}(e_c) = \min\{ROT(LBP_{\mathrm{empty}}(e_c),l) \mid l = 0 \sim 7\} \tag{2}$$

1.2.2 兴趣区域描述

兴趣区域(ROI)是人体信息较为集中的局部区域。在描述中,要将这些 ROI 按内容(其包含的低层单词实例)分成若干类别,每个类别就是一种 ROI 级别的高层视觉单词。

首先要获取 ROI,为此定义单词(及词组)的敏感度 CD(correlation degree),表示一个单词(词组)与色情图像的相关性。设 $F(w \mid porn)$ 表示出现单词 w 的色情图像的比例,$F(w \mid nonporn)$ 是出现 w 的非色情图像的比例,则有

$$CD(w) = F(w \mid porn)\cdot\left(\frac{F(w \mid porn)}{F(w \mid nonporn)}\right)^2 \tag{3}$$

ROI 具体定义为:高敏感度视觉单词相对集中的局部区域。产生一个尺寸为原始图像 1/8 的缩图,称为敏感度分布图(correlation degree map,CDMap),其中每个位置 p 的值为

$$CDM(p) = \sum_{v\in p}[CD(W(v)) + CD(phrase(v))] \tag{4}$$

其中,$W(v)$ 表示单词实例 v 的单词编号,$phrase(v)$ 表示 v 对应的词组编号。然后在 CDMap 中可以快速地通过 K-均值聚类实现区域分割。在二维的 CDMap 中,每个位置对应的数据点的个数就是 $CDM(p)$。聚类数由 CDMap 中的局部最大值的个数决定。最终,以每个聚类的质心为中心,以聚类半

径的 2 倍为边长,在图像中划分出一个矩形 ROI。

进而可以使用图像级的分析手段来处理各个 ROI。本文选择潜在语义概率分析(probabilistic latent semantic analysis,PLSA)模型[19]来分析 ROI 的类别。PLSA 将文本(或 ROI)视为由单词(或视觉单词)组成的集合。通过分析单词共现关系,提取文本(或 ROI)集合中潜在的"话题"(topic),最终一个文本(或 ROI)将归类为某个话题(即类别)。每个被提取出来的话题,就是一种 ROI 级的视觉单词。通过 PLSA 模型,可以估计单词 w_j 与话题 z_k 的概率关系 $P(w_j \mid z_k)$。

对于 ROI d,由于 $P(z_k \mid d) = \dfrac{P(d \mid z_k)P(z_k)}{P(d)}$,假设各种话题的先验概率是相同的,则有 $P(z_k \mid d) \sim P(d \mid z_k)$。为确定 d 归属的话题,计算

$$\begin{aligned} P(d \mid z_k) &= \prod_{w_j} P(w_j, n(d,w_j) \mid z_k) \\ &= \prod_{w_j} P(n(d,w_j) \mid w_j, z_k) P(w_j \mid z_k) \end{aligned} \tag{5}$$

其中 $n(d,w_j)$ 表示 w_j 在 d 中出现的次数。$P(n \mid w_j, z_k)$ 用 Parzen 方法[20]予以估计:

$$P(n \mid w_j, z_k) = \frac{1}{m_{jk} h} \sum_{l=1\sim m_{jk}} K\left(\frac{n - n_l}{h}\right),$$

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2) \tag{6}$$

其中 m_{jk} 表示包含单词 w_j、属于话题 z_k 的训练 ROI 的数量,n_l 表示一个训练 ROI 中包含的 w_j 实例的数量;h 是平滑参数,由于 n 是离散值,所以简单地取为 1。

1.3 敏感度特征

非色情图像中仍然可能存在一些高敏感度视觉单词,但是这些单词的分布规律与色情图像会有所不同。为此在 CDMap 中提取了如下特征来描述图像中敏感单词的分布:(1)CDMap 所有非零点的均值;(2)非零点的值的均方差;(3)最大值;(4)所有局部最大值的均值;(5)所有局部最大值的均方差;(6)超过阈值 t_1 的局部最大值位置的个数;(7)散度(4 维),CDMap 分别在 X 轴和 Y 轴上的质心位置和到质心距离的均值;(8)超过阈值 t_2 的点的个数;(9)非零点的个数。其中 t_1、t_2 是从一个小数据集中统计得到的阈值。

1.4 降维及分类器

检测器所使用的特征向量 $\boldsymbol{x}$ 包括上述所有 SIFT 单词、纹理单词、空单词、词组(包括空词组)、ROI 话题的直方图,以及敏感度特征。其中 ROI 话题的直方图为

$$h(z_j) = \sum_{d \in I, P(z_j \mid d) > 0.1} P(z_j \mid d) \tag{7}$$

I 是输入图像。空单词和空词组所起的作用类似于传统的皮肤区域特征。系统最终采用的分类器是 SVM,其判断函数 $g(\boldsymbol{x})$ 对应 $\boldsymbol{x}$ 到 SVM 分类面的距离。

面对如此长的一个特征向量(超过 1000 维),系统需要降维的处理。更重要的是,还要求图像的特征向量间的距离能体现图像间的语义距离。为此,有必要进行特征向量的映射,使得在新特征空间中 SVM 具有更好的区分能力。设映射转换矩阵为 $\boldsymbol{F}$,新特征向量 $\boldsymbol{y}_i = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{x}_i$。设 c_i 表示训练图像 i 的类别符号(取 ±1),d_e 表示向量之间的欧氏距离,$\boldsymbol{F}$ 按下式求得:

$$\boldsymbol{F} = \arg\max_{F} \frac{\left| \sum_i \left[\sum_j (\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^{\mathrm{T}} SD(i,j) U_i U_j / VD_i(x_j) \right] \right|}{\left| \sum_{i \in porn} \left[\sum_{j \in porn} (\boldsymbol{y}_i - \boldsymbol{y}_j)(\boldsymbol{y}_i - \boldsymbol{y}_j)^{\mathrm{T}} VD_i(\boldsymbol{x}_j) \right] \right|} \tag{8}$$

$$U_i = \begin{cases} 1 & c_i g(\boldsymbol{x}_i) > 1 \\ 2 - c_i g(\boldsymbol{x}_i) & c_i g(\boldsymbol{x}_i) \leqslant 1 \end{cases},$$

$$SD(i,j) = \begin{cases} 1 & c_i \neq c_j \\ 0 & \text{else} \end{cases},$$

$$VD_i(\boldsymbol{x}_j) = 3 - 3\exp\left(-\frac{d_e(\boldsymbol{x}_i, \boldsymbol{x}_j)^2}{2\sigma_i^2}\right) \tag{9}$$

这里 U_i 表示:如果 $\boldsymbol{x}_i$ 临近分类面,则重点考察,其中 $g(\boldsymbol{x})$ 来自于一个临时的基于原始特征向量的 SVM。$SD(i,j)$ 和 $VD_i(\boldsymbol{x}_j)$ 则分别体现着图像 i、j 之间的语义距离和空间距离。式(8)意图在新特征空间中拉大不同类别的图像之间的距离,特别是当它们在原特征空间中彼此接近或靠近分类面时,同时拉近那些彼此距离较远的色情图像。由于内容的多样性,同一大类图像往往分布于不同的子空间中[21]。为此 $VD_i(\boldsymbol{x}_j)$ 的参数 σ_i 设为 $\boldsymbol{x}_i$ 到其前 30 近邻的距离的最大值,从而融合了 $\boldsymbol{x}_i$ 周围的局部子空间结构。代入 $\boldsymbol{y}_i = \boldsymbol{F}^{\mathrm{T}}\boldsymbol{x}_i$,令

$$\boldsymbol{S}_B = \sum_{i \in porn} \left[\sum_{j \in porn} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} VD_i(\boldsymbol{x}_j) \right],$$

$$\boldsymbol{S}_A = \sum_i \left[\sum_j (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} SD(i,j) U_i U_j / VD_i(\boldsymbol{x}_j) \right],$$

有

$$\boldsymbol{F} = \arg\max_{F} \frac{\left|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{S}_A\boldsymbol{F}\right|}{\left|\boldsymbol{F}^{\mathrm{T}}\boldsymbol{S}_B\boldsymbol{F}\right|} \tag{10}$$

该式的解为 $\boldsymbol{F}$ 由 $\boldsymbol{S}_B^{-1}\boldsymbol{S}_A$ 的对应高特征值的特征向量组成[22],$\boldsymbol{F}$ 的具体列数将由实验观察得到。最终针对新生成的特征向量,训练得到高斯核的 SVM 分类器。

2 实验结果

在实验中关注三种类型的图像:色情图像(27707 幅)、人物类图像(26640 幅)以及其它类图像(78439 幅),均来自于互联网和 Corel 图像库。三种图像各随机抽取 10000 幅肤色比例在 5% 以上的图像,一半进行训练,另一半用于测试比较各种特征组合。另外,手工标注了 3000 幅色情图像的身体区域和敏感部位,用来产生各种视觉单词以及肤色模型。

聚类产生了 800 个 SIFT 单词、250 个纹理单词(训练描述向量均来自于标注的敏感部位);构造词组时所用的小型单词表包括 150 个 SIFT 单词和 50 个纹理单词。在构造典型词组时,调整剪枝参数,先建立 2000 个典型词组。然后使用线性 SVM 作为特征选择工具[23],对所有词组进行排序。设定不同的词组集合大小,选取排在前面的词组,产生不同的词组表。然后结合基层单词直方图和敏感度特征,比较不同词组集合的性能(训练得到 SVM)。限于计算复杂度,只选用排名最高的前 600 个词组。类似地,选定 ROI 话题数量为 300。为了证明高层单词具有更低的歧义性和更好的描述能力,基于标注的敏感区域(性器官,来自 1033 幅图像)数据,为各层单词计算了它们与敏感区域的相关性分数:score(w),等于 $F(w \mid \text{敏感区域})/F(w \mid porn)$。图 2 展示了三层视觉单词的相关性分数(经过排序,并各自按等距抽取 300 个)。如图 2 所示,高层单词(ROI 话题,词组)对应的 score(w)曲线更为陡峭,显示高层描述能够更明确地限定哪些单词(编码)与敏感区域相关,而其他单词则明显无关。

接下来要考察各种特征的性能,包括:敏感度特征(CDM,12 维),SIFT 单词直方图(SIFT,800 维),普通单词直方图(BASIC,包括 SIFT、纹理单词 250 维、空单词),词组直方图(PHRA,672 维,包括空词组),ROI 话题直方图(ROI,300 维)。这里所使用的性能指标是 CorrectRate,即在色情类(非色情类)图像的集合中,被判断为色情类(非色情类)的比例。表1

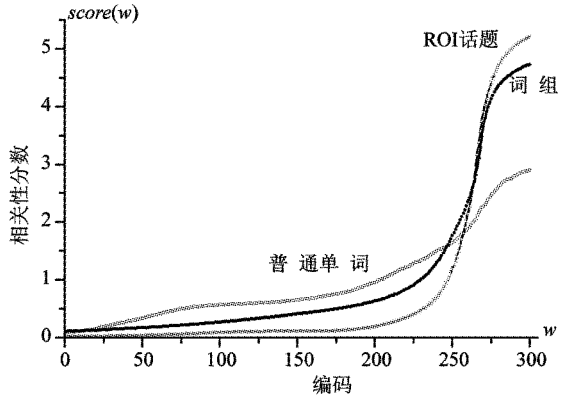

图 2 三层视觉单词在敏感区域中相关性分数的比较

列出各种特征组合的性能,此处未进行特征空间映射。其中 PAIR 代表文献[11]所提出的以单词对为词组的方法。ALL 表示使用了上述所有的特征。从表 1 中可以看到:(1)本文提出的词组由于具有更好的描述能力,性能高于 PAIR;(2)SIFT 方案是典型的基于视觉单词的场景分析方法[10],与之相比,多层视觉单词显著提高了识别性能,尤其是在人物类图像中;(3)ROI 单独的性能不如词组,部分原因是每幅图像中兴趣区域数量有限,造成 ROI 话题直方图比较稀疏;(4)各种特征均有其局限性,最终放在一起产生令人满意的性能。

表 1 各种特征方案的 CorrectRate(%)

特征方案 \ 类别	色情类	人物类	其它类
CDM	79.20	72.88	87.12
SIFT	86.56	75.64	95.68
BASIC	87.28	78.96	96.28
PAIR	82.18	76.96	94.16
词组(600 维)	82.46	83.34	96.12
ROI	82.96	76.82	96.64
CDM + BASIC	88.22	80.54	96.72
CDM + BASIC + PHRA	89.28	82.82	96.80
CDM + BASIC + ROI	89.46	82.08	97.00
BASIC + PHRA + ROI	90.08	83.82	97.56
ALL	90.86	85.00	98.62

使用 20% 的训练数据来计算矩阵 $\boldsymbol{F}$。对比拥有不同列数的映射矩阵 $\boldsymbol{F}$,最终选用 $\boldsymbol{S}_B^{-1}\boldsymbol{S}_A$ 最高 500 个特征值对应的特征向量,排列组成 $\boldsymbol{F}$。

最后,本文对比了三种色情图像检测方案:S_{BD},来自于文献[6],是基于身体模型的检测方法的代表;S_{TRD},来自文献[8],是基于传统类型特征的典型方法;作者以往提出的基于 SIFT 视觉单词分布的检

测方法[15]，S_{WD}。使用相同的训练数据，在总测试集(图像总集中排除训练图像的部分)中取得的结果列于表 2。S_{BD}在色情图像中的性能较差，主要是难以区分那些很近或很远的镜头，尤其是身体部位的特写，但其在姿态端正的人物类图像中具有较好的性能。而传统类型方法 S_{TRD}则明显受肤色检测的影响，所以在肤色区域较多的人物类图像中性能较差。S_{WD}能够较好地分析色情图像中关键区域的视觉单词分布，但由于仅利用了普通视觉单词，其总体性能与本文的方法相比仍有明显差距。

从表 2 可以看到，对于肤色较为少见的“其它类”图像，尽管中心区域的视觉单词都被提取，但带来的混淆很少。最后在一些易于产生误判的图像类别中进行对比。表 3 显示，本文的方法在比基尼照片集中性能略差于 S_{BD}，但在其它类型图像中性能优异。

表 2 与其它检测方法的 CorrectRate 对比(%)

方法 \ 类别	色情类	人物类	其它类
S_{BD}	79.8	85.7	98.1
S_{TRD}	84.6	75.7	95.7
S_{WD}	87.9	81.1	97.2
本文方法	91.3	91.8	99.6

表 3 在易产生误判的图像类别中判为色情的比例(%)

方法 \ 类别	S_{BD}	S_{TRD}	S_{WD}	本文方法
头像(1950)	9.32	13.68	9.80	3.22
比基尼(1839)	12.02	29.58	21.81	13.49
动物(805)	4.97	16.02	7.08	1.25
人造物(1361)	6.47	16.46	8.67	1.10
自然风光(892)	7.96	18.16	12.11	0.20

3 结 论

本文对视觉单词出现规律及其上下文关系，进行了多层次、多角度的分析，有效地识别了色情图像。实验结果证明，相比于传统类型方法，本文的方法不再从根本上依赖肤色检测，能明显降低误检率，尤其是在人物类图像的测试中。由于算法对色情图像进行了深入和全面的分析，其处理速度尚不适用于个人主机的实时保护，但适合应用于对可疑网页的后台分析，有助于肃清互联网中的色情垃圾信息。

参考文献

[1] Jeong C Y, Han S W, Choi S G, et al. An objectionable image detection system based on region of interest. In: Proceedings of 2006 IEEE International Conference on Image Processing, Atlanta, USA, 2006. 1477-1480

[2] Shih J, Lee C, Yang C. An adult image identification system employing image retrieval technique. *Pattern Recognition Letters*, 2007, 28(16): 2367-2374

[3] 温泽逢，袁华. 基于内容的图像过滤新方法. 通信学报，2006，27(11)：280-284

[4] Lee J, Kuo Y, Chung P, et al. Naked image detection based on adaptive and extensible skin color model. *Pattern Recognition*, 2007, 40(8): 2261-2270

[5] 万月亮，李文正，曹元大等. 基于统计肤色模型的敏感图像检测. 高技术通讯，2008，18(6)：596-601

[6] Hu W, Wu O, Chen Z, et al. Recognition of pornographic web pages by classifying texts and images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(6): 1019-1034

[7] 杨金锋，傅周宇，谭铁牛等. 一种新型的基于内容的图像识别与过滤方法. 通信学报，2004，25(7)：93-106

[8] 曾炜，郑清芳，赵德斌. 图片卫士：一个自动成人图像识别系统. 高技术通讯，2005，15(3)：11-16

[9] Arentz W A, Olstad B. Classifying offensive sites based on image content. *Computer Vision and Image Understanding*, 2004, 94(1-3): 295-310

[10] Quelhas P, Monay F, Odobez J M, et al. A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(9): 1575-1589

[11] Zheng Q, Wang W, Gao W. Effective and efficient object-based image retrieval using visual phrases. In: Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, USA, 2006. 77-80

[12] Savarese S, Winn J, Criminisi A. Discriminative object class models of appearance and shape by correlations. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, USA, 2006. 2033-2040

[13] Zhang W, Deng H, Dietterich T G, et al. A hierarchical object recognition system based on multi-scale principal curvature regions. In: Proceedings of the 18th International Conference on Pattern Recognition, Hong Kong, China, 2006. 778-782

[14] Lowe D G. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 2004, 60(2): 91-110

[15] 王宇石，李远宁，高文. 基于局部视觉单词分布的成人图像检测. 北京理工大学学报，2008，28(5)：410-

413

[16] Ojala T, Pietikainen M, Maenpaa T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(7): 971-987

[17] Agarwal A, Triggs B. Hyperfeatures: multilevel local coding for visual recognition. In: Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 2006. 30-43

[18] Yu J, Tian Q. Learning image manifolds by semantic subspace projection. In: Proceedings of the 14th ACM International Conference on Multimedia, Santa Barbara, USA, 2006. 297-306

[19] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001, 42(1/2): 177-196

[20] Webb A. Statistical Pattern Recognition. 2nd edition. Hoboken, USA: John Wiley & Sons Inc, 2002. 106-113

[21] Yuan J, Wu Y, Yang M. Discovery of collocation patterns: from visual words to visual phrases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, USA, 2007. 1-8

[22] Swets D, Weng J. Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1996, 18(8): 831-836

[23] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd edition. San Francisco, USA: Morgan Kaufmann, 2005. 291

Pornographic image detection based on visual words and semantic projection

Wang Yushi[*], Gao Wen[*,**]

([*] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001)

([**] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871)

Abstract

This paper proposes a new method to detect pornographic images using multi-level visual words because the traditional methods suffer from high false positive rates. The method is described as below. First, various visual components of pornographic scenes are coded as visual words. Then, higher-level codes are constructed based on those words, including visual phrases and classes of regions of interest (ROI). Thus images can be represented and analyzed in different appearance levels. The image features are composed of the histograms of those codes, and they are projected into a low-dimensional space to bridge the gap between the semantic similarities and geometric distances of image pairs. The proposed method performs well on a wide range of test data. For example, for human images, the method can decrease the false positives by 40% compared to the baseline methods. The experimental results demonstrate that the multi-level visual words are more effective in analyzing complex pornographic scenes.

Key words: image recognition, pornographic image, visual word, visual phrase, multi-level representation, dimension reduction