用基于视觉单词上下文的核函数对图像分类

王宇石[1)] 高 文[1),2)]

[1)](哈尔滨工业大学计算机科学与技术学院,哈尔滨 150001) [2)](北京大学信息科学技术学院,北京 100871)

摘 要 当前在图像分析领域,将局部特征编码为视觉单词的做法非常流行。基于普通的视觉单词,提出了一种新的能够融合单词多层上下文的核函数。设计中体现了如下信息:1)多层的单词直方图;2)多层的“词组”直方图;3)单词(以及词组)的上下文的类别。然后将该核函数应用于支持向量机,对图像进行分类。在 Corel 图像库等公共测试集上,该方法取得出色的性能。此外,在一个实用性很强的复杂问题中进行了对比:识别成人图像和泳装图像。该方法的识别准确率,比经典方法提高了约 7%。实验结果表明,将核函数度量同视觉单词的多层次描述结合在一起,能够显著提高图像的识别能力。

关键词 图像分类 核函数 支持向量机 视觉单词 多分辨率直方图

中图法分类号:TP391.41 **文献标志码**:A **文章编号**:1006-8961(2010)04-607-10

Kernel-based Image Classification Using the Context of Visual Words

WANG Yushi [1)], GAO Wen [1),2)]

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

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

Abstract In recent literature of image analysis, it has been very popular to code local features into visual words. We propose a novel kernel which fuses multi-level contexts of visual words. Besides the histogram pyramid of words, our kernel also incorporates the histogram pyramid of visual phrases (the local co-occurrence patterns of words) and the context classes of those words and phrases. Then support vector machines using the kernel are trained to perform image classification. Our method performs well on a wide range of test data, such as the Corel dataset. The method is also tested in a challenging problem, the discrimination of pornographic images from bikini ones. The classification accuracy of our method is 7% higher than that of the baseline method. Experimental results demonstrate that the performance of image classification can be improved by the integration of kernel based measurements and the multi-level representation of visual words. In the future work, more compact and efficient representation of contexts should be researched.

Keywords image classification, kernel, support vector machine, visual word, multi-resolution histogram

0 引 言

图像分类是计算机视觉领域一个重要的分支。当前,伴随着数字技术的迅猛发展和普及,互联网成为了一个巨大的数字图像库,大大丰富了人们的日常生活。作为一个极具挑战性的问题,图像分类可以应用于对个人或专业的图库(包括视频库)进行组织与检索。在本文中,图像分类既包括对图像场景的分类,也可以是对图像的中心物体进行分类。图像分类早期的做法是提取图像的全局描述特征,例如边缘特征、局部颜色特征、傅里叶变换系数等。

基金项目:国家高技术研究发展计划(863)项目(2003AA142140);国家自然科学基金项目(60702035)

收稿日期:2008-11-12;**改回日期**:2009-02-11

第一作者简介:王宇石(1978—),男。哈尔滨工业大学计算机学院博士研究生。主要研究方向为图像识别,改进了基于视觉单词的图像识别方法。E-mail: yswang@jdl.ac.cn

这类方法中代表性的工作有文献[1]和文献[2]。进而出现了基于图像局部区域的分类方法,这些工作探讨了局部区域和全局语义的关联[3-5]。

近年来,基于“视觉单词”的图像分析方法非常流行。所谓视觉单词,是对图像局部兴趣点(例如某种突出的视觉元素)进行适当地描述,并将描述向量量化为编码,这些编码即为视觉单词[6-7](简称为单词)。利用图像中视觉单词的出现规律,可以判断图像的类别。最简单的分析方式就是提取视觉单词的直方图[8-9]。由于将图像的视觉元素量化为单词,在某种程度上使得图像具有和文本类似的形式,部分研究者开始用文本语义分析模型来识别图像[9-12]。另一个研究方向是,建立图像关于视觉单词的更高层级的描述,即从单词的局部共现规律着手分析,例如“词组”[13-14],或多层视觉单词[15]。

上述将图像视为“单词包”的策略,实质是在比较两个编码集合的相似性。研究者们发现,基于核函数的分类方法非常适于此项工作[16-17]。所谓核函数,是一种衡量输入数据点之间相似性的函数,能够将普通的输入特征映射到非线性空间中。核函数能够高效地表达复杂的分类面,并具有优良的推广能力[16],例如支持向量机(SVM)的使用已很普及。在对于核函数的研究中,早期的工作存在如下局限:或者计算复杂度过高(达到特征数的平方或立方)[18-20];或者核函数不具有正定性[18,21],这导致支持向量机的优化无法得到唯一的解。文献[22]、[23]的方法使用特定的概率模型来描述特征的分布;Qi 等人的方法也是类似的策略[24],并将隐马尔可夫模型融入到图像的建模中。Grauman 等人的工作实现了一个计算高效(线性复杂度)且具有正定性的核函数[25],他们将单词描述向量空间进行逐层细分,得到一个金字塔式的单词直方图,并利用直方图的“相交”得到核函数的值。他们证明了基于直方图金字塔构成的核函数能够高效地区分不同的图像类别。但是,这种方法将图像的单词集合视为无序的集合,没有利用视觉单词的局部共现关系。文献[26]则相反,不是逐层细分特征向量空间,而是将图像本身进行逐层分割,形成对单词位置分布的金字塔式描述。

本文提出了一种新的基于核函数的图像分类方法,核心工作是设计了一种核函数,该核函数融入了图像局部视觉特征(单词)的多层次分析。分析视觉单词的局部相关性,能够提高对图像的描述能力。为此,分两个层次分析了单词间局部的共现规律:1)针对单词的相邻关系,引入了一种简洁、高效的“词组”模型;2)每一类图像中存在着自身独特的视觉元素,发现这类视觉元素集中的兴趣区域(ROI),并将之归类为不同的“话题”(即局部场景类别)。然后,对于词组模型和 ROI 的话题模型,通过多重的粒度实现了相应的直方图金字塔,并将这些单词的上下文描述,融入到核函数的设计中。实验结果表明,将基于直方图金字塔的核函数,同视觉单词的多层次分析结合在一起,能够显著提高图像的识别能力。

1 图像分类方法

本文提出了一种新的基于核函数的图像分类方法,在此核函数中,体现了以下 3 个方面的图像距离:1)视觉单词的总体出现情况;2)视觉词组(视觉单词的典型相邻关系)的总体出现情况;3)在较大的图像区域内(ROI),分析其中视觉单词(词组)的上下文关系,依据这些上下文关系,给上述单词(词组)赋以权重。然后将证明该核函数符合 Mercer 条件[16],从而能够确保实现 SVM 的优化。最终使用 SVM 作为分类器。

1.1 提取视觉单词

本文采用视觉单词稀疏的提取方式,使用高斯差分(DoG)兴趣点检测子[27]搜索图像中局部发生突出变化的位置。相比于其他常见的工具,例如 Harris 角点检测器、Salient 检测子等,DoG 能够检测多种类型的兴趣点,并具有适中的计算复杂度[6,9]。在 DoG 算法中,需要计算图像中各个位置(x,y)在各级尺度因子(s)下的高斯均值$G(x,y,s)$,是计算量负担最重之处。本文用平板均值代替了高斯均值,在实现中使用式(1)显著降低了 DoG 的计算复杂度:

$$\begin{aligned}G(x,y,s)=&(Sum(x+s,y+s)-\\&Sum(x+s,y-s)-Sum(x-s,y+s)+\\&Sum(x-s,y-s))/(2s+1)^2\end{aligned}\tag{1}$$

$$\begin{aligned}Sum(a,b)=&\sum_{x=0\sim a,y=0\sim b}I(x,y)=Sum(a-1,b)+s(a,b);\\&s(a,b)=s(a,b-1)+I(a,b)\end{aligned}\tag{2}$$

$I(x,y)$为图像灰度值,$Sum(a,b)$是一个速算表,按式(2)的方法递归求得[28],表示 $\boldsymbol{I}$ 中位置(a,b)左上部分的像素灰度和作者前面的工作[29]已经验证,简化后的 DoG 算子同样具有足够的检测能力。

在采集图像中的兴趣点之后,需要在其周围区域建立描述向量,刻画图像在此处的局部形态。在文献中研究者提出了各种描述子,其中最广为使用的是尺度不变特征变换(scale invariant feature transform,SIFT)算法[27],这是一种 128 维的局部梯度的方向-分布描述子。该算子将图像的局部区域等分成 4×4 的分区,在每个分区中提取梯度方向直方图,其中梯度方向被量化为 8 个方向。所有分区的梯度方向直方图合并在一起,构成 128 维的局部描述向量。其他形式的描述子,在适当的应用中,同样可以使用。

随后,对从训练集中提取的 SIFT 描述向量集合进行聚类,产生的每个聚类作为一个视觉单词。这是一个描述向量量化的过程,通常采用的手段是 K-均值聚类。本文采用文献[17]的多层聚类方式,构造一个树结构的单词表 $T^{(w)}$:第 1 层只有一个节点(w_{11}),代表整体描述向量空间;在第 2 层中,利用 K-均值聚类,把向量空间分成 k 个子空间(即产生这一层上的 k 个单词:$w_{21},\cdots,w_{2k}$);在下一层,则把这 k 个子空间,各自再分成 k 个子空间,共 k^2 个子空间($w_{31},\cdots,w_{3k^2}$);如此重复,直到第 L_w 层,这一层共有 k^{L_w-1} 个子空间。在测试时,对于图像的某个局部描述向量 $\boldsymbol{v}$,从根节点开始,在每一层中都要归入相应的子空间。例如在第 2 层,把 $\boldsymbol{v}$ 归属为距离最近的子空间 $\boldsymbol{g}$;然后在 $\boldsymbol{g}$ 的所有 k 个下层子空间中,又使 $\boldsymbol{v}$ 归属为距 $\boldsymbol{v}$ 最近的儿子空间;如此重复向下,直到最底层,相当于 $\boldsymbol{v}$ 从 $T^{(w)}$ 的根节点前进到某个叶子节点。

1.2　建立视觉词组

考察视觉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[15],还能降低单词的歧义性[14]。例如,不同的事物可能会具有相似的结构(人眼和某些树叶),如果被归类为同一单词,则分析其邻域背景内其他单词的分布(是五官,还是大量其他树叶),有助于提高识别的精确性。为此,本文首先借助一种简洁、高效的“词组”模型,从视觉单词的相邻关系着手进行描述。并形成一种多层次的机制,就如同 2.1 节的视觉单词那样,能够在不同的粒度上刻画这种相邻关系,便于后面建立多层直方图。

选定 $T^{(w)}$ 的某一层的所有 V 个单词,作为构造词组用的单词表(详情请参看 3.1 节关于方法实现的部分)。该算法将以其中的每一种单词为中心,构造一组典型词组。以单词 w 为例,对于图像中 w 的某个实例 P,设 $W(P)$ 表示 P 对应的视觉单词编号。在 P 周围 2 倍于 P 的尺度因子的距离内,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。由于图像中单词分布可能并不均匀,为考察邻域内单词分布的密度定义了一种“空邻居”(EN)。如图 1 所示,在 P 周围的每个等分扇形内,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 $\boldsymbol{S}_w$ 是以 w 为中心的局部单词直方图训练集,通过对 $\boldsymbol{S}_w$ 进行聚类,获取以 w 为中心的典型词组。考虑到上述局部单词直方图非常稀疏,并且为了易于查表,使用 Liu 等人提出的聚类树(clustering tree,CLTree)[30] 作为聚类工具。

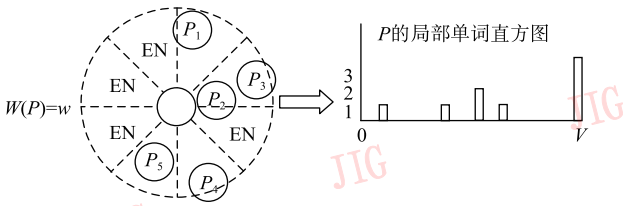


图 1　创建词组表的示意图(以单词 w 为例)

Fig. 1　Illustration of constructing the phrases for a word w

CLTree 是一种以建立决策树为架构的聚类方法,利用 $\boldsymbol{S}_w$ 产生一株决策树 T_w,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 $\boldsymbol{C}_w$。在 CLTree 算法中只要指定了总体的剪枝参数[30],就不需要规定在 $\boldsymbol{S}_w$ 中聚类时应该产生多少个聚类。最终,总的词组表是 $\bigcup_w \boldsymbol{C}_w$。

文献[13]以相邻的单词对作为词组。本文构造词组的方法,相比于文献[13],除了能简洁、有力地描述相邻单词的共现关系,还进一步描述了单词局部共现的频率(直方图)和密度。

更为重要的是,本文的视觉词组描述算法,能够对单词局部相邻关系进行多粒度描述,便于后面建立多层次直方图。即可以给每个词组实例在不同的层次(粒度)中赋予对应的词组编号,如同在 1.1 节的 $T^{(w)}$ 中一样。总体词组表 $T^{(p)}$ 由 ν 株 CLTree 树组成。第 1 层的词组集合为 $\boldsymbol{C}(1)$,等同于产生词组的单词表——其中每个词组仅要求以各个单词为中心,不考虑邻域情况。即 $\boldsymbol{C}(1)$ 中的 ν 株树都只有根节点。显然,通过不同的剪枝参数,CLTree 可以产

生粒度不同的聚类(词组)。设采用苛刻的剪枝条件,生成词组表 $\boldsymbol{C}(2)$(包含 ν 株 CLTree 树此时所有的叶子节点);选取较为宽松的剪枝参数,则生成词组表 $\boldsymbol{C}(3)$。对于其中的某词组 p_{3j}($p_{3j}\in\boldsymbol{C}(3)$),必有 $p_{3j}\in\boldsymbol{C}(2)$ 或 p_{3j} 是 $\boldsymbol{C}(2)$ 某词组的后代(在对应的 CLTree 树中)。以此类推,可以得到第 i 层的词组表 $\boldsymbol{C}(i)$。注意,不要将这里词组的上下层关系和 CLTree 树里的节点父子关系弄混。对于 $\boldsymbol{C}(i)$ 层的某词组 p_{ij},设其对应所在 CLTree 树的某个节点,但它的某儿子节点 c_{ij} 未必收录在 $\boldsymbol{C}(i+1)$ 中,而可能是 $\boldsymbol{C}(i+1)$ 某些词组在此树中的祖先节点。

1.3 通过 ROI 话题来描述上下文

前面已经建立了关于视觉单词及其相邻关系(词组)的多层次描述体系。尽管词组能够在一定程度上刻画图像视觉元素的上下文关系,但限于词组描述的复杂度,仅包含了近邻单词的共现规律。为了能更多地利用较大范围内的上下文关系,本文提出,利用 ROI 话题分布来描述图像局部区域的语义。

设 C_{interet} 是当前关注的图像类。兴趣区域(ROI)在本文中定义为如下的局部区域,其相对集中地包含了与 C_{interet} 相关的视觉信息。为了要获取 ROI,需要定义单词(及词组)的相关度(CD),表示一个单词(词组)与 C_{interet} 的相关性。设 $F(w\mid C_{\text{interest}})$ 表示出现单词 w 的 C_{interest} 类图像的比率,$F(w\mid\overline{C}_{\text{interest}})$ 是出现 w 的非 C_{interest} 类图像的比率,则有

$$CD(w)=F(w\mid C_{\text{interest}})\cdot\left(\frac{F(w\mid C_{\text{interest}})}{F(w\mid\overline{C}_{\text{interest}})}\right)^2 \tag{3}$$

注意,这里的单词(词组)也是来自于 $T^{(w)}$($T^{(p)}$)特定的一层,不是针对所有层次的单词(词组)。

ROI 具体定义为:与 C_{interest} 类高度相关的视觉单词相对集中的局部区域。产生一个尺寸为原始图像 1/16 的缩图,称为 CDMap(Correlation degree map),其中每个位置 p 的值为

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

式中,$phrase(\nu)$ 表示单词实例 ν 的词组编号。然后在 CDMap 中可以快速地通过 K-均值聚类算法提取 ROI:在 2 维的 CDMap 中,每个位置对应的数据点的个数就是 $CDM(p)$;聚类数可以由 CDMap 中的局部最大值的个数决定;最终,得到的每个聚类就对应一个 ROI,属于该聚类的单词实例就属于该 ROI。图 2 展示了 ROI 提取的 3 个实例。在每一组的两幅图像中,右图展示了左图的 CDMap 以及 K-均值分割结果。

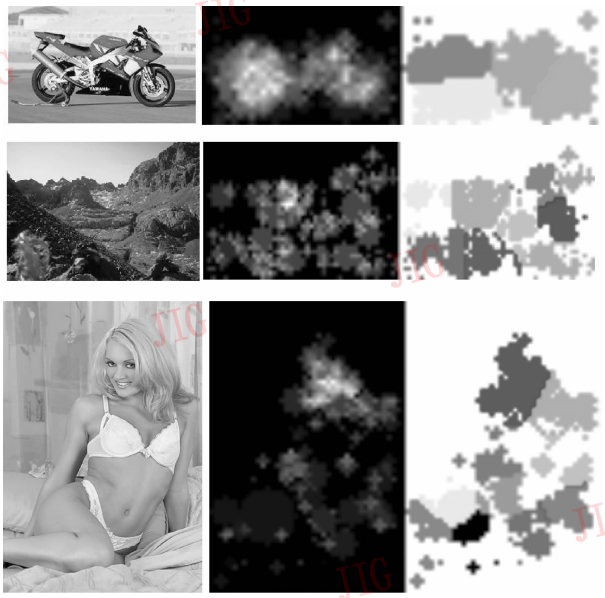


图 2 ROI 提取实例

Fig. 2 Extraction of ROIs

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[31]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”,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 $P(w_j\mid z_k)$。

对于图像中某个 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

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

式中,$n(d,w_j)$ 表示 w_j 在 d 中出现的次数。$P(n\mid w_j,z_k)$ 用 Parzen 方法[32]予以估计:

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

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

式中,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,$K(x)$ 相当于一个高斯核。最终通过 $\arg\max_{z_k}(P(d \mid z_k))$ 判断 d 的主话题。

作为一种文本(图像)聚类工具,PLSA 模型同样也存在聚类粒度变化的问题。在对 ROI 的分析中,如果设定较少的话题数量(较粗的粒度),则一个话题可能涉及目标类图像较多变化的场景内容;反之则一个话题可能只对应目标类图像特定场景或局部的内容。设总体的 ROI 话题表为 $T^{(t)}$,并具有 L_{topic} 个级别,每个级别具有不同的粒度。其中第 l 级具有的话题数为 $N_l^{(t)}$,且有 $N_{l-1}^{(t)} < N_l^{(t)} < N_{l+1}^{(t)}$。设单词实例 ν 所处的 ROI 的第 l 级的话题编号为 $t^l(\nu)$。

1.4　新核函数

本文提出的核函数如下所示:

$$K(\boldsymbol{I}_a, \boldsymbol{I}_b) = K_w(\boldsymbol{I}_a, \boldsymbol{I}_b) + \alpha \cdot K_p(\boldsymbol{I}_a, \boldsymbol{I}_b) \tag{7}$$

式中,$\boldsymbol{I}_a, \boldsymbol{I}_b$ 表示两幅图像,$K_w(\boldsymbol{I}_a, \boldsymbol{I}_b)$ 代表从单词共现的角度来衡量图像的相似性,而 $K_p(\boldsymbol{I}_a, \boldsymbol{I}_b)$ 则表示从词组(即局部单词相邻关系)共现的角度来衡量图像的相似性。α 是平衡两方面作用的权重系数,需要在具体的应用中,设为 $K_w(\boldsymbol{I}_a, \boldsymbol{I}_b)$ 均值和 $K_p(\boldsymbol{I}_a, \boldsymbol{I}_b)$ 均值的比,以平衡两方面在 $K(\boldsymbol{I}_a, \boldsymbol{I}_b)$ 中作用的比重。

为了 $K_w(\boldsymbol{I}_a, \boldsymbol{I}_b)$ 定义的方便,首先定义如下函数:

$$H_{i,j}^{(w)}(\boldsymbol{I}_a, \boldsymbol{I}_b) = \sum_{l=1}^{L_{\text{topic}}} \sum_{m=1}^{N_l^{(t)}} Q_{l,m}^{(t)} \cdot \min\left[n_{l,m}^{(\text{topic})}(S_{i,j}^{(w)}(\boldsymbol{I}_a)), n_{l,m}^{(\text{topic})}(\boldsymbol{S}_{i,j}^{(w)}(\boldsymbol{I}_b))\right] \tag{8}$$

式中,$\boldsymbol{S}_{i,j}^{(w)}(\boldsymbol{I})$ 表示图像 $\boldsymbol{I}$ 中归属为单词 w_{ij}($T^{(w)}$ 中第 i 层第 j 个单词)的单词实例集合。$n_{l,m}^{(\text{topic})}(\boldsymbol{S}) = |\{\nu \mid t^l(\nu) = m, \nu \in \boldsymbol{S}\}|$。而 $Q_{l,m}^{(t)}$ 表示第 l 层第 m 个话题 $t_{l,m}$ 的权重,定义为

$$Q_{l,m}^{(t)} = \exp(-T_{l,m}/\sigma_{\text{topic}}) \tag{9}$$

这里 $T_{l,m}$ 表示 $t_{l,m}$ 平均在每幅(训练)图像中出现的次数,σ_{topic} 表示所有 $T_{l,m}$ 的均值。

然后 $K_w(I_a, I_b)$ 定义为

$$K_w(I_a, I_b) = \sum_{i=1}^{L_w} \sum_{j=1}^{k^{i-1}} Q_{ij}^{(w)} \left(H_{i,j}^{(w)}(\boldsymbol{I}_a, \boldsymbol{I}_b) - \sum_{c=1}^{k} H_{i+1,kj-k+c}^{(w)}(\boldsymbol{I}_a, \boldsymbol{I}_b)\right) = \sum_{i=1}^{L_w} \sum_{j=1}^{k^{i-1}} (Q_{ij}^{(w)} - P_{ij}^{(w)}) H_{i,j}^{(w)}(\boldsymbol{I}_a, \boldsymbol{I}_b) \tag{10}$$

式中,L_w 表示 $T^{(w)}$ 的层数,$Q_{ij}^{(w)}$ 表示 w_{ij} 的权重,而 $P_{ij}^{(w)}$ 表示 w_{ij} 的父单词的权重。注意 $H_{L_w+1,\bullet}^{(w)}(\ )$ 一律取为 0,$P_{11}^{(w)} = 0$。式(10)的目的是考察:图像 $\boldsymbol{I}_a$ 和 $\boldsymbol{I}_b$ 在 w_{ij} 中"相合"的单词实例数量,与在 w_{ij} 的各儿子单词中相合的数量有什么差别。毫无疑问,在儿子单词中相合的实例,在父单词中一定相合,而在 w_{ij} 中要考察有多少新相合的实例。这种相合性分数既基于直方图的相交,同时又受到各单词实例所属上下文的话题的影响。对于上下文相似的单词,它们的相合应该带来更高的分数。

各单词以 $Q_{ij}^{(w)}$ 为权重,合计起来衡量 $\boldsymbol{I}_a$ 和 $\boldsymbol{I}_b$ 的相似性。在文献[17]中,$Q_{ij}^{(w)}$ 具有和单词聚类半径成反比的形式,则 $Q_{ij}^{(w)}$ 定义如下:

$$Q_{ij}^{(w)} = \exp(-A_{ij}/\sigma) \tag{11}$$

式中,A_{ij} 表示 w_{ij} 的聚类半径,而 σ 表示训练集中单词实例间的平均向量距离。

$K_p(\boldsymbol{I}_a, \boldsymbol{I}_b)$ 具有和 $K_w(\boldsymbol{I}_a, \boldsymbol{I}_b)$ 非常类似的形式,只不过现在不再是单词树,而是在一组词组树的多个层次上计算直方图($n(\ )$)。

$$H_{i,j}^{(p)}(\boldsymbol{I}_a, \boldsymbol{I}_b) = \sum_{l=1}^{L_{\text{topic}}} \sum_{m=1}^{N_l^{(t)}} Q_{l,m}^{(t)} \cdot \min\left[n_{l,m}^{(\text{topic})}(\boldsymbol{S}_{i,j}^{(p)}(\boldsymbol{I}_a)), n_{l,m}^{(\text{topic})}(\boldsymbol{S}_{i,j}^{(p)}(\boldsymbol{I}_b))\right] \tag{12}$$

$$K_p(\boldsymbol{I}_a, \boldsymbol{I}_b) = \sum_{i=1}^{L_p} \sum_{j=1}^{N_i^{(p)}} Q_{ij}^{(p)} \left(H_{i,j}^{(p)}(\boldsymbol{I}_a, \boldsymbol{I}_b) - \sum_{p_{(i+1)m} \text{是} p_{ij} \text{的后代}} H_{i+1,m}^{(p)}(\boldsymbol{I}_a, \boldsymbol{I}_b)\right) = \sum_{i=1}^{L_p} \sum_{j=1}^{N_i^{(p)}} (Q_{ij}^{(p)} - P_{ij}^{(p)}) H_{i,j}^{(p)}(\boldsymbol{I}_a, \boldsymbol{I}_b) \tag{13}$$

L_p 表示词组所分的层数,$N_i^{(p)}$ 表示在第 i 层即 $\boldsymbol{C}(i)$ 中词组的数量。其余的量和前面定义 $K_w(\boldsymbol{I}_a, \boldsymbol{I}_b)$ 时类似,只是上标相应换成"(p)"表示是词组。其中,

$$Q_{ij}^{(p)} = \exp(-B_{ij}/\rho) \tag{14}$$

B_{ij} 表示归属词组 p_{ij} 的训练实例数量,ρ 表示所有 B_{ij} 的均值。

1.5　新核函数满足 Mercer 条件

核函数需要满足半正定条件,从而能够进行凸函数优化(例如在支持向量机的优化中)。设

$\boldsymbol{a}$、$\boldsymbol{b}$ 是两个向量。在文献[16]中定义,如果一个核函数 K 满足下列条件(Mercer 条件),则是半正定的:

$$K(\boldsymbol{a},\boldsymbol{b})=\langle \boldsymbol{\Phi}(\boldsymbol{a}),\boldsymbol{\Phi}(\boldsymbol{a})\rangle \tag{15}$$

这里〈,〉表示点积,即能够转化为特征空间中两个向量函数($\boldsymbol{\Phi}$)的点积。而且对于已符合 Mercer 条件的函数,其和、与正数之积均为符合 Mercer 条件的函数[16]。在文献[33]中已经得知,min 函数是 Mercer 函数,则 $H_{i,j}^{(w)}(\boldsymbol{I}_a,\boldsymbol{I}_b)$ 和 $H_{i,j}^{(p)}(\boldsymbol{I}_a,\boldsymbol{I}_b)$ 均为 Mercer 函数。显然在设计时就已限定,各级单词(词组)的权重 Q 必定不小于其父单词(词组)的权重。于是 $K_w(\boldsymbol{I}_a,\boldsymbol{I}_b)$ 和 $K_p(\boldsymbol{I}_a,\boldsymbol{I}_b)$ 均为 Mercer 函数,最终 $K(\boldsymbol{I}_a,\boldsymbol{I}_b)$ 为 Mercer 函数。所以,新核函数可以与 SVM 结合使用,具有出色的区分能力。

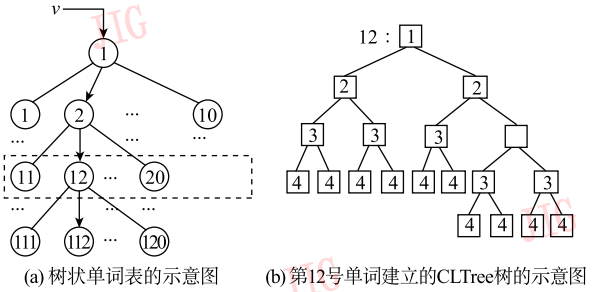
1.6 核函数的实现

从式(10)和式(13)可以看到,$K(\boldsymbol{I}_a,\boldsymbol{I}_b)$ 的计算需要面对数量惊人的直方图扫描。单词和词组不但需要分很多层级,而且单词树和词组树中每一个节点(对应一个单词或词组),都似乎要包含一个关于话题的直方图。实际上,根本无须如此惊人的直方图扫描工作,根据文献[25]的思想,只需要把各个单词、词组实例进行整数编码,然后再进行整数的基数排序(Radix sort)即可。

以计算 $H_{i,j}^{(w)}(\boldsymbol{I}_a,\boldsymbol{I}_b)$ 为例,对于一个单词实例 ν,实际上为其提供一个 L_w+L_{topic} 位的“整数”编码。具体地说:在前 L_w 位中,是 k 进制的,保存 ν 在 $T^{(w)}$ 各层中所属单词的编号;在后 L_{topic} 位中,各位是 $N_l^{(t)}$ 进制的,表示 ν 所属 ROI 在各话题层所属的话题编号。对于基数排序来说,很容易处理这样各位进制混杂的“整数”。在进行基数排序时,后 L_{topic} 位如果基数(对应 $N_l^{(t)}$)太高,可以进行如下优化:把后 L_{topic} 位转化为 10 进制的整数,再进行排序,所增位数有限,但可以方便地执行基数排序。当把图像中所有的单词实例的编码排序完成后,计算两幅图像的 $K_w(\boldsymbol{I}_a,\boldsymbol{I}_b)$ 的核心模块是:扫描两个编码队列,逐个对比两个队列里的当前编码;遇到相同编码,即产生直方图对应元素的计数;否则继续前进,直到把两个有序编码队列扫描完毕。详细原理请参看文献[17]。

对于各个词组实例,也是分别为其提供一个 L_p+L_{topic} 位的“整数”编码。不同的是,前 L_p 位不是 k 进制的,而是 N_{BNPC} 进制的。N_{BNPC} 是“最多子词组”的儿子词组数(biggest number of phrases' children)。

作为总结,用一个虚拟的例子来形象地解释第 1.1—1.4 节中出现的概念,图 3、图 4 展示了相应的示意图。图 3(a)中,虚线框表示词组的建立是以此层的单词表为基础,箭头指示了单词实例 ν 在各级中归属的单词。图 3(b)中,每个方框表示一个词组。框中的数字 l 表示词组的层数(即在第 L_p-l+1 次剪枝时形成的叶子节点)。其中有的节点没有数字,表示它不是一个词组。因为在各次剪枝中,其均不是叶子节点。图 4 中,ν 周围的区域表示 ν 所属的 ROI,其在 3 个话题层中分别被归属为某个话题。右侧各个实心圈表示某层的一个话题,虚线框中所有的话题表示针对 ROI 的多级话题模型。总之,对于单词实例 ν:1)设 $k=10$,$L_w=4$;ν 在第 2 层的编号是 2(总编号是 1~10),在第 3 层的编号是 12(总编号是 1~100),在第 4 层的编号是 112(总编号是 1~1 000)。2)设词组是以 1)中第 3 层那 100 个单词为基础建立的,且 $L_p=4$。3)设 $L_{topic}=3$,$N_1^{(t)}=1$,$N_2^{(t)}=10$,$N_3^{(t)}=100$。



(a) 树状单词表的示意图 (b) 第12号单词建立的CLTree树的示意图

图 3 树状单词表和词组 CLTree 树的示意图

Fig. 3 Illustration of the hierarchical word table and a CLTree for the 12th word

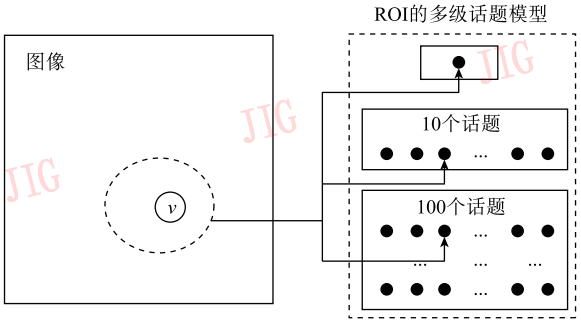


图 4 一个单词实例 ν 的 ROI 话题归属示意图

Fig. 4 Illustration of the ROI topics related to a word instance ν

2　实验结果

将在两组数据中验证本文的图像分类方法的性能。第 1 种来自 Corel 图像库,该数据集包含大量类别已标注的图像。此外,还在一个实用性较强的问题中验证了本文的方法:收集了 3 000 幅包含各种姿势、距离的色情图像和 3 000 幅泳装类图像,作为第 2 个实验数据集 $\boldsymbol{D}_{\text{body}}$。

2.1　方法实现

系统的实现涉及如下细节:

1)在每幅图像中,提取 300 个兴趣点(单词实例),每个兴趣点在每个单词层都产生一个标号,即一个兴趣点有 L_{w} 个单词标号。

2)对于 2.1 节所述的单词树 $T^{(w)}$,根据文献[17]的做法,设 $k=10, L_{\text{w}}=5$。

3)考虑到计算复杂度,本文选用第 3 层单词(共 100 个)来产生 2.2 节所述的词组,即共有 100 株 CLTree 树。然后每个兴趣点实际又有了 L_{p} 个词组标号。

4)$N_1^{(\text{p})}=100$(因为有 100 株 CLTree)。对于 L_{p} 和 $k_{\text{p}}=N_i^{(\text{p})}/N_{i-1}^{(\text{p})}$(即通过设定剪枝参数,以固定比例增加 $\boldsymbol{C}(i)$ 的词组数),限于计算复杂度,有如下组合:$k_{\text{p}}=5, L_{\text{p}}=3,4,5$;$k_{\text{p}}=10, L_{\text{p}}=3,4$;共 5 种组合。

5)$N_1^{(t)}=1$(对应所有场景)。对于 L_{topic} 和 $k_{\text{t}}=N_l^{(t)}/N_{l-1}^{(t)}$(即以固定比例增加各层话题的数量),限于计算复杂度,有如下组合:$k_{\text{t}}=5, L_{\text{topic}}=3,4$;$k_{\text{t}}=10, L_{\text{topic}}=3,4$;共 4 种组合。连同 4)、5),共 20 种组合,需要在各个实验中寻找性能优良的组合。

6)关于 1.3 节中计算的单词(词组)相关度,在单词方面使用第 4 层的 1 000 个单词;在词组方面:$k_{\text{p}}=5$ 时选择第 3 层的词组(2 500 个),$k_{\text{p}}=10$ 时选择第 2 层的词组(1 000 个);由于此处并没有直接利用词组的语义分析图像,所以粒度适当即可。

针对两种经典方法比较了性能:1)文献[9]是典型的基于视觉单词包的图像分析方法,以视觉单词(只有一层)的直方图作为图像特征(Word histogram),使用 SVM 分类;2)文献[25],使用多层单词的直方图金字塔构成 SVM 的核函数(pyramid match kernel,PMK)。

2.2　在 Corel 图像库中的性能

Corel 图像库中包含着大量各种场景类别的图像。为了验证本文的方法在场景识别中的性能,选取了如下的图像类别:头像(上半身)、人(远景为主)、城市、天空、山岭、大海、海岸、树林、草地、荒漠、人文景观(如文物建筑)、洞穴、室内家居、玩具、鱼类、昆虫、哺乳动物、鸟类、瓜果、植物花卉共 20 个大类别。

在建构(多层)词组表时,依据不同的 k_{p} 共产生了两个词组表供选择($k_{\text{p}}=5$ 时,$L_{\text{p}}=5$;$k_{\text{p}}=10$ 时,$L_{\text{p}}=4$;可以只使用部分层的词组)。在建构话题模型时,同样是依据不同的 k_t 共产生了两个话题模型供选择($k_t=5$ 时,$L_{\text{topic}}=4$;$k_t=10$ 时,$L_{\text{topic}}=4$;可以只使用部分层的话题)。每类取 50 幅,按照上述规则建立单词、词组、话题的多层次模型。在随后的实验中,各类另行选取 200 ~ 500 幅图像。在此数据集中,为每类单独产生自己的 SVM 分类器。为了在每类图像的识别中,选定 2.1 节 4)、5)所述的参数,针对每一种参数组合都建立一个 SVM,从而发现性能最好的参数。然后基于选定的最优参数,重复进行 5 次实验(即总共重复随机抽取了 5 次训练集),取分类准确率的均值。

测试得到的平均识别准确率如表 1 所示,结果显示,本文提出的核函数在场景分析中,相比于传统方法,明显具有更出色的性能。从中可以看到,在增加了多层词组和 ROI 上下文模型后,能够显著改善识别性能。这说明,在单词的不同层次的邻域内,分析其他单词的分布信息,有助于降低单词的歧义性,突出了不同类别图像间的差异。

表 1　在 Corel 测试集中各类的平均检测准确率

Tab. 1　Mean classification accuracies in the Corel dataset

/%

训练数据比例	10%	20%	33%	50%
Word histogram	40.7	47.8	52.5	57.4
PMK	50.2	59.0	65.1	69.9
本文核函数	57.8	64.6	69.3	74.1

此外,在参数选择中发现,词组的层数普遍不需要达到最大。一般在最下层划分出 5 000 ~ 10 000 个词组即可。图 5 展示了分类混淆矩阵(以 50% 的数据训练),其中元素 (i,j) 表示第 i 类图像的分类器将第 j 类图像判断为第 i 类的比例,灰度值的高低表示比例的高低。在总共 20 类场景中,本文的方法在 15 类上取得了优势,只有 3 类和对比方法性能相当:即大海、荒漠、天空等场景。这主要是因为分析这些类别,比较偏重于颜色、纹理等方面的信息。事

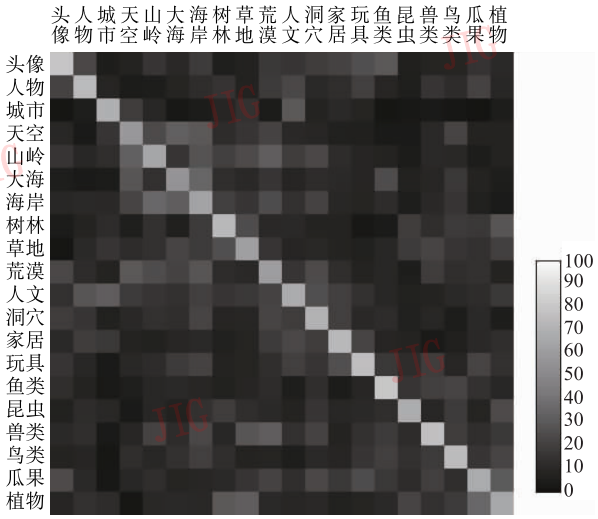


图 5 每类图像的分类器在 Corel 数据集各类中的分类结果(%)

Fig. 5 Classification results of the classifiers in the Corel dataset (%)

实上,这些特征的局部描述子可以按类似的方式融入到本文提出的核函数中。

2.3 对色情类和泳装类图像的分类

最后在 $\boldsymbol{D}_{\text{body}}$ 中进行一个单纯的两类区分的实验。在某些应用中,不一定需要把特定类别从所有图像中识别出来,而是将两种混杂在一起的图像区分开。例如,在色情图像的检测中,大量裸露部分身体(如泳装照片)的合法图片被误识为色情的。如果能在初步的检测结果中,进一步区分色情图像与非色情的人体图像,则能显著提高对色情图像的监控能力。

在实验中,两类图像各随机取出 20%,用于创建单词、词组、话题的多层次模型。余下的图像进行接下来的实验。在余下的图像中,每次随机选取同样数量的两类图像进行训练,并分别尝试使用 20% 和 50% 的图像训练 SVM。两种情况选择的参数均为:$k_{\text{p}}=5$, $L_{\text{p}}=4$; $k_{t}=5$, $L_{\text{topic}}=3$。显然由于两种人体图像内容接近,过细的话题分割导致过度的拟合。两组实验各自重复 5 次,取检测准确率的平均值。3 个算法的实验对比结果如表 2 所示。由于没有采用更丰富的颜色、纹理、边缘等方面的识别特征,所以性能并不突出。但实验结果表明,本文的多层次单词语义的挖掘,能在相当程度上区分两种极为相似的图像,能够捕捉较为细微的类间差别(例如对人体敏感器官的描述)。并且使用较少的图像进行训练,本文的核函数仍能取得约 70% 的识别准确率。在引入了词组和 ROI 话题后,对色情图像的识别能力显著提高。图 6 展示了第 3 层(25 个)话题的出现频率,显示两类图像在多数话题上分布比较接近,但仍存在少数话题能对两者进行显著区分。

表 2 在 $\boldsymbol{D}_{\text{body}}$ 中的分类准确率

Tab. 2 Classification accuracies in $\boldsymbol{D}_{\text{body}}$

/%

	色情	泳装
Word histogram (20% 训练)	56.9	45.6
PMK (20% 训练)	57.5	65.8
本文核函数 (20% 训练)	67.1	71.6
Word histogram (50% 训练)	67.9	59.0
PMK (50% 训练)	70.1	77.9
本文核函数 (50% 训练)	78.3	82.4

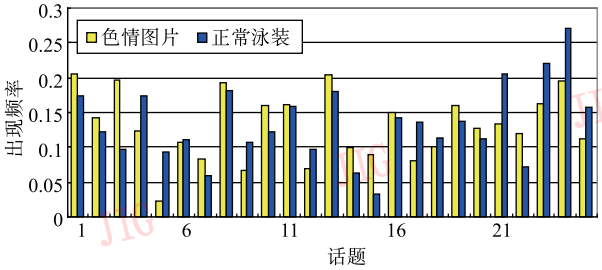


图 6 在 $\boldsymbol{D}_{\text{body}}$ 中第 3 层 ROI 话题的出现频率

Fig. 6 Frequencies of ROI topics of the third level in $\boldsymbol{D}_{\text{body}}$

2.4 计算复杂度的分析

最后比较一下计算复杂度。计算量比较大的一个是视觉单词的提取,此处 3 种方法的计算复杂度是一样的。相比于视觉单词的提取,提取词组和 ROI 的复杂度可以忽略不计。另一个计算量大的地方是在 SVM 分类器中,需要进行大量的核函数的计算。在每一次核函数的计算中,Word histogram 的复杂度是 $O(N_{\text{w}})$,其中,N_{w} 是单词表中单词的个数;PMK 的计算复杂度是 $O(m\cdot L_{\text{w}})$,其中 m 是图像中单词实例的个数;本文计算核函数的方法与 PMK 一脉相承,主要区别在于:1)每个单词实例都对应一个词组实例,使得图像单词包的元素数量翻了一倍;2)单词实例编码增加了 L_{topic} 位(1.6 节)。最终本文提出的核函数的计算复杂度为 $O(m\cdot(L_{\text{w}}+L_{\text{p}})\cdot L_{\text{topic}})$,但比起传统的复杂度为 $O(m^2)$ 的核函数,依然很有竞争力。实验使用 3.0 GHz 的 Intel 主机,在实际测试中,PMK 比较每两幅图像距离的平均耗时是 1.24×10^{-4} s,而本文的方法是 8.12×10^{-4} s。

3　结　论

为了提高基于视觉单词的核函数的图像识别能力,本文将视觉单词的上下文融入到核函数的设计中。为描述上下文,在多个层次上分析了视觉单词的共现规律:1)词组,描述单词的局部相邻关系;2)ROI 话题,与特定场景相关的区域的类别。在核函数的设计中,融合以下 3 个方面:1)多层视觉单词的直方图金字塔。2)多层词组的直方图金字塔。3)分析各单词(词组)的实例所对应的 ROI 话题,从而在识别中对上下文近似的单词(词组)给予重视。实验结果表明,在充分考察了单词的(多层次)上下文关系之后,相比于传统方法,新的核函数具有更好的性能,在多种测试数据中取得出色的分类结果。

参考文献(References)

[1] Vailaya A, Figueiredo M, Jain A, et al. Image classification for content-based indexing [J]. IEEE Transactions on Image Processing, 2001, 10(1): 117-130.

[2] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope [J]. International Journal of Computer Vision, 2001, 42(3): 145-175.

[3] Naphade M, Huang T. A probabilistic framework for semantic video indexing, filtering and retrieval [J]. IEEE Transaction on Multimedia, 2001, 3(1): 141-151.

[4] Serrano N, Savakis A, Luo J. A computationally efficient approach to indoor/outdoor scene classification [C] // Proceedings of the 16th International Conference on Pattern Recognition. Quebec, Canada, IEEE Press, 2002: 146-149.

[5] Vogel J, Schiele B. Natural scene retrieval based on a semantic modeling step [C] //Proceedings of International Conference on Image and Video Retrieval. Dublin, Ireland: Springer, 2004: 207-215.

[6] Mikolajczyk K, Leibe B, Schiele B. Local features for object class recognition [C] //Proceedings of the Tenth IEEE International Conference on Computer Vision. Beijing, China, IEEE Press, 2005: 1792-1799.

[7] Jiang W, Er G, Dai Q H. Similarity-based online feature selection in content-based image retrieval [J]. IEEE Transactions on Image Processing, 2006, 15(3): 702-712.

[8] Perronnin F, Dance C, Csurka G, et al. Adapted vocabularies for generic visual categorization [C]//European Conferece on Computer Vision. Graz, Austria, Springer, 2006: 464-475.

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

[10] Zhang R, Zhang Z. Effective image retrieval based on hidden concept discovery in image database [J]. IEEE Transactions on Image Processing, 2007, 16(2): 562-572.

[11] Wang G, Zhang Y, Li F. Using dependent regions for object categorization in a generative framework [C]//Proceedings of 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA, IEEE Press, 2006: 1597-1604.

[12] Nowozin S, Tsuda K, Uno T, et al. Weighted substructure mining for image analysis [C]//IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, USA: IEEE Press, 2007: 1-8.

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

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

[15] Agarwal A, Triggs B. Hyperfeatures: multilevel local coding for visual recognition [C] //European Conference on Computer Vision. Graz, Austria: Springer, 2006: 30-43.

[16] Shawe-Taylor J, Cristianini N. Kernel Methods for Pattern Analysis [M]. Cambridge: Cambridge University Press, 2004.

[17] Grauman K, Darrell T. Approximate correspondences in high dimensions [C]//Proceedings of the 20th Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada, MIT Press, 2006: 505-512.

[18] Wallraven C, Caputo B, Graf A. Recognition with local features: the kernel recipe [C] //Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France, Springer, 2003: 257-264.

[19] Lyu S. Mercer kernels for object recognition with local features [C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, California, USA, IEEE Press, 2005: 223-229.

[20] Sahbi H, Audibert J. Context-dependent kernel design for object matching and recognition [C] //IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, Alaska, USA: IEEE Press, 2008: 1-8.

[21] Wolf L, Shashua A. Learning over sets using kernel principal angles [J]. Journal of Machine Learning Research, 2003, 4(10): 913-931.

[22] Kondor R, Jebara T. A kernel between sets of vectors [C]// Proceedings of International Conference on Machine Learning. Washington, D C, USA, AAAI Press, 2003: 361-368.

[23] Moreno P, Ho P, Vasconcelos N. A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications [C]//Proceedings of the 17th Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia,

Canada, MIT Press, 2003: 1385-1392.

[24] Qi G, Hua X, Rui Y. A joint appearance-spatial distance for kernel-based image categorization [C]//IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, Alaska, USA, IEEE Press, 2008: 1-8.

[25] Grauman K, Darrell T. The pyramid match kernel: efficient learning with sets of features [J]. Journal of Machine Learning Research, 2007, 8(4): 725-760.

[26] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories [C]// Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE Press, 2006: 2169-2178.

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

[28] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features [C] //Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai, Hawaii, USA: IEEE Press, 2001: 511-518.

[29] Wang Yushi, Li Yuanning, Gao Wen. 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成人图像检测. 北京理工大学学报, 2008, 28(5): 410-413.]

[30] Liu B, Xia Y, Yu P S. Clustering through decision tree construction [C]//Proceedings of the 9th International Conference on Information and Knowledge Management. McLean, Virginia, USA, ACM, 2000: 20-29.

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

[32] Webb A. Statistical Pattern Recognition [M]. Second Edition. Hoboken, USA: John Wiley & Sons Incorporation, 2002: 106-113.

[33] Odone F, Barla A, Verri A. Building kernels from binary strings for image matching [J]. IEEE Transactions on Image Processing, 2005, 14(2): 169-180.