软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2017,28(7):1847−1861 [doi: 10.13328/j.cnki.jos.005069]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

# 三角形约束下的词袋模型图像分类方法*

汪荣贵, 丁 凯, 杨 娟, 薛丽霞, 张清杨

(合肥工业大学 计算机与信息学院,安徽 合肥 230009)

通讯作者: 杨娟, E-mail: yangjuan@hfut.edu.cn

摘 要: 视觉词袋模型广泛地应用于图像分类与图像检索等领域.在传统词袋模型中,视觉单词统计方法忽略了视觉词之间的空间信息以及分类对象形状信息,导致图像特征表示区分能力不足.提出了一种改进的视觉词袋方法,结合显著区域提取和视觉单词拓扑结构,不仅能够产生更具代表性的视觉单词,而且能够在一定程度上避免复杂背景信息和位置变化带来的干扰.首先,通过对训练图像进行显著区域提取,在得到的显著区域上构建视觉词袋模型.其次,为了更精确地描述图像的特征,抵抗多变的位置和背景信息的影响,该方法采用视觉单词拓扑结构策略和三角剖分方法,融入全局信息和局部信息.通过仿真实验,并与传统的词袋模型及其他模型进行比较,结果表明,该方法获得了更高的分类准确率.

关键词: 词袋模型;显著区域;空间拓扑结构;三角剖分;图像分类

中图法分类号: TP391

中文引用格式: 汪荣贵,丁凯,杨娟,薛丽霞,张清杨.三角形约束下的词袋模型图像分类方法.软件学报,2017,28(7):1847−1861. http://www.jos.org.cn/1000-9825/5069.htm

英文引用格式: Wang RG, Ding K, Yang J, Xue LX, Zhang QY. Image classification based on bag of visual words model with triangle constraint. Ruan Jian Xue Bao/Journal of Software, 2017,28(7):1847−1861 (in Chinese). http://www.jos.org.cn/1000-9825/5069.htm

## Image Classification Based on Bag of Visual Words Model with Triangle Constraint

WANG Rong-Gui, DING Kai, YANG Juan, XUE Li-Xia, ZHANG Qing-Yang

(School of Computer and Information, Hefei University of Technology, Hefei 230009, China)

**Abstract**: Bag of visual words model is widely used in image classification and image retrieval. In traditional bag of words model, the statistical method of visual words ignores the spatial information and object shape information, resulting lack of ability to distinguish between image features. In this paper, an improved bag of words method is proposed to combine with salient region extraction and visual words topological structure so that it is can not only produce more representative visual words to certain extent, but also avoid the disturbance of complex background information and position change. First of all, the significant areas of training image are extracted and the bag of visual words model is built on the significant area. Secondly, in order to describe the characteristics of the image more accurately and resist the changing location and the influence of background information, the strategies of visual words topological structure and Delaunay triangulation method are utilized and integrated into the global information and local information. Simulation experiments are performed to compare with the traditional bag of words and other models, the results demonstrate that the proposed method obtained a higher classification accuracy.

**Key words**: BoW (bag of words); saliency; topological structure; Delaunay triangulation; image classification

随着互联网和多媒体技术的飞速发展,数字图像、视频等多媒体信息的数量呈爆炸式增长.为了准确、高

* 基金项目: 安徽省自然科学基金(J2014AKZR0055); 中国博士后科学基金(2014M561817); 安徽省科技攻关重大项目(1301041025)

Foundation item: Natural Science Foundation of Anhui Province, China (J2014AKZR0055); China Postdoctoral Science Foundation (2014M561817); Special Fund for Key Program of Science and Technology of Anhui Province, China (1301041025)

收稿时间: 2015-12-02; 修改时间: 2016-03-03; 采用时间: 2016-03-22; jos 在线出版时间: 2016-05-03

CNKI 网络优先出版: 2016-05-04 08:44:17, http://www.cnki.net/kcms/detail/11.2560.TP.20160504.0844.009.html

效地组织、管理和检索图像,需要计算机准确地理解图像内容.图像分类是解决图像理解问题的重要途径,对多媒体检索技术的发展有重要的推动作用.作为图像处理的基础,图像特征表示是该领域的重要研究内容,其性能直接关系到图像分类、对象识别等问题的处理结果.

Sivic 等人[1]提出的视觉词袋模型广泛应用于物体和场景分类中,并且展现出优秀的性能.该模型借鉴了文本分析领域中的相关概念和思想,将词袋模型(bag of words,简称 BoW)引入到了计算机视觉领域,取得了巨大成功.它无需分析图像中的具体目标组成,而是应用图像的整体统计信息,将量化后的图像底层特征视为视觉单词,通过图像的视觉单词分布来表达图像内容.

传统的 BoW 模型在图像分类与检索等领域都取得了较好的效果,但该方法假设视觉单词之间相互独立,使用的视觉单词统计方法忽略了视觉单词的空间位置关系和形状等信息,在表示图像信息方面还存在局限性.近年来以 BoW 为基础产生了多种改进的图像特征表示方法.Lazebnik 等人[2]提出空间金字塔匹配方法(spatial pyramid matching,简称 SPM),该方法将图像均匀划分成不同粒度的网格形成图像层次,并计算各网格内视觉单词出现频率的直方图统计,将不同层次中各个网格的统计向量进行叠加形成高维特征表示.SPM 方法在直方图相交函数的基础上使用了 Grauman 等人[3]提出的金字塔匹配核函数进行图像分类,即对不同层次中匹配的视觉单词进行加权.Lu 等人[4]使用马尔可夫网络模型捕捉视觉单词之间的垂直和水平空间关系,形成表示图像的二阶马尔可夫模型,利用图像上下文、一致性和多样性,并使用标签传播方法,建立了基于图的半监督交互式图像分类构架.江悦等人[5]引入图像上下文信息,通过统计图像中的相异词对和视觉词群形成图像特征.刘硕研等人[6]通过计算视觉单词之间的语义共生概率,同时引入马尔可夫随机场理论,提高了视觉单词的语义准确性.近年来,除了上述图像表示方法外,还有众多加入视觉短语的图像特征表示方法.Li 等人[7]基于 BoW 模型提出了上下文词袋(contextual bag-of-word,简称 CBoW)模型,通过计算视觉单词在图像类别上分布的 KL 分布距离形成视觉词组,并加入视觉单词之间的语义关系,同时也应用 *N*-gram 方法将视觉单词组成视觉短语进行图像表示.Zhang 等人[8]提出了描述性视觉单词(descriptive visual word,简称 DVW)和描述性视觉短语(descriptive visual phrase,简称 DVP)方法,该方法以某个视觉单词为中心取相应半径范围内的视觉单词组成视觉短语.Zheng 等人[9]通过统计近邻视觉单词共同出现的频率来提取视觉短语,并将其应用于图像检索.Sivic 等人[10]使用 *k* 近邻方法(*k*-nearest neighbor,简称 *k*-NN)形成视觉短语用于图像检索.Wang 等人[11]使用 meanshift 方法聚类视觉单词,并通过 FP-Growth 算法挖掘有意义的空间视觉单词组合,用于图像分类.Zhou 等人[12]研究部分重复的网络图像检索,利用空间编码方法对图像中局部特征的空间关系进行编码,有效地发现并解决了图像之间局部特征的错误匹配问题.Tian 等人[13]结合用户定义的感兴趣区域和特征空间分布来进行基于内容的图像检索,通过相关性反馈结合了更多的图像空间信息.虽然当前的方法取得了较好的效果,但容易受到图像特征多变的位置和背景信息的干扰,且在结合视觉单词空间信息方面,所生成的视觉单词特征表达能力和区分度不足.因此,本文通过结合显著区域提取和视觉单词的空间位置关系以及形状信息,可以使图像特征表示更有区分性和代表性,从而提高分类性能.

研究发现,人眼视觉在对物体进行识别时,往往只对某一显著区域感兴趣[14],可以通过提取图像的显著区域并且只对显著区域进行处理来避免多余的计算,避免了背景中杂乱信息的干扰.显著性区域是图像中最能引起用户注意、最能表现图像内容的区域.由于人类视觉系统和注意机制的共性,使得图像中某些区域总能显著地吸引人视觉的注意,这些区域往往含有丰富的信息.通过图像的某些底层特征提取图像中的显著性区域,符合人的主观评价,是视觉上重要的区域.近年来出现了多种显著性区域提取方法,Itti 等人[15]在快速场景分析工作中提出了基于显著性的视觉注意模型,其主要受早期灵长类动物视觉系统行为和神经元结构的启发,利用多尺度图像特征组合成单一的视觉显著图.Harel 等人[14]提出一种自底向上的视觉显著性模型,利用生物启发的过滤器提取图像的特征向量,不同尺度的特征图生成激活图,基于生物模型和计算能力实现归一化以突出图像显著性区域.Hou 等人[16]通过分析输入图像的对数谱,在谱域中提取图像的谱残差,提出一种快速的方法在空间域中构建对应的显著性图.Ma 等人[17]提出一种图像注意力分析框架,利用模糊增长算法来模拟人的感知过程,通过研究人类感知的对比过程,得到一种基于对比的显著图.由于当前的显著性区域提取方法具有低分辨率、边界

模糊和计算复杂度高等问题,一些方法在目标对象边界处产生较高的显著值而不是均匀地覆盖整个对象.本文中我们使用显著性区域提取方法[18]来估计像素中央-周围颜色与亮度的对比度,具有均匀突出显著性区域、边界轮廓分明、分辨率高和计算效率高等优点.

本文在视觉词袋模型的基础上,结合视觉显著度,引入图像视觉单词拓扑结构相似性约束[19−21]进行图像分类.基于图像显著区域,首先,对视觉单词的位置关系进行全局约束,通过近似最近邻搜索可以得到匹配的视觉单词集[22−24].其次,采用 Delaunay 三角剖分方法,构建 DT 三角网络,对视觉单词局部位置关系进一步约束,应用于图像分类中.通过引入全局位置信息和局部关系信息改进传统的视觉词袋模型,可以提高图像分类的效率.

本文第 1 节介绍相关的准备工作,包括图像分类的基本方法以及显著性区域的提取.第 2 节详细描述融合空间信息的改进 BoW 方法,包括全局限制与局部约束.第 3 节主要在 Caltech101、COIL_100 和 Caltech256 这 3 个公开的图像集中对本文方法进行验证,并与其他方法进行比较,该方法能够有效地提高图像分类效果.最后总结全文,并对未来值得关注的改进方向进行初步探讨.

## 1 相关工作

### 1.1 基于 BoW 的图像分类方法

词袋模型最初应用于文本处理领域,用来对文档进行分类和识别.词袋模型因为其简单、有效的优点而得到了广泛的应用.其基本原理为将文档看作是无序的关键词的集合,通过统计每个关键词在单个文档中出现的频率来对文档进行向量表示,从而进行分类.对应到图像分类技术中,应用词袋模型进行图像分类包括以下 3 个步骤:特征提取和描述,视觉词典构造,训练分类器进行分类和识别.

应用词袋模型进行图像分类的实现过程可以描述如下.

步骤 1. 特征提取和描述.给定一幅图像,特征提取和描述环节的主要任务是从图像中抽取具有代表性的全局和局部特征作为对该图像的描述.传统方法中大多数采用 SIFT 描述子来实现这一过程,每一个特征用 128 维的向量来加以表示.

步骤 2. 生成视觉词典.采用 $k$-means 聚类方法对由步骤 1 得到的大量特征点进行聚类,将聚类的中心定义为视觉单词,所有的视觉单词进行组合则构成视觉词典,视觉词典的大小即为视觉单词的个数.传统方法中一般采用直接表示的方法,将图像表示成基于视觉单词的统计直方图.

步骤 3. 训练分类器进行分类.SVM 是较常用且实现较为简单的分类器之一.其核心思想是通过寻找最优超平面对空间中的不同特征集进行划分.该分类器最初只应用于两类的分类问题中,现在已逐渐应用于解决多类高维分类问题,并取得了良好的效果.它可以描述为如下优化问题:

$$\min_{\omega,\xi}\left\{\frac{1}{2}\|\omega\|^2+c\sum_{i-1}^{n}\xi_i\right\}.$$

约束条件为 $y_i(\omega\cdot x_i-b)\geq 1-\xi_i,\xi_i\geq 0,i-1,...,n$.其中,$\omega$为与超平面垂直的向量,$c$ 为惩罚因子,$\xi_i$ 为稀疏变量,$y_i$ 的值为 1 或者−1,表示数据点所属类别.

本文主要引入显著性区域,结合视觉单词的空间位置信息来改进传统视觉词袋模型,并采用 SVM 分类器实现图像的分类.

### 1.2 显著区域提取

本文选择高斯差分滤波器作为带通滤波器[18].由于该滤波器能够有效地逼近拉普拉斯高斯滤波器,因此被广泛应用于边缘检测、感兴趣点检测和显著性区域检测.式(1)给出高斯差分滤波器,其中,$\sigma_1,\sigma_2(\sigma_1>\sigma_2)$是高斯滤波器的标准方差:

$$DoG(x,y)=\frac{1}{2\pi}\left[\frac{1}{\sigma_1^2}\mathrm{e}^{-\frac{(x^2+y^2)}{2\sigma_1^2}}-\frac{1}{\sigma_2^2}\mathrm{e}^{-\frac{(x^2+y^2)}{2\sigma_2^2}}\right]=G(x,y,\sigma_1)-G(x,y,\sigma_2) \tag{1}$$

一个高斯差分滤波器是简单的带通滤波器,它的通频带宽度由$\sigma_1$:$\sigma_2$ 控制.如果定义$\sigma_1=\rho\sigma$和$\sigma_2=\sigma$,那么$\rho=\sigma_1/\sigma_2$,若将多个高斯差分滤波组合,那么这些具有标准方差的高斯差分滤波器的总和表述如下:

$$\sum_{n=0}^{N-1}G(x,y,\rho^{n+1}\sigma)-G(x,y,\rho^{n}\sigma)=G(x,y,\sigma\rho^{N})-G(x,y,\sigma) \tag{2}$$

其中,$N$ 为正整数,上式的结果实质上就是两个高斯值的差,它们的标准方差的比例是 $R=\rho^N$.如果假设通过$\sigma_1$ 和$\sigma_2$的变化来确保$\rho$为常数 1.6(为了检测边缘的需要),那么需要增加一些边缘检测算子在不同的图像尺度上的输出,这样就保证了整个显著性区域都可以得到突出,而不只是显著性区域的边缘或者其中心部分得到突出.

在计算显著图时,需要选择适当的$\sigma_1$ 和$\sigma_2$ 来确保带通滤波器保留所需要的原始图像空间频率.给定足够长的滤波器和$\sigma_1$ 与$\sigma_2$ 之间足够大的差分,带通滤波器通频带就可以近似看成来自两个连续的高斯滤波器.由于$\sigma_1>\sigma_2$,那么,$\omega_{lc}$的大小由$\sigma_1$控制,$\omega_{wc}$的大小由$\sigma_2$控制.在实际应用中,这些滤波器的长度不可能足够长,虽然实现很简单,但是近似就不够精确.

为了在标准方差下实现大的比例值,$\sigma_1$ 被设置成无穷大,为了消除噪声和纹理的高频信息以及计算上方便,使用小的高斯核,这些小的高斯核的二项式滤波器能够很好地拟合离散的高斯值.

对于 $W\times H$ 大小的图像 $I$,如下式来计算其显著图 $S$:

$$S(x,y)=\left|I_u-I_{\omega_{hc}}(x,y)\right| \tag{3}$$

其中,$I_u$ 是图像 $I$ 的算术平均灰度值,$I_{\omega_{hc}}$ 是为了消除纹理细节和噪声将该图像经过高斯模糊后的值.因为对这两者之间的差值大小感兴趣,因此就用绝对值这个范数来表示,这样计算较为方便.由于对图像不需要下采样,所以可以直接得到完整分辨率的显著图.

本文采用自适应阈值 $Th$,该值设定为图像平均显著值的两倍,如下式所示:

$$Th=\frac{2}{W\times H}\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}S(x,y) \tag{4}$$

其中,$W$ 和 $H$ 分别是显著图的宽和高,$S(x,y)$是坐标为$(x,y)$的像素点对应的显著值.

## 2 改进的视觉词袋模型

### 2.1 定位图像显著区域

将图像分成$m\times n$块,使用上节的显著区域提取方法 $S(x,y)=\left|I_u-I_{\omega_{hc}}(x,y)\right|$统计每块中的显著度,将每块区域内的显著值记录在一个 $m\times n$ 矩阵中,如图 1 所示.如果分块内平均显著值≥$Th$,认为该分块区域为图像显著子区域,其中,$Th$ 为判断分块内显著值的阈值.最后输出每个显著子区域的位置,得到图像的显著区域.

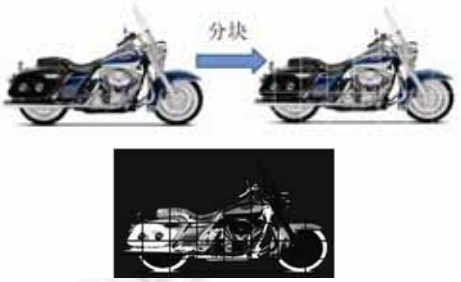



Fig.1 Salient region extraction based on image blocks

图 1 基于图像分块的显著区域提取

### 2.2 特征提取与图像表示

在底层特征提取阶段,通过提取大量的 SIFT 特征,最大限度地对图像进行底层描述,防止丢失过多的有用信息,这些底层描述中的信息主要靠后面的特征编码和特征汇聚得到抽象和简并.通过观察图像的 SIFT 特征可

以看到,虽然提取的 SIFT 特征并不完全相同,但在相同姿态下,所提取的大部分特征的相对位置是大体一样的,相同姿态下不同物体的SIFT特征维持了很好的几何结构.采用*K*-means聚类算法,对图像的局部特征进行聚类,所有视觉单词形成一个视觉词典,每个聚类中心被看作是词典中的一个视觉单词,聚类中心个数即为词典大小.本文中一幅图像 $I$ 表示为视觉单词 $w$ 的集合,每个单词对应的有其在图像中的位置 $r\cdot I\{(w_1,r_1),(w_2,r_2),\ldots,(w_m,r_m))\}$.其中,位置 $r$ 可以由兴趣点和区域检测子检测到,单词通过局部特征聚类得到.

### 2.3 高层特征的核函数

我们定义单个单词所表示的特征为一阶特征,两个单词、3 个单词表示的特征分别为二阶特征和三阶特征[25,26],依次类推.视觉单词的不同空间分布代表不同的特征,这里采用$\Sigma$为视觉单词词汇表,$n$ 阶特征 $f_n$ 为从$\Sigma$中选取的 $n$ 个具有特定空间分布的单词.核函数可表示为

$$K(x,y)=\sum_{n=1}^{\infty}w_nK_n(x,y) \tag{5}$$

其中,权值 $w_n=u^{1-n}$,$u$ 在 0 和 1 之间.当 $u$ 接近 0 时,我们将更大的权重赋予高阶特征.

$$K_n(x,y)=\langle\phi_n(x),\phi_n(y)\rangle=\sum\nolimits_{f_n}\langle\phi_{f_n}(x),\phi_{f_n}(y)\rangle \tag{6}$$

$\phi_n(x)$表示 $n$ 阶特征,$\phi_{f_n}(x)$ 表示 $n$ 阶特征 $f_n$ 在图像中所出现的次数.$n$ 阶特征核函数 $K_n(x,y)$则表示该 $n$ 阶特征所匹配的次数.

### 2.4 全局拓扑约束

#### 2.4.1 全局拓扑描述

接下来我们引入图像拓扑结构[20]这一概念,所谓图像的拓扑结构,是指组成图像的基本要素及其相互关系,基本要素主要包括图像中的顶点、边界线、区域及其灰度等.

图像分类过程中视觉单词在两幅相似图像之间的分布具有拓扑结构的稳定性,满足如下拓扑相似性约束条件.

1) 角度顺序约束.任意单词邻域各单词与其连线所形成夹角的顺序基本相同.

2) 邻域约束.若参考图像中单词 $W_1^1$ 是单词 $W_2^1$ 的邻域,则待匹配图像中匹配单词 $W_1^2$ 仍然为单词 $W_2^2$ 的邻域.参考图像中单词 $W_1^1$ 前 $N$ 个最近相邻单词,在待匹配图像中匹配点应属于前 $2N$ 个最近相邻单词之内.

3) 三角形位置约束.三角形与点的位置关系不变,参考图像中点 $W_{i,4}^1$ 位于 $\Delta W_{i,1}^1W_{i,2}^1W_{i,3}^1$ 之内/外,则待匹配图像中匹配点 $W_{i,4}^2$ 仍应位于 $\Delta W_{i,1}^2W_{i,2}^2W_{i,3}^2$ 之内/外.

一幅图像由 $n$ 个视觉单词组成,即 $n$ 个局部区域特征.我们可以获取到所有视觉单词的空间位置和视觉单词种类.这些信息可以被记录为一个配置矩阵:

$$M_{ij}=\left(1+\frac{\mathrm{d}w_{ij}^2}{\sigma_{\mathrm{d}w}^2}\right)\frac{\mathrm{d}_{ij}^2}{\sigma_d^2},i,j=1,\ldots,n \tag{7}$$

其中,$\mathrm{d}w_{ij}^2=\left\|\vec{f}_i-\vec{f}_j\right\|$ 表示视觉单词 $i$ 和视觉单词 $j$ 特征向量之差的平方,$\vec{f}_i$ 和 $\vec{f}_j$ 为它们的特征向量,$\sigma_{\mathrm{d}w}^2$ 为视觉单词所表示的特征向量差值的平方的平均值.$\mathrm{d}_{ij}^2=\left\|x_i-x_j\right\|^2$ 表示视觉单词 $i$ 和视觉单词 $j$ 所在方格中心点的欧式距离的平方.$\sigma_d^2$ 是所有包含视觉单词的方格中心平均欧氏距离的平方.$M$ 具有尺度、形变和旋转不变性,矩阵 $M$ 的第 $i$ 行代表视觉单词 $i$ 与其他视觉单词之间的相对距离.我们也可以称第 $i$ 行是视觉单词 $i$ 的配置向量.矩阵 $M$ 是对称的,如果特征向量之差 $\mathrm{d}w_{ij}$ 很小,那么 $M_{ij}$ 主要取决于它们的空间距离.反之,特征差别大,$M_{ij}$ 主要取决于特征向量差.

特征分解揭示了矩阵的子空间结构,矩阵 $M$ 的特征分解如下:

$$M=Q\Lambda Q^{-1},$$

其中,$Q$ 是包含特征向量的标准正交矩阵,$\Lambda$是对角矩阵,对角元素值为降序的特征值.$M^{-1}=M\times Q=Q\times\Lambda\times Q^{-1}\times Q$,特

征向量可表示为 $\vec{r}_1..\vec{r}_n$.

$$Q=\begin{pmatrix}\vec{r}_1\\ \vec{r}_2\\ \vdots\\ \vec{r}_n\end{pmatrix}.$$

在图像分类中,两幅相似的图像,它们的特征空间表示应该是相似的,如图 2 所示,全局的限制要求特征空间的配置矩阵相似.图像之间的相似性匹配就是要找到一个有序的配置向量使得特征配置矩阵的相似度最大化.

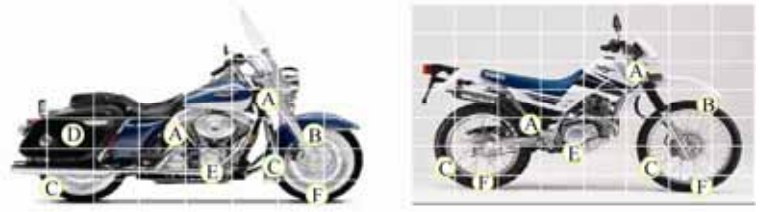



Fig.2 Spatial distribution of visual words between similar images

图 2 相似图像的视觉单词空间分布

2.4.2 相似度计算

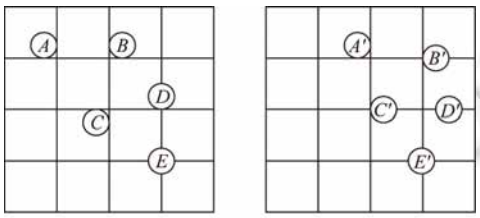
假设有两幅图像 $I_1$ 和 $I_2$,$I_1$ 中有 $m$ 个视觉单词,$I_2$ 中有 $n$ 个视觉单词.通过特征分解,我们可以得到每个视觉单词的子空间表示 $\vec{r}_i^{(1)}, i=1,\ldots,m$ 和 $\vec{r}_j^{(2)}, j=1,\ldots,n$. 对于 $I_1$ 中的视觉单词,其表示的是一个 $n$ 维向量,图像 $I_2$ 中的视觉单词,是一个 $m$ 维的向量.现在令 $L$=min($m$,$n$),选取前 $L$ 个向量元素截断 $\vec{r}$ ,为了使向量的每个元素对于向量差异的影响相同,可以对向量进行特征值归一化.

为了进行匹配,需要找到两个视觉单词 $\vec{r}$ 之间的最小化距离,计算最小化距离公式如下:

$$D_{ij}=\left\|\vec{r}_i^{(1)}-\vec{r}_j^{(2)}\right\| \tag{8}$$

图像 $I_1$ 中的视觉单词 $i$ 和图像 $I_2$ 中的视觉单词 $j$ 只有当 $D_{ij}$ 最小时才匹配.经过全局拓扑约束后,可以得到匹配的视觉单词集.

如图 3(a)和图 3(b)所示,从(a)到(b),经过了平移、旋转、尺度变换和变形.区域($B$,$C$,$D$)和($B'$,$C'$,$D'$)的形状相似,$E$ 和 $E'$是一样的.对于图 3(a)和图 3(b),我们可以得到归一化的特征空间配置矩阵,进行相似度的计算.




(a) (b)

Fig.3 Similar images with similar global topological constraints

图 3 相似图像具有相似的全局拓扑约束

### 2.5 局部位置约束

2.5.1 局部位置描述

在计算几何中,对于一个散射点模型,其相似关系可以通过 Voronoi 图表达.实际上,一个 Voronoi 图记录了所有邻近点集的信息.因为 Delaunay 三角剖分和 Voronoi 图互为偶图,也就是一一对应的,因此也包含了相同的信息.本文将 Delaunay 三角剖分[27~29]引入视觉词袋模型,以此来结合视觉单词之间的局部空间位置关系.

在视觉单词构建完成的基础上,对一幅图像的视觉单词集 $S$ 进行三角剖分,端点在 $S$ 中,而且只相交于端点,将 $S$ 的凸包分割成多个三角形.要满足 Delaunay 三角剖分的定义,必须符合两个重要的准则.

(1) 空圆特性:Delaunay 三角网是唯一的(任意 4 点不能共圆),在 Delaunay 三角形网中任一三角形的外接圆范围内不会有其他点存在.

(2) 最大化最小角特性:在散点集可能形成的三角剖分中,Delaunay 三角剖分所形成的三角形的最小角最大.

Delaulay 三角剖分具备如下优异特性.

(1) 最接近:以最近邻的 3 点形成三角形,且各线段(三角形的边)皆不相交.

(2) 唯一性:不论从区域何处开始构建,最终都将得到一致的结果.

(3) 最优性:任意两个相邻三角形形成的凸四边形的对角线如果可以互换的话,那么两个三角形 6 个内角中最小的角度不会变大.

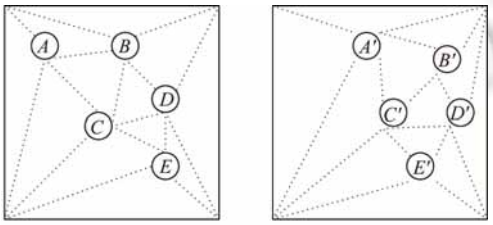
(4) 最规则:若将三角网中的每个三角形的最小角进行升序排列,则 Delaunay 三角网的排列得到的数值最大.

(5) 区域性:新增、删除、移动某一个顶点时只会影响临近的三角形.

(6) 具有凸多边形的外壳:三角网最外层的边界形成一个凸多边形的外壳.

鉴于上述优点,Delaunay 三角剖分很适合描述点集的局部位置关系,对于 $n$ 个点,计算时间复杂度为 $O(n\log n)$.通过全局的位置限制,得到匹配的视觉单词集,基于这些视觉单词,我们可以完成 Delaunay 三角剖分.通过三角剖分来匹配邻近的点,所有三角形内未匹配的点只能在其他图像的三角形内寻求匹配.只要有新的匹配区域,就可以继续进行 Delaunay 三角剖分,在这一步,其他的局部限制方法也可以加进来.

将计算几何的三角剖分方法引入视觉词袋模型,通过对视觉单词集进行 Delaunay 三角剖分,把图像空间上位置相近的视觉单词按照一定规则相连,得到三角形网络.如图 4 所示,对视觉单词进行 Delaunay 三角剖分(包括图像的 4 个顶点),将图像分成了多个三角形区域.然后基于该网络寻找若干参考单词对,进行匹配.DT 网络最大限度地避免了出现狭长、尖锐的三角形连接,并具有唯一性.当图像视觉单词表示给定时,用 Delaunay 三角剖分对该点进行三角剖分时,所产生的三角形网格是唯一的,与视觉单词的排序无关,只是与视觉单词的拓扑结构有关.可以使得少数视觉单词的缺失、个别干扰视觉单词的出现和视觉单词定位偏差等因素对划分结果的影响尽量限定在一个小的局部范围内.图 4(a)中如果视觉单词 $A$ 在图 4(b)中没有对应点,这表明单词 $A$ 要么对于图 4(b)来说是一种真正缺失的情况,要么对于图 4(b)来说是一个伪点,但这都不影响在相应位置右下方找到两对相似三角形.而对于图 4(a)和图 4(b)中对应的视觉单词 $B$ 和视觉单词 $C$,虽然视觉单词位置有偏差,而这也影响到它们与周围若干视觉单词之间的结构,但并不影响在视觉单词周围找到若干匹配的三角形对.在视觉单词集中插入一个单词,它仅仅影响到外接圆包含该插入点的三角形,对 Delaunay 三角形网格只造成局部影响.只要图像的一些区域没有严重损坏,都可以进行准确定位,从而有可能得以正确地识别.




(a) (b)

Fig.4 Delaunay triangle subdivision on visual word sets

图 4 视觉单词集进行 Delaunay 三角剖分

Delaunay 三角剖分还可以避免产生不稳定和瘦小的三角形.DT 网络具有极好的结构稳定性,它保证了 Delaunay 三角剖分得到的三角形不是狭长的,狭长的三角形会导致不稳定性甚至产生错误结果.对几种我们熟知的拓扑结构进行比较后发现,Delaunay 三角剖分对于随机的位置扰动具有最好的结构稳定性.

2.5.2 基于三角形限制的图像匹配

两幅待匹配图像之间具有几何相似性约束,且 Delaunay 三角形网络可以唯一地表示其特征,所以可以利用两幅图像 Delaunay 三角网中三角形之间的相似度进行匹配.

三角形相似性匹配包括以下步骤.

1) 相似边

在两幅图像的 Delaunay 三角网络中,任意选择一个三角形的边,分别为 $L_1$ 和 $L_2$,边长为 $l_1l_2$.边 $L_1$ 连接视觉单词 $w_{11}$ 和 $w_{12}$,边 $L_2$ 连接视觉单词 $w_{21}$ 和 $w_{22}$.

判断两条边是否相似,需满足以下条件.

(1) $|l_1-l_2|<T_1$.

(2) $(w_{11}=w_{21})\&(w_{12}=w_{22})\|(w_{11}=w_{22})\&(w_{12}=w_{21})$.

其中,阈值 $T_1$ 根据实验情况设定,若以上条件满足,则 $L_1$ 和 $L_2$ 为相似边.如果不满足则从参考图像的 DT 网络中找下一条边和 $L_1$ 进行匹配,直到图中所有三角边和 $L_1$ 都比对过.如果所有边都比较过,且没有找到一对相似边,则表示两幅图像不匹配.

2) 相似三角形

在三角网络中,边的结构包含 4 个标志位 $w_1,w_2,l,r$.$w_1$ 和 $w_2$ 为两端的顶点单词,$l$ 和 $r$ 对应边的左右是否有三角形.根据一条边的左右标志,可以找到这条边的左右邻接三角形.现有如图 5 所示的 $Tri_1(l_{11},l_{12},\theta_{11},\theta_{12},w_1)$和 $Tri_2(l_{21},l_{22},\theta_{21},\theta_{22},w_2)$两个三角形,$l_{11}$ 和 $l_{12}$ 分别为视觉单词 $P$ 与 $w_{11}$ 和 $w_{22}$ 的距离,$l_{21}$ 和 $l_{22}$ 分别为视觉单词 $Q$ 与 $w_{21}$ 和 $w_{22}$ 的距离.$\theta_1,\theta_{11},\theta_{12}$ 为 $Tri_1$ 对应的 3 个夹角,$\theta_2,\theta_{21},\theta_{22}$ 为 $Tri_2$ 对应的 3 个夹角,$w_1$ 和 $w_2$ 为视觉单词类型.另外,基于三角形内角相似即可判断三角形相似的原理,可以引入三角形的模糊相似度判断[29].三角形 $Tri_1$ 的一个内角为 $\theta_1$,那么在另一个三角形 $Tri_2$ 的对应角与该角的差异性变化应该遵循下式的高斯分布:

$$\mathrm{d}(\theta_2)=\mathrm{e}^{-k(\theta_2-\theta_1)^2} \tag{9}$$

这里,$\theta_2$ 表示另外一个三角形的对应角,$\mathrm{d}(\theta_2)$表示这两个角的差异值.其中大于 0 的系数 $k$ 可以由下式计算:

$$\begin{cases} k=\dfrac{1}{2\times\sigma^2} \\ \sigma=\dfrac{\theta_1\times P_{\theta_1}}{3} \end{cases} \tag{10}$$

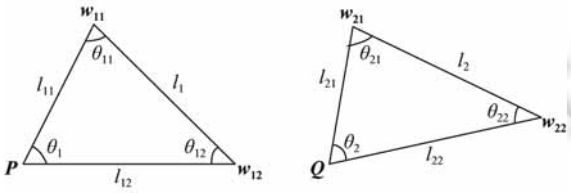



Fig.5 The comparison of similar triangle model

图 5 相似三角形的比较模型

式(10)是指 $\theta_1\times P_{\theta_1}$ 就是高斯分布的 $3\sigma$点,这里我们令 $P_{\theta_1}$ 为 50%.这样对角度的相似度可以由下式得到:

$$S_{\theta_1}=\cos^{n_s}(k_s(1-\mathrm{d}(\theta_2))) \tag{11}$$

基于余弦曲线的特征,我们可以定义 $n_s=3,k_s=\dfrac{\pi}{2}$. 对于一对三角形,可以先利用上式计算它们的内角相似度,然后利用下式计算整个三角形的相似度.

$$S_t=k_1S_1+k_2S_2+k_3S_3 \tag{12}$$

这里,$k_i(i=1,2,3)$表示两个三角形相应内角的相似度权值,我们取 $k_1=k_2=k_3=\dfrac{1}{3}$, 也就是说,3 个内角对三角形相似度的影响是相同的.匹配相似三角形是通过模糊相似度阈值大于一个阈值 $T_S$ 来获得的.

判断两个三角形是否相似,需要验证以下条件.

(1) $|l_{11}-l_{21}|<T_1 \,\&\, |l_{12}-l_{22}|<T_1$;

(2) $|\theta_{11}-\theta_{21}|<T_2 \,\&\, |\theta_{12}-\theta_{22}|<T_2 \,\&\, |\theta_1-\theta_2|<T_2$;

(3) $w_1=w_2$;

(4) $S_t>T_S$,

其中,$T_1$,$T_2$ 和 $T_S$ 均是根据实验选取的阈值,相似三角形比较步骤如下.

Step 1. 选取一条相似边,如果有左右邻接三角形,则对其进行判断,若满足以上 4 个条件,则认为这两个三角形相似.

Step 2. 若不满足以上条件,则返回上一步骤,处理下一对相似边.

Step 3. 当所有的相似边处理完成,算法结束,统计得到的相似三角形并记录其对应的顶点单词.

3) 三角形内点的匹配

在全局拓扑限制下,还存在一些未匹配的视觉单词,这些视觉单词位于三角形内部.对于三角形内部点,只能在其他图像的对应三角形内部寻找匹配点[27].

视觉单词限制在 $\Delta abc$ 和 $\Delta a'b'c'$中,对于视觉单词 $P_i$ 在 $\Delta abc$ 中,它和 $\Delta abc$ 三条边之间的关系为

$$P_i=\alpha+\beta(b-a)+\gamma(c-a) \tag{13}$$

$\beta,\gamma$ 分别是$(b–a)$和$(c–a)$的尺度系数,已知 $\Delta abc$ 的 3 条边 $a=[x_a,y_a,1]^T$,$b=[x_b,y_b,1]^T$ 和 $c=[x_c,y_c,1]^T$. 参数 $K$ 可以由下式计算:

$$K=\begin{bmatrix}\alpha\\ \beta\\ \gamma\end{bmatrix}=\begin{bmatrix}x_a & x_b & x_c\\ y_a & y_b & y_c\\ 1 & 1 & 1\end{bmatrix}^{-1}\begin{bmatrix}x_i\\ y_i\\ 1\end{bmatrix} \tag{14}$$

其中,$P_i=[x_i,y_i,1]^T$,$\alpha=1-\beta-\gamma$.如果 $\Delta abc$ 和 $\Delta a'b'c'$相似,$P_i$ 和 $\Delta a'b'c'$中的单词 $P_j$ 相匹配,则 $\Delta a'b'c'$的 3 条边也与参数 $K$ 相关.经过平移、尺度、旋转或者仿射变换后,估计的匹配单词 $P_j$ 可以通过 $P_i$ 的相关计算得到:

$$P_j=\begin{bmatrix}x_j\\ y_j\\ 1\end{bmatrix}=\begin{bmatrix}x'_a & x'_b & x'_c\\ y'_a & y'_b & y'_c\\ 1 & 1 & 1\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\\ \gamma\end{bmatrix} \tag{15}$$

假定 $\Delta abc$ 的 3 条边$[x_a,y_a,1]^T$,$[x_b,y_b,1]^T$ 和$[x_c,y_c,1]^T$ 通过矩阵 $H$ 变换为对应参照图像中 $\Delta a'b'c'$的 3 条边 $[x'_a,y'_a,1]^T$,$[x'_b,y'_b,1]^T$ 和 $[x'_c,y'_c,1]^T$ .同样地,$\Delta abc$ 内的视觉单词点$[x_i,y_i,1]$可以通过 $H[x_i,y_i,1]^T$ 转换为 $[x'_i,y'_i,1]^T$, 相关参数可以等价地表示为

$$K_H=\begin{bmatrix}\alpha'\\ \beta'\\ \gamma'\end{bmatrix}=\begin{bmatrix}x'_a & x'_b & x'_c\\ y'_a & y'_b & y'_c\\ 1 & 1 & 1\end{bmatrix}^{-1}\begin{bmatrix}x'_i\\ y'_i\\ 1\end{bmatrix}=\left(H\begin{bmatrix}x_a & x_b & x_c\\ y_a & y_b & y_c\\ 1 & 1 & 1\end{bmatrix}\right)^{-1}\left(H\begin{bmatrix}x_i\\ y_i\\ 1\end{bmatrix}\right) \tag{16}$$

可以证明 $K_H=K$.由于点的齐次坐标第 3 个维度保持为 1,不需要进行归一化,三角形约束具有平移、旋转、尺度和仿射变换不变性.由于三角剖分将图像平面划分为很小的块,所以对于噪声和扭曲更加鲁棒.

三角形内的视觉单词可能不止一个,因此假设视觉单词集为 $w=\{w_j\}$,$P_j$ 位于单词集 $w$ 中,$P_i$ 和待匹配单词 $w_j$ 的相似性得分由下式计算:

$$S_{i,j}=1.5^{-(dist_{i,j}/R)^2}\times D_i^T D_{w_j},\ j=1,2,\ldots,|w| \tag{17}$$

$dist_{i,j}$ 表示 $P_i$ 和 $\omega_j$ 的欧氏距离,$D_i$ 和 $D_{w,j}$ 分别表示 $P_i$ 和 $w_j$ 的视觉单词特征向量.其中,$1.5^{-(dist_{i,j}/R)^2}\in(0,1]$ 代表估计的待匹配单词 $P_j$ 和单词 $w_j$ 的空间距离,点积 $D_i^T D_{w_j}$ 计算特征向量的相似性.如果 $w_j$ 的相似性得分高于一个预定的阈值 $\tau$,则将该单词视为 $P_i$ 的一个匹配,且 $P_i$ 至多只有一个匹配,阈值 $\tau$ 将在实验中选择.

通过对待匹配图像 $P_A$ 的所有视觉单词进行处理后,将会得到一个与 $P_B$ 的匹配集合 $T$.

最终计算不同阶特征的核函数 $K(x,y)$,用于 SVM 训练.

## 3 实验结果与分析

实验在 Windows 7 操作系统(CPU 双核 2.60GHz,内存 4G),Visual Studio 2010 开发工具上由 C++语言实现.实验采用 Caltech101,COIL_100 和 Caltech256 这 3 个数据库验证本文算法.Caltech101 数据库包含 102 类共 9 145 幅图片,涵盖了由人到物的多种物体种类,每一类包含 40~800 张图片.COIL_100 数据库包含 100 种物体类别,每类物体包含 72 幅图片,类内图像具有复杂几何变换和仿射变换等特点.Caltech256 数据集包括 257 类共 30 607 幅图片,每个类别至少包含 80 张图片,相较于 Caltech101 数据集,物体种类更多,多样性增加.

我们采用的实验框架如下,首先进行图像分块与显著性区域提取,提取 128 维的 SIFT 特征描述子,使用标准的 $K$-means 聚类算法生成视觉单词表,其中,视觉单词容量参数 $K$ 会在实验中讨论,然后结合单词的空间位置信息,构建视觉单词的空间几何模型,最后应用 SVM 进行图像分类.我们重复实验 10 次,每次随机选择训练图像集和测试图像集,获得平均准确率和偏差.每个类别随机选择多幅图像作为训练样本,剩下的用于测试.

在实验中,我们主要考虑影响分类性能的几个因素,包括显著区域的提取、分块数、视觉词典个数以及 SVM 核函数等.

首先,本文算法在提取特征之前进行显著性区域提取.应用显著区域提取方法,可以在训练集较小的情况下达到较好的分类性能,通过对图像进行显著区域提取,然后根据该区域进行特征提取和聚类,由此产生的视觉单词更具有代表性,能更精确地描述图像,而且能在一定程度上避免复杂背景信息和位置变化带来的干扰和影响.部分图像显著区域的提取效果如图 6 所示.传统方法在训练图像较少的情况下性能不太稳定,这是由于类内图像具有差异性所致.

接下来我们对分块数大小的选取进行分析,分块大小主要影响显著区域的大小和个数.选取分块大小为 4×4、5×5、10×10,分别计算不同分块大小对图像分类准确率的影响,分块过大,选取的显著性区域仍会包含大面积的背景区域.分块过小,使得包含一些显著度不高的目标区域被排除在外.在 Caltech101 数据集中将所有图像归一化为 200×200,分块数为 5×5 较好.

Fig.6 Extract the saliency region of image

图 6 图像显著区域提取

分析了显著性区域和分块数之后,我们将开始在 Caltech101 数据集上进行实验.在 Caltech101 数据集上选择 5 类物体,包括 airplane、euphonium、motorbike、camera、watch.训练集每类随机选择 30 张图像,测试集中每类随机选择 50 张图像.视觉词典大小为 500,实验重复 10 次,得到平均分类精确度,如图 7 所示.

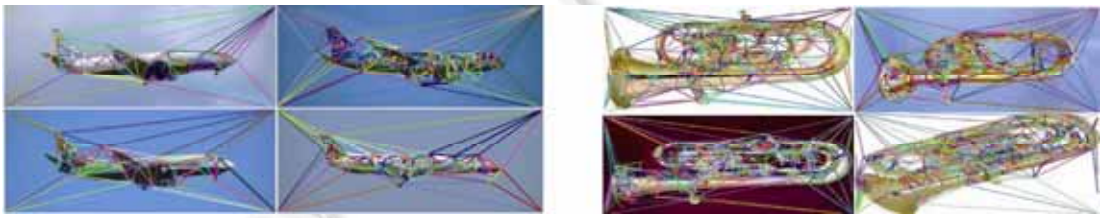


Fig.7 Spatial geometry information modeling

图 7 空间几何信息建模

本文工作的主要目的是结合视觉单词的空间位置信息,以改进视觉词袋模型应用于图像分类中的性能.空间信息编码方法(SC)[12]以及用户自定义感兴趣区域与空间分布相结合的方法(ROI+SL)[13]均利用了图像特征的空间信息,因此,我们运用 SC 算法以及 ROI+SL 算法改进 BoW 用于图像分类实验中,并和本文算法进行比较,分析本文算法的有效性.

这部分我们针对视觉词典大小对分类性能的影响进行分析,这里采用不同大小的视觉词典{50,100,200,300,500,1000}.如图 8 所示,$x$ 轴表示词典大小,$y$ 轴表示分类准确率,采用不同数目的视觉单词对分类性能产生一定的影响.根据分类准确率曲线,我们可以得出使用本文算法、SC 算法和 ROI+SL 算法的分类性能均有所提升,而且性能均好于传统的 BoW 模型.当视觉单词数目过小时,由于不同语义的图像块可能被标记为相似的视觉单词,使得无论是应用传统 BoW 模型还是改进的方法,它们的分类性能都相对较低.随着视觉单词数目的逐步增加,图像分类性能有所提升,由于传统 BoW 模型仅仅应用视觉单词的统计信息,忽略了视觉单词的空间信息,其分类准确率明显低于改进的方法,这说明,结合视觉单词的空间信息取得了效果.当视觉单词数目为 500 时,传统的 BoW 方法得到的分类准确率仅为 52.7%,SC 算法得到的分类准确率为 62.4%,ROI+SL 算法得到的分类准确率为 58.2%,而本文算法得到的分类准确率为 71.2%,加入空间信息后,分类准确率有很大的提升.

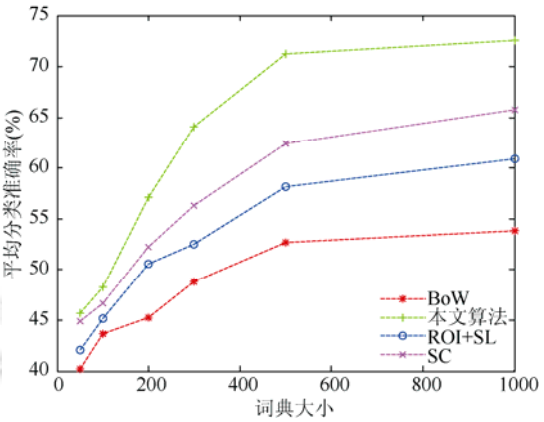



Fig.8 Dictionary capacity's impact on the average classification performance

图 8 词典容量对平均分类性能的影响

通过和 SC 算法以及 ROI+SL 算法的实验对比,可以看出本文算法在结合视觉单词空间位置信息上的优越性.SC 方法有效地编码了图像特征的空间位置关系,通过二进制的空间图描述特征对的相对空间位置,对局部特征施加几何限制.这种方法潜在要求待匹配图像具有重复的图像块,也就是说,具有相同或非常相似的匹配特征点的空间构型.由于存在无法避免的量化误差,该方法使用空间验证来消除错误的匹配.ROI+SL 算法结合用户自定义的感兴趣区域和空间分布进行特征匹配,基于用户定义感兴趣区域和图像分块的重叠百分比赋予图像特征不同的权重,图像的相似性计算通过单个图像块的相似性距离的线性组合计算得出.以上两种方法的分类准确率均低于本文算法,其中,SC 算法对于具有相同或非常相似的重复图像块的图像均有良好的分类效果,但是对于 Caltech101 数据库内类内差别较大的图像类别,分类效果提升得不会很明显,图像的仿射变换以及局部特征描述子的偏移也会对实验效果产生影响.ROI+SL 算法具有空间依赖性,用户选取的感兴趣区域对于算法效果影响显著,人工选取感兴趣区域虽然可靠性高,但是选取效率低,依赖于人机交互,过高的空间依赖导致图像分类效果提升有限.例如,一幅左下角有车的图像可能很难与右下角有车的图像分为同一类别.本文算法首先运用自适应的方法提取显著性区域,基于视觉单词拓扑结合策略,利用矩阵的特征分解有效地提取了关于视觉单词位置的全局信息.引入 Delauany 三角剖分方法进行局部空间位置约束,图像空间上位置相近的视觉单词按照一定规则相连,得到 DT 网络,DT 网络具有良好的结构稳定性,可以使得少数视觉单词的缺失、个别干扰视觉单词的出现和视觉单词定位偏差等因素对分类结果的影响尽量限定在一个小的局部范围内.本文算法结合了一阶、二阶和三阶单词特征,将视觉单词的空间几何信息融入传统的视觉词袋模型,这里提到的 $n$ 阶特征($n$=1,2,3),我们将在下文进行讨论.

本文算法不仅与 SC 算法和 ROI+SL 算法进行了比较,还与文献[30,31]中的图像分类实验结果进行了对比.文献[30]中主要通过对视觉单词的全局空间分布进行建模,并利用图像中相同视觉单词对的方向信息,改进传统 BoW 模型.文献[31]中提出空间金字塔匹配的改进方法,捕获更精确的图像特征空间信息.从实验结果可以看出,加入空间信息的图像特征表示方法具有较好的分类性能,其中,文献[30]的分类准确率为 67.1%,文献

[31]的分类准确率为 65.93%,SC 算法的分类准确率为 62.4%,ROI+SL 算法的分类准确率为 58.2%.对比 BoW 的实验结果分别提升 14.4%,13.23%,9.7%和 5.5%.本文算法的分类准确率相比 BoW 提升了 18.5%,分类准确率高于以上其他方法,见表 1.

**Table 1** Classification results for the Caltech101 dataset

**表 1** Caltech101 图像库上的实验结果

| 方法 | 平均准确率/(%) |
|---|---|
| BOW | 52.7 |
| SC[12] | 62.4 |
| ROI+SL[13] | 58.2 |
| 文献[30] | 67.1 |
| 文献[31] | 65.93 |
| 本文算法 | **71.2** |

在本文算法中,我们使用 Delaunay 三角剖分对单词的空间几何信息进行建模,几何信息建模的效果如图 7 所示.第 2.3 节已经介绍了高阶特征的定义,实验中,我们不仅单独使用 $n$ 阶特征的核函数 $K_n(n=1,2,3)$进行分类实验,还使用 1 到 $n$ 阶核函数的加权和去验证算法,其中,$u=0.02$.我们发现,无论是单独使用 $n$ 阶特征的核函数还是进行加权和,从一阶(视觉单词)到二阶或三阶特征,效果都获得了提升.这表明,对视觉单词的空间几何信息建模起到了重要作用.然而随着阶数的增加,一些图像将不会出现共现的特征.分类效果的精确度将会随着我们提升阶数而有所提高(尽管会在某一阶数停止),我们发现虽然高阶核函数的精确度会有所下降,将它和其他阶特征结合在一起,但高阶特征仍能在一定程度上提升分类性能,事实上,高阶特征更具区分能力.

另外,随着阶数的上升,某些类别的图像类内相似性降低,原因在于某些类内图像具有更多变的目标结构、尺度和旋转,导致共现的高阶特征会减少,影响了分类性能.

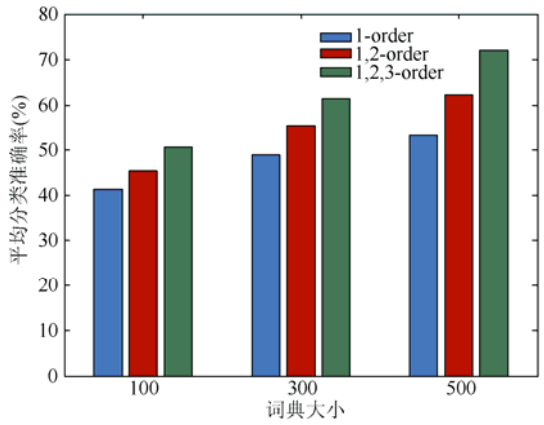



Fig.9 The influence of different order number for the average classification performance

图 9 不同阶数对于平均分类性能的影响

在分类过程中,我们应用 SVM 核函数,对于不同的核赋予不同的权重.如图 9 所示,增加特征阶数后,可以看到性能的提升.实验中,我们采用的词典大小为 500.我们尝试将词典大小降到 100 进行图像分类,在这种情况下,某些单个的视觉单词将可能变得无意义,在使用低阶特征进行分类时准确率很低.然而对于 SVM,当提高特征的阶数后,得到的性能非常接近我们使用 500 个视觉单词时所获得的性能.这表明,即使局部特征本身的区分能力不足,通过提高阶数也能够创造具有区别能力的特征.当将一阶、二阶和三阶特征结合时,可以达到最好的分类性能.

接下来将在 COIL_100 数据集上进行分类实验,在 COIL_100 数据集上随机选择 6 类物体,训练集中每类随机选择 20 张图像,测试集中每类随机选择 50 张图像.视觉词典大小为 500,实验重复 10 次,得到平均分类精确度.如图 10 所示,为 COIL_100 数据集上图像视觉单词空间几何信息建模结果.

本文实验还应用于 Caltech256 数据集上,我们在 Caltech256 数据集中选择 10 类物体,包括 butterfly、calculator、camel、necktie、elephant 等类别.训练集每类随机选择 30 张图像,测试集中每类随机选择 50 张图像.视觉词典大小为 1 000,实验重复 10 次,得到平均分类精确度.

表 2 为 COIL_100 图像库中使用不同的方法进行图像分类的实验结果.将本文算法与传统的 BoW 模型以及 SC 算法[12]、ROI+SL 算法[13]和文献[32]中的方法进行了实验对比,可以看出,加入空间信息的特征表示显示了良好的分类性能,SC 算法编码了图像特征的空间位置关系,ROI+SL 算法基于用户自定义感兴趣区域,结合了图像特征进行了空间分布,均在一定程度上提升了图像特征的区分力和表达能力,提高了分类准确率.本文算法

较传统 BoW 模型的分类准确率提升 5.3%,效果最佳.表 3 为在 Caltech256 图像库中使用不同的特征表示方法进行图像分类的实验结果,Caltech256 图像库中类内变化更多样,分类难度有所增加.表中结果显示本文方法取得了较好的实验结果,与传统的 BoW 方法和空间金字塔匹配(SPM)方法[2]、SC 算法、ROI+SL 算法相比,分类性能分别提高了 10.3%、7.4%、5.3%和 13%.文献[33]所采用的卷积神经网络方法效果最佳,平均准确率为 70.6%.然而,基于卷积神经网络的方法需要大量的样本和时间来学习特征,因此针对一些小的数据集,由于不需要预先的监督学习这一步骤,本文算法仍然能够起到重要的作用,效果更好.

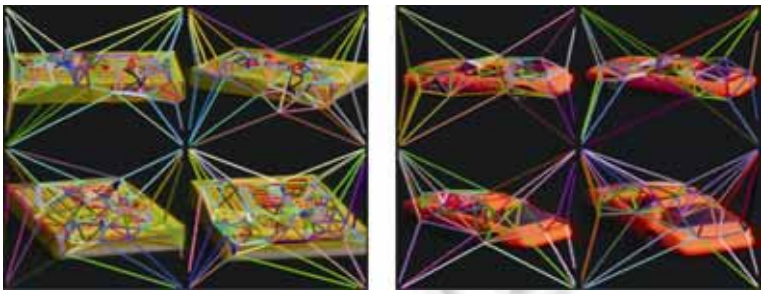


Fig.10 The spatial geometry information model of visual words

图 10 视觉单词空间几何信息建模

**Table 2** Classification results for the COIL_100 dataset

**表 2** COIL_100 图像库上的实验结果

| 方法 | 平均准确率(%) |
|---|---|
| BOW | 91.3 |
| SC[12] | 93.2 |
| ROI+SL[13] | 91.6 |
| 文献[32] | 90.52 |
| 本文算法 | **96.8** |

**Table 3** Classification results for the Caltech256 dataset

**表 3** Caltech256 图像库上的实验结果

| 方法 | 平均准确率(%) |
|---|---|
| BOW | 31.2 |
| SPM[2] | 34.1 |
| SC[12] | 36.2 |
| ROI+SL[13] | 28.5 |
| 文献[33] | 70.6 |
| 本文算法 | **41.5** |

本文算法主要针对传统视觉词袋模型进行改进,结合视觉单词的空间位置信息,提高图像特征表示的区分能力,并应用于图像分类.在采用词袋模型之前,使用显著性区域检测进行筛选,分类器是改进的 SVM.从表 1~表 3 中均可以看出本文算法取得了良好的效果,克服了图像的几何、旋转和仿射变换的影响,排除了复杂背景区域的干扰,对于类内差别较大的数据集,仍然能够避免背景干扰,有效应对类内变化,取得良好的分类效果.

## 4 总 结

视觉词袋模型广泛地应用于图像分类领域.在传统词袋模型中,视觉单词统计方法忽略了视觉词之间的空间信息以及分类对象的形状信息,导致图像特征表示区分能力不强.本文提出了一种改进的视觉词袋方法,结合显著区域提取和视觉单词空间位置结构,采用视觉单词拓扑结构策略和三角剖分方法,融入全局信息和局部信息,不仅能够产生更具代表性的视觉单词,而且能够在一定程度上避免复杂背景信息和位置变化带来的干扰.实验结果表明,本文算法将空间信息加入视觉词袋模型,并结合高阶特征,能够提高图像分类的精确度,而且与现有方法相比具有优越性.

我们还可以看到本文方法的改进之处,包括结合其他的核函数来改进视觉单词空间位置模型,一些其他的编码方法也可以引入,例如局部软分配方法[34]可以用于软相似度加权.最后,空间信息还可以由多样的信息提供,例如颜色和纹理等,这是一个值得研究的方向.

**References**:

[1] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos. In: Proc. of the 9th IEEE Int'l Conf. on Computer Vision. IEEE, 2003. 1470–1477. [doi: 10.1109/ICCV.2003.1238663]

[2] Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. of the 2006 IEEE Conf. on Computer Vision and Pattern Recognition. IEEE Computer Society, 2006,2:2169–2178. [doi: 10.1109/CVPR.2006.68]

[3] Grauman K, Darrell T. The pyramid match kernel: Discriminative classification with sets of image features. In: Proc. of the 10th IEEE Int'l Conf. on Computer Vision, ICCV 2005. IEEE, 2005,2:1458–1465. [doi: 10.1109/ICCV.2005.239]

[4] Lu ZW, Ip H. Combining context, consistency, and diversity cues for interactive image categorization. IEEE Trans. on Multimedia, 2010,12(3):194–203. [doi: 10.1109/TMM.2010.2041100]

[5] Jiang Y, Wang RS, Wang C. Scene classification with context pyramid features. Journal of Computer-Aided Design and Computer Graphics, 2010,22(8):1366–1373 (in Chinese with English abstract).

[6] Liu SY, Xu D, Feng SH, Liu D, Qiu ZD. A novel visual words definition algorithm of image patch based on contextual semantic information. Dianzi Xuebao, 2010,38(5):1156–1161 (in Chinese with English abstract).

[7] Li T, Mei T, Kweon IS, Hua XS. Contextual bag-of-words for visual categorization. IEEE Trans. on Circuits and Systems for Video Technology, 2011,21(4):381–392. [doi: 10.1109/TCSVT.2010.2041828]

[8] Zhang SL, Tian Q, Hua G, Huang QM, Li SP. Descriptive visual words and visual phrases for image applications. In: Proc. of the 17th ACM Int'l Conf. on Multimedia. ACM, 2009. 75–84. [doi: 10.1145/1631272.1631285]

[9] Zheng QF, Wang WQ, Gao W. Effective and efficient object-based image retrieval using visual phrases. In: Proc. of the 14th Annual ACM Int'l Conf. on Multimedia. ACM, 2006. 77–80. [doi: 10.1145/1180639.1180664]

[10] Sivic J, Zisserman A. Efficient visual search of videos cast as text retrieval. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009,31(4):591–606. [doi: 10.1109/TPAMI.2008.111]

[11] Wang MY, Zhang CL, Song Y. Extraction of image semantic features with spatial-range mean shift clustering algorithm. In: Proc. of the 10th IEEE Int'l Conf. on Signal Processing (ICSP). IEEE, 2010. 906–909. [doi: 10.1109/ICOSP.2010.5655732]

[12] Zhou WG, Lu YJ, Li HQ, Song YB, Tian Q. Spatial coding for large scale partial-duplicate web image search. In: Proc. of the 18th ACM Int'l Conf. on Multimedia. ACM, 2010. 511–520. [doi: 10.1145/1873951.1874019]

[13] Tian Q, Wu Y, Huang TS. Combine user defined region-of-interest and spatial layout for image retrieval. In: Proc. of the 2000 Int'l Conf. on Image. IEEE, 2000,3:746–749. [doi: 10.1109/ICIP.2000.899562]

[14] Harel J, Koch C, Perona P. Graph-Based visual saliency. In: Advances in Neural Information Processing Systems. 2006. 545–552. https://nips.cc/Conferences/2007

[15] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. on Pattern Analysis & Machine Intelligence, 1998,(11):1254–1259. [doi: 10.1109/34.730558]

[16] Hou XD, Zhang LQ. Saliency detection: A spectral residual approach. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2007. IEEE, 2007. 1–8. [doi: 10.1109/CVPR.2007.383267]

[17] Ma YF, Zhang HJ. Contrast-Based image attention analysis by using fuzzy growing. In: Proc. of the 11th ACM Int'l Conf. on Multimedia. ACM, 2003. 374–381. [doi: 10.1145/957013.957094]

[18] Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-Tuned salient region detection. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2009. IEEE, 2009. 1597–1604. [doi: 10.1109/CVPR.2009.5206596]

[19] Li Y, Tsin YH, Genc Y, Kanade T. Object detection using 2D spatial ordering constraints. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2005. IEEE Computer Society, 2005,2:711–718. [doi: 10.1109/CVPR.2005.253]

[20] Ma J, Ahuja N. Region correspondence by global configuration matching and progressive Delaunay triangulation. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. IEEE, 2000,2:637–642. [doi: 10.1109/CVPR.2000.854932]

[21] Zeng D, Shi H, Zhang Q. Feature matching between images with multiple similar contents. Journal of Computer-Aided Design & Computer Graphics, 2011,23(10):1725–1733 (in Chinese with English abstract).

[22] Jegou H, Douze M, Schmid C. Product quantization for nearest neighbor search. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011,33(1):117–128. [doi: 10.1109/TPAMI.2010.57]

[23] Kalantidis Y, Avrithis Y. Locally optimized product quantization for approximate nearest neighbor search. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2014. 2321–2328. [doi: 10.1109/CVPR.2014.298]

[24] Wei SK, Xu D, Li XL, Zhao Y. Joint optimization toward effective and efficient image search. IEEE Trans. on Cybernetics, 2013, 43(6):2216–2227. [doi: 10.1109/TCYB.2013.2245890]

[25] Zhang YM, Jia ZY, Chen T. Image retrieval with geometry-preserving visual phrases. In: Proc. of the 2011 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). IEEE, 2011. 809–816. [doi: 10.1109/CVPR.2011.5995528]

[26] Zhang YM, Chen T. Efficient kernels for identifying unbounded-order spatial features. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2009. IEEE, 2009. 1762–1769. [doi: 10.1109/CVPR.2009.5206791]

[27] Guo XJ, Cao XC. Good match exploration using triangle constraint. Pattern Recognition Letters, 2012,33(7):872–881. [doi: 10.1016/j.patrec.2011.08.021]

[28] Tang G. Study on improved fingerprint identification algorithm based on Delaunay triangulation [Ph.D. Thesis]. Shenyang: Northeastern University, 2008 (in Chinese with English abstract). [doi: 10.7666/d.y1842528]

[29] Li GH, Zhou DX, Dong L, Liu YH, Cai XP. Effective corner matching based on Delaunay triangulation. Signal Processing, 2007,23(5):695–698 (in Chinese with English abstract). [doi: 10.3969/j.issn.1003-0530.2007.05.012]

[30] Khan R, Barat C, Muselet D, Ducottet C. Spatial orientations of visual word pairs to improve bag-of-visual-words model. In: Proc. of the British Machine Vision Conf. BMVA Press, 2012. 89.1–89.11. http://bmvc2012.surrey.ac.uk/

[31] Zhang E, Mayo M. Improving bag-of-words model with spatial information. In: Proc. of the 25th Int'l Conf. of Image and Vision Computing, IVCNZ. IEEE, 2010. 1–8. [doi: 10.1109/IVCNZ.2010.6148795]

[32] Han XH, Chen YW, Ruan X. Multilinear supervised neighborhood embedding of a local descriptor tensor for scene/object recognition. IEEE Trans. on Image Processing, 2012,21(3):1314–1326. [doi: 10.1109/TIP.2011.2168417]

[33] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Proc. of the Computer Vision, ECCV 2014. Springer Int'l Publishing, 2014. 818–833. [doi: 10.1007/978-3-319-10590-1_53]

[34] Liu LQ, Wang L, Liu XW. In defense of soft-assignment coding. In: Proc. of the 2011 IEEE Int'l Conf. on Computer Vision (ICCV). IEEE, 2011. 2486–2493. [doi: 10.1109/ICCV.2011.6126534]

## 附中文参考文献:

[5] 江悦,王润生,王程.采用上下文金字塔特征的场景分类.计算机辅助设计与图形学学报,2010,22(8):1366–1373.

[6] 刘硕研,须德,冯松鹤,刘镝,裘正定.一种基于上下文语义信息的图像块视觉单词生成算法.电子学报,2010,38(5):1156–1161.

[21] 曾丹,史浩,张琦.多相似内容图像的特征匹配.计算机辅助设计与图形学学报,2011,23(10):1725–1733.

[28] 唐果.基于 Delaunay 三角剖分的指纹识别改进算法的研究[博士学位论文].沈阳:东北大学,2008. [doi: 10.7666/d.y1842528]

[29] 李赣华,周东祥,董黎,刘云辉,蔡宣平.基于 Delaunay 三角化的有效角点匹配算法.信号处理,2007,23(5):695–698. [doi: 10.3969/j.issn.1003-0530.2007.05.012]

汪荣贵(1966－),男,安徽池州人,博士,教授,博士生导师,主要研究领域为视频大数据,多媒体技术,人工智能,模式识别.

薛丽霞(1976－),女,博士,副教授,主要研究领域为图像处理,视频大数据,多媒体技术.

丁凯(1992－),男,硕士,主要研究领域为图像处理,视频大数据.

张清杨(1990－),男,硕士,主要研究领域为进化计算,图像处理,人工智能.

杨娟(1983－),女,博士,讲师,主要研究领域为智能信息处理.