文章编号:1001-9081(2014)S2-0280-03

基于文本和内容的图像检索算法

顾 昕,张兴亮,王 超,陈思媛,方 正*
(厦门大学 机电工程系,福建 厦门 361005)
(* 通信作者电子邮箱 fangzheng38@163.com)

摘 要: 为了提高图像检索的效率,提出一种基于文本和内容的图像检索算法。该算法采用稠密的尺度不变特征转换(DSIFT)构造视觉单词的方式来描述图像内容,依据基于概率潜在语义分析(PLSA)模型的图像自动标注方法获取的视觉语义对查询图像进行初步检索,在此结果集上对筛选出的语义相关图像按内容相似度排序输出。在数据集 Corel1000 上的实验结果表明,该算法能够实现有效的图像检索,检索效率优于单一的基于内容的图像检索算法。

关键词: 图像检索;稠密的尺度不变特征转换;概率潜在语义分析;自动标注;视觉语义
中图分类号: TP391.41 **文献标志码:** A

Image retrieval algorithm based on text and content

GU Xin, ZHANG Xingliang, WANG Chao, CHEN Siyuan, FANG Zheng*
(*Department of Mechanical and Electrical Engineering, Xiamen University, Xiamen Fujian* 361005, *China*)

Abstract: In order to improve the efficiency of image retrieval, an image retrieval algorithm based on text and content was proposed. This method used Dense Scale-Invariant Feature Transform (DSIFT) feature to construct visual words to describe image content, roughly searched query image according to the visual semantics acquired by automatically annotating based on Probabilistic Latent Semantic Analysis (PLSA) model, then sorted the filtered semantically related images according to the similarity of content. Experimental results in Corel1000 database demonstrate that the proposed algorithm can retrieve images effectively and achieve better performance than the algorithm only based on image content.

Key words: image retrieval; Dense Scale-Invariant Feature Transform (DSIFT); Probabilistic Latent Semantic Analysis (PLSA); automatic annotation; visual semantic

0 引言

随着数字成像技术的飞速发展和互联网的普及,各种各样的图像数量正以惊人速度增长,日益丰富的图像资源使用户难以在浩如烟海的数据中找到其真正需要的信息,因而有效的图像检索技术成为近年来研究界关注的热点之一。

现有的图像检索技术主要分两种:基于文本的图像检索(Text-Based Image Retrieval, TBIR)[1]以及基于内容的图像检索(Content-Based Image Retrieval, CBIR)[2]。TBTR 主要依赖于图像的标注信息进行检索,但是面对数以万计的图像数据集,手工进行图像标注的代价太过昂贵,使得这种检索方案渐已不能满足现实的应用需要;CBIR 主要利用特征提取和高维索引技术进行图像检索,但是由于语义鸿沟[3]的存在,视觉特征相似的图像很可能在语义上是不相关的,这就使得很多情况下基于内容的图像检索结果难以满足用户的信息需求。单纯的 TBIR 和 CBIR 均有各自的缺陷,于是很多研究者开始研究融合这两种技术的图像检索方法或系统[4-7],这些工作都显著地提高了图像检索性能,然而它们在获取图像文本语义时多借助于人工的注释或是采用适当的算法提取 Web 中相关的文本信息,因此其应用受到了一定的限制。

本文提出了一种基于文本和内容的图像检索算法,该算法使用了基于概率潜在语义分析(Probablistic Latent Semantic Analysis,PLSA)模型的图像自动标注方法,有效解决了手工标注需耗费大量劳力的问题,图像的检索采取了先依文本信息粗检索后按内容相似度排序的方式,取得了较为满意的检索结果。

1 基于文本和内容的图像检索

1.1 图像的内容描述

为了分析图像的内容,本文采用尺度不变特征变换(Scale-Invariant Feature Transform,SIFT)算子[8]来描述图像的局部特征,它具有较强的图像缩放、旋转、仿射变换鲁棒性,而且还对视角变化、亮度变化、噪声可以保持一定程度的稳定性,能够很好地表达图像的细节特征。由于在图像检索中,不同类别图像的 SIFT 特征点的个数差异较大,不能够直接提取 SIFT 描述子,本文使用稠密 SIFT(Dense SIFT, DSIFT)[9]来描述图像的局部特征。与传统的 SIFT 算法相比,DSIFT 算法不需要高斯差分计算检测极值点,从而可以省去十分耗时的这一步;而且该描述子无需旋转标准化从而省去了旋转计算,只需设定适当网格进行运算即可提取该特征。

本文采用 DSIFT 构造视觉单词的方式来描述图像内容,其步骤如下:

1) 把每幅图像缩放至统一的大小,将所有图像分解为 $E \times E$ 的子块,间距 F 像素采样,计算采样点相对于子块窗口

收稿日期:2013-12-13;修回日期:2014-01-29。基金项目:中央高校基本科研业务专项资金资助项目(2013121018);福建省自然科学基金资助项目(2012J01413);大学生创新创业训练项目(DC2014009)。

作者简介:顾昕(1988-),男,江苏泰兴人,硕士研究生,主要研究方向:数字图像处理;张兴亮(1988-),男,安徽六安人,硕士研究生,主要研究方向:红外分析;王超(1988-),男,河南杞县人,硕士研究生,主要研究方向:生物医学工程;陈思媛(1979-),女,湖北武汉人,助理工程师,硕士,主要研究方向:机电工程;方正(1976-),男,湖北武汉人,副教授,博士,主要研究方向:生物医学工程、精密仪器。

的 SIFT 描述子;

2) 将描述子通过聚类生成字典,构造视觉单词;

3) 图像的每个子块根据字典标注为最接近的单词,统计整个图像的单词出现频率生成直方图,即得到图像的内容基于词袋(Bag Of Words, BOW) [10] 的描述。

1.2　PLSA 模型

PLSA 方法[11] 起源于自然语言处理研究,它通过计算文档中共现词的分布来分析文档语义。本文采用图像处理的语言来描述该模型。假设一组训练集包含 n 幅图像 $D = \{d_1, d_2, \cdots, d_n\}$,在上节图像的内容基于词袋描述的基础上,每幅图像的若干个局部区域可被包含 m 个视觉单词的词汇表 $W = \{w_1, w_2, \cdots, w_m\}$ 量化表示,因此训练图像的集合就可以由一个单词图像的互共现矩阵来表示,矩阵的每个元素 $n(d_i, w_j)$ 表示单词 w_j 在图像 d_i 中出现的次数,引入一个隐含 k 个变量的集合 $O = \{o_1, o_2, \cdots, o_k\}$,表示训练集的潜在语义集合。

PLSA 假设图像和视觉词汇之间及其包含的潜在语义都是条件独立的,对应的联合概率模型可由式(1) 来表示:

$$P(d_i, w_j) = P(d_i)P(w_j \mid d_i), P(w_j \mid d_i) = \sum_{r=1}^{k} P(o_r \mid d_i)P(w_j \mid o_r) \tag{1}$$

其中: $P(d_i)$ 表示图像 d_i 出现的概率,$P(w_j \mid o_r)$ 表示词汇在潜在语义上的概率分布,$P(o_r \mid d_i)$ 表示图像的潜在语义概率分布。

PLSA 算法中,潜在变量的估计使用的是最大期望(Expectation Maximum, EM) 算法,即通过最大化式(2) 所示的对数似然函数计算 PLSA 参数,从而拟合潜在语义模型。

$$L = \sum_{i=1}^{n}\sum_{j=1}^{m} n(d_i, w_j) \operatorname{lb} P(d_i, w_j) \tag{2}$$

该算法在 E 步骤中,通过对图像集模拟训练,计算 (d_i, w_j) 时潜在语义 o_r 的后验概率,具体的计算如式(3) :

$$P(o_r \mid d_i, w_j) = \frac{P(o_r)P(d_i \mid o_r)P(w_j \mid o_r)}{\sum_{l=1}^{k} P(o_l)P(d_i \mid o_l)P(w_j \mid o_l)} \tag{3}$$

在 M 步骤中,利用上一步的期望值来最大化当前的参数估计,其计算过程如式(4) ~(6) :

$$P(w_j \mid o_r) = \frac{\sum_{i=1}^{n} n(d_i, w_j)P(o_r \mid d_i, w_j)}{\sum_{ms=1}^{m}\sum_{i=1}^{n} n(d_i, w_{ms})P(o_r \mid d_i, w_{ms})} \tag{4}$$

$$P(o_r \mid d_i) = \frac{\sum_{j=1}^{m} n(d_i, w_j)P(o_r \mid d_i, w_j)}{\sum_{i=1}^{n}\sum_{ms=1}^{m} n(d_i, w_{ms})P(o_r \mid d_i, w_{ms})} \tag{5}$$

$$P(o_r) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} n(d_i, w_j)P(o_r \mid d_i, w_j)}{\sum_{i=1}^{n}\sum_{j=1}^{m} n(d_i, w_j)} \tag{6}$$

经过 E、M 步迭代,满足收敛条件时停止即可得到 $P(w_j \mid o_r)$ 和 $P(o_r \mid d_i)$ 的分布。

对于参数 $P(w \mid o)$ 和 $P(o \mid d)$,若已知其中一个分布 $P(w \mid o)$ (或 $P(o \mid d)$) ,则另一个分布 $P(o \mid d)$ (或 $P(w \mid o)$) 可以采用叠入算法[12] 计算得到,该算法在迭代过程中固定已知参数,不断更新未知参数,使得式(2) 中的似然函数最大。

1.3　图像的自动标注

为实现图像的自动标注,本文采用了一种自适应不对称学习方法[13] 来获取图像的视觉语义。将图像看作是一系列潜在主题的混合,假设训练集图像对应的文本标注词主题 t 的个数为 g,基于词袋的视觉词主题 s 的个数为 h,则图像混合的主题数 $k = g + h$。

本文使用自适应不对称学习方法进行图像自动标注的步骤如下:

1) 根据 PLSA 模型分别计算得到图像文本模态和视觉模态对应的主题分布 $P_w(t \mid d)$ 和 $P_v(s \mid d)$。

2) 计算每幅图像的视觉词分布熵 $H(v(d_i))$,根据经验公式(7) 得到每幅图像的文本模态和视觉模态对描述该图像贡献的百分比分别为 α_{wi} 和 α_{vi}:

$$\begin{cases} \alpha_{vi} = \begin{cases} 1, & H(v(d_i)) \leqslant 3 \\ \exp(3 - H(v(d_i))), & H(v(d_i)) > 3 \end{cases} \\ \alpha_{wi} = 1 - \alpha_{vi} \end{cases} \tag{7}$$

3) 根据式(8) 融合每幅图像的文本主题分布 $P_w(t \mid d_i)$ 和视觉主题分布 $P_v(s \mid d_i)$,得到混合的主题分布 $P(o \mid d_i)$:

$$P(o_r \mid d_i) = \begin{cases} \alpha_{wi} p_w(t_r \mid d_i), & r = 1, 2, \cdots, g \\ \alpha_{vi} p_v(s_{r-g} \mid d_i), & r = g+1, g+2, \cdots, g+h \end{cases} \tag{8}$$

4) 对混合的主题分布 $P(o \mid d_i)$ 使用叠入算法,学习得到文本词和视觉词在混合主题上的分布 $P(w \mid o)$ 和 $P(v \mid o)$,这两个参数对训练集外的图像也是有效的。

5) 对于给定的待标注图像 d_{new},计算其基于词袋的表示 $v(d_{new})$,再次使用叠入算法处理 $v(d_{new})$ 与学习得到的参数 $P(v \mid o)$,得到图像的主题分布 $P(o \mid d_{new})$,根据式(9) 计算得到概率 $P(w \mid d_{new})$,并对此进行排序选取具有最大后验概率的若干文本标注词作为 d_{new} 的标注。

$$P(w \mid d_{new}) = \sum_{r=1}^{k} P(w \mid o_r)P(o_r \mid d_{new}) \tag{9}$$

1.4　基于文本和内容的图像检索系统

基于文本和内容的图像检索算法的思路是: 对检索库图像(训练集) 分别进行文本和内容描述存入数据库中,如图 1 虚线以上所示; 在进行图像检索时,首先对待查询图像(测试集) 进行自动标注得到其文本信息并与库中文本比较初步检索筛选出语义相关图像集,然后根据其自动获取的视觉内容对语义相关图像进行相似性度量并排序,返回检索结果,其过程如图 1 虚线以下所示。

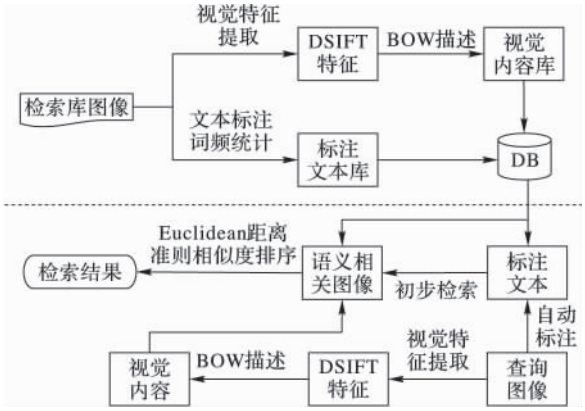

图 1　图像检索系统结构

2　实验结果与分析

为了评估基于文本和内容的图像检索算法的性能,本实验开发了一个原型系统,该系统能够实现有效的图像检索功

能。图像的文本库与内容库的建立以及 PLSA 模型的拟合过程均采用离线方式执行,图像的检索采用在线方式执行。

2.1 实验设计

本实验使用的数据集是 Corel1000,该图像库由 10 个不同类别的子集组成,包括: Arica、Beach、Buildings、Bus、Dinosaurs、Elephants,、Flowers、Horses、Mountains、Food,每个子集包含相同语义内容的图像 100 幅,每幅图像的标注词数是 5,图像集中的 1 000 幅图像共包括有 167 个不同的标注词。为了保证检索中每类图像数目相同,本文从该图像库的 10 个子集中分别随机选取 97 幅图像,合计 970 幅图像作为检索图像库(训练集),而其余 30 幅图像作为查询图像(测试集)。实验采用了 DSIFT 特征提取算法采集图像的局部特征矢量,每幅图像平均有 1 500 个左右的局部特征,通过 K-means 聚类算法将所有向量的特征聚到 1 000 个不同的类中,于是图像特征 $|V| = 1\,000$。在采用 PLSA 模型进行语义主题学习时,使用 45 个潜在主题学习视觉模态信息,使用 70 个潜在主题学习文本模态信息。图 2 为本实验的一个检索示例,其中左上角为查询图像,其余的 35 幅为检索结果图像,其相关性由左至右、由上至下逐渐减小。

图 2 一个图像检索示例

2.2 性能评价方法

查准率和查全率是信息检索中的标准评价方法,现已被广泛用于图像检索中。本文即采用 Precision-Recall 准则作为系统性能评价标准,定义为精度 $Precision = B/A$,召回率 $Recall = B/C$。其中: A 指检索返回的图像数,B 指检索结果中与查询图像相关的图像数,C 指检索图像库中与查询图像相关的图像总数。

2.3 实验结果及分析

在相同检索条件下,将本文算法与基于内容的图像检索算法进行比较,两种算法的 Precision-Recall 曲线如图 3 所示。

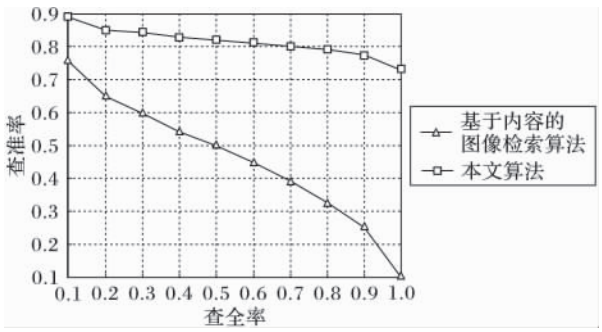

图 3 两种算法查准率-查全率比较

图 3 表明本文算法检索精度明显高于基于内容的图像检索算法精度,并且本文算法的精度随召回率的增加下降较慢,因此本文算法的检索性能更加稳定。

为了分析 2 种算法对不同类别图像的检索结果,表 1 展示了检索返回的图像数为 30 时,各个类别图像的平均检索精度。

表 1 返回图像数为 30 时 2 种算法各类检索精度 %

类别	基于内容的图像检索算法	本文算法
Arica	66.67	78.89
Beach	45.56	87.78
Buildings	38.89	82.22
Bus	93.33	97.78
Dinosaurs	98.89	100.00
Elephants,	68.89	87.78
Flowers	84.44	97.78
Horses	88.89	100.00
Mountains	17.78	46.67
Food	41.11	58.89

由表 1 可知,本文算法在大部分类别的检索精度都明显高于基于内容的图像检索算法精度,两种算法的平均精度分别为 64.44% 和 83.78%。

3 结语

本文提出一种基于文本和内容的图像检索算法,首先采用稠密 SIFT 构造视觉单词的方式来描述图像的内容; 然后基于 PLSA 模型使用自适应的不对称学习方法融合并学习视觉模态和文本模态信息实现图像的自动文本标注; 最后采取先依文本信息粗检索再按内容相似度排序的方式查询图像。该算法充分发挥了文本与内容之于图像检索各自的优势,实验结果表明本文算法能够实现较为有效的图像检索。

参考文献:

[1] BACH J R, FULLER C, GUPTA A, *et al*. Virage image search engine: an open framework for image management [C] // Proceedings of the SPIE 2670. Bellingham: SPIE, 1996: 76 - 87.

[2] DATTA R, JOSHI D, LI J, *et al*. Image retrieval: Ideas, influences, and trends of the new age [J]. ACM Computing Surveys, 2008, 40(2): Article No. 5.

[3] SMEULDERS A W M, WORRING M, SANTINI S, *et al*. Content-based image retrieval at the end of the early years [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(12): 1349 - 1380.

[4] ZHAO R, GROSKY W I. Narrowing the semantic gap-improved text-based Web document retrieval using visual features [J]. IEEE Transactions on Multimedia, 2002, 4(2): 189 - 200.

[5] JING F, LI M, ZHANG H J, *et al*. A unified framework for image retrieval using keyword and visual features [J]. IEEE Transactions on Image Processing, 2005, 14(7): 979 - 989.

[6] HE R, JIN H, TAO W, *et al*. Unifying keywords and visual features within one-step search for Web image retrieval [M] // Advances in Multimedia Information Processing - PCM 2006. Heidelberg: Springer Berlin, 2006: 527 - 536.

[7] 谢琳. 融合文本语义和视觉内容的 Web 人像图片检索[D]. 北京: 北京交通大学, 2008.

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

[9] LAZEBNIK S, SCHMID C, PONCE J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories [C] // CVPR2006: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2006, 2: 2169 - 2178.

(下转第 313 页)

每一帧图像用第 1 章的方法进行光标位置自动定位并代入标校数学模型得到当前帧标校值 K_i。若对某个样本的自动定位结果满足第 2.1 节所述的 DET 评价法原则说明定位成功,准确率 λ 为满足此条件的样本个数与总样本个数的比例,定位精度 $\mu = 1 - \max_i\{|K_i - K_t|\}/K_t$。

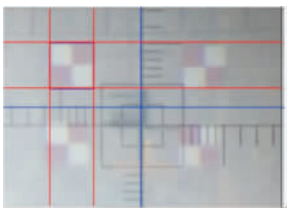

图 9　在计算机采集的图像中人工标定光标位置

本文实验中,选择了 3 个光照条件不同(晴天、阴天、雨天)的标校时刻,采用上文所述的方法得到 3 个标称值 K_{t1}、K_{t2}、K_{t3}。为采集若干帧图像进行实验,在 4 个光标均在视场范围内的条件下,随机指定天线方位和俯仰,每组角度取一幅图像及其倒镜图像作为实验样本。实验得出 $\lambda = 97.297\%$、$\mu = 98.667\%$,而 PCNN 方法的 $\lambda = 74.324\%$;进一步对上述实验样本可分别统计采用本文方法和采用人工方法进行标校的随机误差[14]曲线(图 10),可以看出由于受视差、主观因素的影响,人工标校结果的一致性差、精度低,最大随机误差达到 22.5762″(图 10(a)),本文方法的最大随机误差为 0.2292″(图 10(b)),进一步计算可得人工方法的定位精度 $\mu = 81.392\%$,明显低于本文方法。本文算法在主频 3 GHz、双核计算机平台上的平均时间复杂度为 1.43 s,远低于人工进行一次标校的耗时(10 min 以上)。

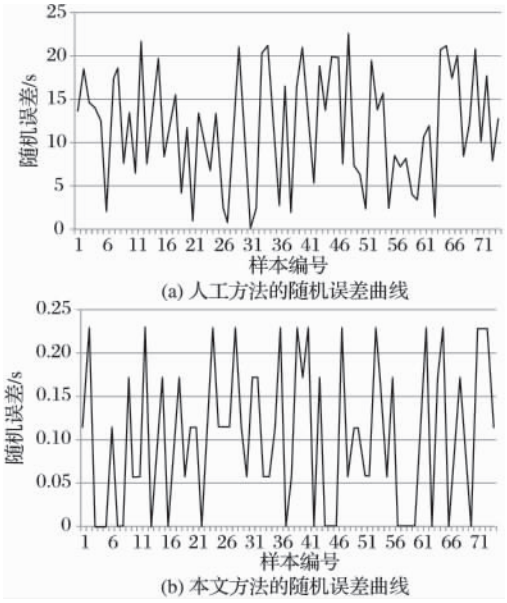

图 10　随机误差曲线

3　结语

本文立足于解决雷达光电误差参数人工标校方法存在的诸多局限性,从研究光标的特征出发,提出了一种新的计算机视觉方法实现光标的自动定位。与 PCNN 方法相比,本文方法综合利用了光标的局部颜色分布和方向梯度特征,对光照变化、刻度遮挡的影响表现出更强的鲁棒性。本文方法有效避免了人工标校中视差和主观因素影响,提高了标校的精度。本文方法耗时低,具备工程应用中可接受的时间复杂度,能够在任务窗口附近实施标校,保证了标校结果的即时性。下一步的研究方向是: 解决天线存在抖动情况下光标的自动定位方法。

参考文献:

[1] 康德永, 傅敏辉, 赵文华, 等. 基于恒星测量的船载雷达轴系误差修正参数动态标定[J]. 电讯技术, 2013, 35(7): 141 – 145.
[2] 杨斌峰. 地面测控雷达角度标校技术[J]. 现代电子技术, 2005, 28(17): 47 – 49.
[3] 王辉, 苏茂君, 段炳玺, 等. 航天测控天线光电标参数自动标定系统[J]. 微计算机信息, 2009, 25(7): 270 – 272.
[4] BASU M. Gaussian-based edge-detection methods—a survey[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, 2002, 141(10): 322 – 324.
[5] WEI D, ZHAO Y, CHENG R, *et al*. An enhanced histogram of oriented gradient for pedestrian detection[C]// Proceedings of 2013 Fourth International Conference on Intelligent Control and Information Processing. Piscataway: IEEE, 2013: 459 – 463.
[6] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]// Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2005, 1: 886 – 893.
[7] YLIOINAS J, HADID A, GUO Y, *et al*. Efficient image appearance description using dense sampling based local binary patterns [C]// ACCV 2012, LNCS 7726. Heidelberg: Springer Berlin, 2013: 375 – 388.
[8] DALAL N. Finding people in images and videos[D]. Grenoble: Institute National Polytechnique De Grenoble, 2006: 30 – 33.
[9] COMANICIU D. An algorithm for data-driven bandwidth selection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(2): 281 – 288.
[10] 陈星星, 张荣. 基于多尺度相位特征的图像检索方法[J]. 电子与信息学报, 2009, 31(5): 1193 – 1196.
[11] FELZENSZWALB P. Learning models for object recognition[C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, DC: IEEE Computer Society, 2001: 1056 – 1062.
[12] WU B, NEVATIA R. Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors [J]. International Journal of Computer Vision, 2007, 31(2): 10 – 12.
[13] GORGUCCI E, BECHINI R, BALDINI L, *et al*. The influence of antenna radome on weather radar calibration and its real-time assessment[J]. Journal of Atmospheric and Oceanic Technology, 2013, 30(4): 676 – 689.
[14] WU W C-H, YEH M-Y, PEI J. Random error reduction in similarity search on time series: a statistical approach[C]// Proceedings of the 2012 IEEE 28th International Conference on Data Engineering. Piscataway: IEEE, 2012: 858 – 869.

(上接第 282 页)

[10] SIVIC J, ZISSERMAN A. Video Google: a text retrieval approach to object matching in videos [C] // ICCV2003: Proceedings of the Ninth IEEE International Conference on Computer Vision. Washington, DC: IEEE Computer Society, 2003: 1470 – 1477.
[11] BOSCH A, ZISSERMAN A, MUOZ X. Scene classification using a hybrid generative/discriminative approach [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(4): 712 – 727.
[12] HOFMANN T. Unsupervised learning by probabilistic latent semantic analysis [J]. Machine Learning, 2001, 42(1/2): 177 – 196.
[13] 李志欣, 施智平, 李志清, 等. 融合语义主题的图像自动标注[J]. 软件学报, 2011, 22(4): 802 – 812.