# 改进视觉词袋模型的快速图像检索方法①

张祯伟, 石朝侠

(南京理工大学 计算机科学与工程学院, 南京 210094)

**摘 要**: 视觉词袋模型在基于内容的图像检索中已经得到了广泛应用, 传统的视觉词袋模型一般采用 SIFT 描述子进行特征提取. 针对 SIFT 描述子的高复杂度、特征提取时间较长的缺点, 本文提出采用更加快速的二进制特征描述子 ORB 来对图像进行特征提取, 建立视觉词典, 用向量间的距离来比较图像的相似性, 从而实现图像的快速检索. 实验结果表明, 本文提出的方法在保持较高鲁棒性的同时, 明显高了图像检索的效率.

**关键词**: 视觉词袋模型; 局部特征; ORB; 图像检索

## Fast Image Retrieval Method Using Improved Bag of Visual Words Model

ZHANG Zhen-Wei, SHI Chao-Xia

(School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China)

**Abstract**: Bag of visual words model based on content-based image retrieval has been widely used, traditional bag of visual words model generally uses the SIFT descriptors for feature extraction. In view of the high complexity of SIFT descriptors and the long time of feature extraction, this paper proposes to use a faster binary feature descriptor ORB for the image feature extraction, creating visual dictionary, using the distance between two vectors to compare the image similarity, so as to achieve fast image retrieval. Experimental results show that the method proposed in this paper can improve the efficiency of image retrieval obviously, while maintains a relatively high robustness.

**Key words**: bag of visual words; local features; ORB; image retrieval

图像检索技术是机器视觉领域中关注对大规模数字图像进行检索和匹配的研究分支. 它是图像拼接、目标跟踪、运动分析、对象识别、视觉导航等研究方向的研究基础. 近些年伴随着嵌入式系统处理能力及存储容量的 快速提升, 智能终端、机器人等嵌入式系统对大数据量图像检索需求日益增加, 从海量数据中快速检索和匹配到所需的信息已具有很大的研究价值.

在基于内容的图像检索中, 视觉词袋模型(Bag of Visual Word, BoVW)[1]已经成为一种比较常见的方法. 词袋模型(Bag of Words, BoW)最初应用于文档处理领域, 将文档表示成顺序无关的关键词的组合, 通过统计文档中关键词出现的频率来进行匹配. 作为一种源自文本检索的模型, 视觉词袋模型近年来在计算机视觉研究领域表现出良好的适用性, 成为计算机视觉研究的通用模型. BoVW 首先在视频检索的研究中被系统地阐述与应用, 近几年来, 计算机视觉领域的研究者们成功地将该模型的思想移植到图像处理领域, 通过对图像进行特征提取和描述, 得到大量特征进行处理, 从而得到用来表示图像的关键词, 并在此基础上构建视觉词典进而图像可以类似于文本的表示方法即统计基本词汇出现的频数, 将图像表示成一个向量, 利用该向量进行图像的检索. 传统的词袋模型一般采用 SIFT(Scale-Invariant Feature Transform)特征描述子[2], SIFT 算法可以适应图像缩放、旋转、平移等变化, 并且能克服噪声光照变化的影响. 但是 SIFT 算法的计算量比较大, 无法满足系统实时性的要求. 针对 SIFT 描述子的高复杂度问题本文提出了采用更加快速的二进

① 基金项目:国家自然科学基金(61371040)

收稿时间:2016-03-14;收到修改稿时间:2016-04-14 [doi:10.15888/j.cnki.csa.005464]

制特征描述子 ORB[3]来对图像进行特征抽取, 然后利用 BoVW 模型进行建模, 将每一副图像用一个二进制串来表示, 进行图像的检索. 实验表明, 该方法不仅保持了较高的图像检索准确率, 而且大大提高了图像的检索速度.

## 1 视觉词袋模型

BOW 算法起源于基于语义的文本检索算法, 是一种有效的基于语义特征提取和描述的识别算法. 该算法忽略文本的结构信息和语法信息, 仅仅将其看做是若干个词汇的集合, 文本内的每个词的出现都是独立的, 提取其中的语义特征, 构建单词词汇表, 根据每个文本与词汇表的关系, 统计文本中相应单词的出现频率, 形成一个词典维度大小的单词直方图, 经过这样文本到向量运算问题的转化, 最后实现文本检索. 将对文本处理的词袋模型过渡到图像处理领域, 便形成了视觉词袋模型.

### 1.1 算法流程

其实现过程大致分为四个步骤: 首先提取图像中的特征描述子; 然后通过聚类算法将训练图片得到特征描述子进行相似点聚类, 每个聚类中心代表一个视觉单词; 将图像的局部视觉特征映射到视觉单词表并用一个特征向量表示, 特征向量的每一维对应一个视觉单词的权重之和. 最后利用图像生成的向量进行图像检索. 算法流程如图 1 所示.

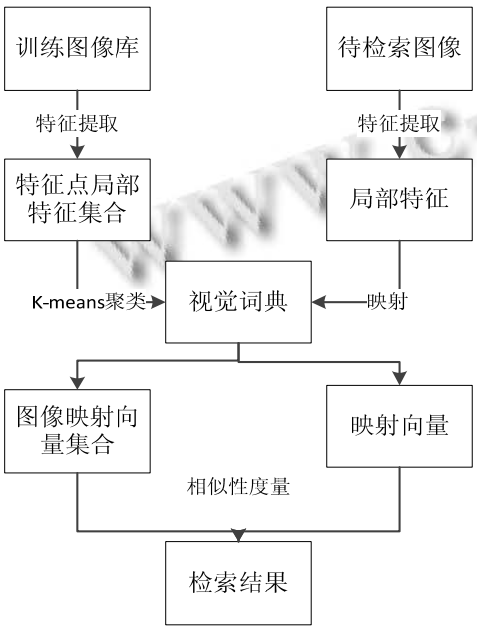


图 1 视觉词袋模型流程

根据图 1, 应用词袋模型进行图像检索的具体实现过程可以描述如下:

(1) 特征提取和描述.

视觉词袋模型往往选取图像底层的 SIFT 特征, 该特征具有旋转、尺度、平移等不变性, 同时对仿射变换, 噪声存在一定的稳定性. SIFT 特征计算主要分为图像特征点的选取和图像特征区域的描述两个部分. 图像特征点的选取步骤如下: 首先对图像建立一个图像金字塔模型, 然后对图像在相邻尺度空间上做差分, 选取尺度空间中的极值点, 最后将极值点周围的一定范围的区域作为特征区域.

(2) 视觉词典构造.

BOW 算法通常采用 k-means 算法对提取的特征进行聚类生成视觉词典. k-means 算法是一种经典的聚类算法, 是典型的基于原型的目标函数聚类方法的代表, 它是数据点到原型的某种距离作为优化的目标函数, 利用函数求极值的方法得到迭代运算的调整规则.

视觉词典构造主要步骤如下:

① 给定待聚类的图像 SIFT 描述子数据集,随机选取 K 个对象作为初始聚类中心.

② 求出 SIFT 描述子数据集中的每个数据与各个聚类中心的距离, 按照最小化原则将数据点划入最近邻聚类中心所在的类簇.

③ 重新计算每个类簇的中心.

④ 重复步骤 2、3, 当各个聚类中心不再改变时算法结束.

(3) 生成视觉直方图

该过程是将每幅图像所有的 SIFT 特征描述子分配到视觉词典的各个维度上, 生成各自的视觉单词直方图. 在分配的过程中, 采用最近邻算法, 每幅图像中的每个 SIFT 特征向量与哪一个视觉词距离最近, 就将该视觉词对应的维度高度加 1, 直到将所有的 SIFT 描述子向量分配完为止, 经过这一系列处理后, 每一幅图像都能用一个 k 维的视觉词直方图表示, 将所有图像的视觉词直方图归一化处理后就可以进行下一步的.

### 1.2 权值的计算

在文本信息检索中, TF-IDF[4]是一种常用的加权方案. TF-IDF 的主要思想是: 如果某个词或短语在一篇文章中出现的频率 TF 高, 并且在其他文章中很少出现, 则认为此词或者短语具有很好的类别区分能力,

适合用来分类. TF 表示词条在文档 d 中出现的频率, 如果一个词条在一个类的文档中频繁出现, 则说明该词条能够很好代表这个类的文本的特征, 这样的词条应该给它们赋予较高的权重, 并选来作为该类文本的特征词以区别与其它类文档.

IDF 的主要思想是: 如果包含单词 $F_i$ 的文档越少, 也就是 $n_i$ 越小, IDF 越大, 则说明单词 $F_i$ 具有很好的类别区分能力. 假设训练集中的图片总数为 $N$, $n_i$ 表示包含单词 $F_i$ 的图片数目. 类似于文本检索当中的逆文档频率 $idf$, 定义为:

$$idf_i = \lg \frac{N}{n_i} \tag{1}$$

即该单词被赋予的权值, 它表明了该单词对于区分不同图像时作用的大小.

## 2 基于ORB特征的视觉词袋模型

视觉词袋模型通常选取图像底层的 SIFT 特征, 该特征具有旋转、尺度、平移等不变性, 同时对仿射变换, 噪声存在一定的稳定性. 为了进一步提高算法实时性, 本文采用 ORB 算法进行特征提取. ORB 算子基于 BRIEF 算子提出, 是对 BRIEF 算子的改进. 文献[3]指出, ORB 算法的速度比 SIFT 要快两个数量级, 同时在不考虑图像尺度变化的情况下, 其匹配性能并不逊色于 SIFT.

### 2.1 图像特征提取和描述

构建视觉词汇表之前, 首先要从图像中提取出具有代表性的全局特征或局部特征, 作为对该图像的“描述”. 这些被提取的特征应该具有较强的稳定性, 能够抵抗光照、视角尺度等因素带来的不利影响. BOW 通常采用局部特征来生成视觉词汇表的候选特征, 在图像识别和物体匹配的过程中, 由于 ORB 描述子计算速度上的优势, 本文采用 ORB 描述子来提取和描述图像的特征点.

#### 2.1.1 特征点提取

ORB(oriented FAST and rotated BRIEF) 是基于 FAST[5]特征检测和 BRIEF 描述子[6]改良的. 该算法使用 FAST 角点检测来提取特征点, FAST 算法的角点定义为在像素点周围邻域内有足够多的像素点与该点处于不同的区域, 在灰度图像中, 即为有足够多的像素点的灰度值与该点灰度值差别够大. 以候选特征 D 为中心, 比较中心点 D 的灰度值与以 D 点为中心的圆周上所有点灰度值之间的大小, 如果圆周上与 D 点灰度值相差足够大的点个数超过一定数值, 则认为候选点 D 为特征点. FAST 角点检测仅仅比较灰度值大小, 具有计算简单、速度较快的优点, 但其检测出的特征点既不具备尺度不变性也不具备旋转不变性.

FAST 不提供角点的度量, 对边缘的响应较大, 因此 ORB 采用 Harris 角点度量的方法按照 FAST 特征点的 Harris 角点响应值对 FAST 特征点进行排序. 如需要提取 N 个特征点, 首先将阈值设置的足够大以得到更多的特征点, 然后根据 Harris 响应值排序, 最后选出响应值最大的 N 个特征点.

由于 FAST 特征点是不带有方向性的, ORB 的论文中提出了一种利用灰度质心法来解决这个问题, 灰度质心法假设角点的灰度与质心之间存在一个偏移, 这个向量可以用于表示一个方向. 对于任意一个特征点 O 来说, 我们定义 O 的邻域像素的矩为:

$$m_{pq} = \sum_{x,y} x^p y^q I(x,y) \tag{2}$$

其中 I(x,y)为点(x,y)处的灰度值. 那么我们可以得到图像的质心为:

$$C = \left(\frac{m_{10}}{m_{00}}, \frac{m_{01}}{m_{00}}\right) \tag{3}$$

特征点与重心的夹角定义为 FAST 特征点的方向: 构造一个从 O 到 C 的向量 $\overrightarrow{OC}$, 则特征点的方向为:

$$\theta = a\tan 2(m_{01}, m_{10}) \tag{4}$$

为了提高方法的旋转不变性, 需要确保 x 和 y 在半径为 r 的圆形区域内, 即 $x, y \in [-r, r]$, r 等于邻域半径.

#### 2.1.2 特征点描述

ORB 中使用 BRIEF 描述子对检测到的特征点进行描述, 并解决了 BRIEF 本身不具有旋转不变性的首要缺陷. 在以关键点为中心的图像块内比较采样点对的灰度值, 得到一个 n 位二进制数, 该 n 位二进制数即为关键点的特征描述子, n 的典型值为 256.

ORB 采用的是 BRIEF 描述子, 它的基本思想是是图像特征点邻域可以用相对少量的灰度对比来表达, 每个图像块由一系列二进制测试构成的位串来表示, 其计算简单、快速. 考虑一个平滑的图像块 p, 一个二进制测试 $\tau$ 定义为:

$$\tau(p; x, y) = \begin{cases} 1, p(x) < p(y) \\ 0, p(x) \geq p(y) \end{cases} \tag{5}$$

其中 p(x)是图像块 p 在点 x 处的灰度值. 特征点被定义

为一个由 n 个二进制测试构成的向量:

$$f_n(p)=\sum_{1\le i\le n}2^{i-1}\tau(p;x_i,y_i) \tag{6}$$

BRIEF 中图像邻域的准则仅考虑单个像素, 所以对噪声敏感. 为了解决这个缺陷, ORB 中每个测试点采用的是 31×31 像素邻域中的 5×5 子窗口, 其中子窗口的选择服从高斯分布, 再采用积分图像加速计算.

ORB 选择了 BRIEF 作为特征描述方法, 但是 BRIEF 是没有旋转不变性的, 所以需要给 BRIEF 加上旋转不变性, 把这种方法称为“Steered BREIF”. 对于任何一个特征点来说, 它的 BRIEF 描述子是一个长度为 n 的二值码串, 这个二值串是由特征点周围 n 个点对(2n 个点)生成的, 将这 n 个点对 $(x_i,y_i)$ 组成一个矩阵 S

$$S=\begin{pmatrix}x_1,x_2,\ldots,x_n\\ y_1,y_2,\ldots,y_n\end{pmatrix} \tag{7}$$

使用邻域方向 $\theta$ 和对应的旋转矩阵 $R_\theta$, 构建 S 的一个校正版本 $S_\theta$

$$S_\theta=R_\theta S \tag{8}$$

其中

$$R_\theta=\begin{pmatrix}\cos\theta & \sin\theta\\ -\sin\theta & \cos\theta\end{pmatrix} \tag{9}$$

此时 Steered BRIEF 描述子变为:

$$g_n(p,\theta)=f_n(p)\,|\,(x_i,y_i)\in S_\theta \tag{10}$$

ORB 根据式(8)中求得的方向参数提取 BRIEF 描述子. 但是由于环境的因素和噪声的引入, 特征点方向会发生变化, 随机像素块对的相关性会比较大, 从而降低描述子的判别性. ORB 采取贪心算法寻找相关性较低的随机像素块对, 一般选取 256 个相关性最低像素块对, 构成一个 256bit 的特征描述子.

由于生成的特征点描述子为二进制码串形式, 因此使用 Hamming 距离对特征点匹配较为简单. 计算机中计算汉明距离可以简单地通过异或进行计算. 汉明距离计算效率非常高.

假设上节得到 ORB 特征 256bit 二进制描述子 $K_1$、$K_2$ 两个特征点的描述子分别为:

$$K_1=x_0x_1\mathrm{L}\ x_{255},\quad K_2=y_0y_1\mathrm{L}\ y_{255}$$

通过汉明距离之间的异或之和表征两个 ORB 特征描述子的相似程度, 用 $D(K_1,K_2)$ 表示:

$$D(K_1,K_2)=\sum_{i=0}^{255}x_i\oplus y_i \tag{11}$$

$D(K_1,K_2)$ 越小代表相似程度越高, 反之相似程度低.

### 2.2 生成视觉单词

在提取到图像的 ORB 描述子之后, 需要进行视觉词典的构建. 该过程通常分为两步来完成. 首先将代表图像局部特征的描述子转换为视觉词, 一个视觉单词可以看作图像中相似的特征点的集中代表, 该过程是通过聚类算法实现的. 最终得到的聚类中心就是我们所期望的视觉单词, 聚类中心的个数就是视觉词典的大小. 根据聚类的视觉单词来建立每张图像的视觉词直方图, 该过程称为映射.

视觉词袋模型中单词数目的选取出现在特征描述的量化过程中, 常见的量化方法是 k-means 聚类, 词汇数目即对应的聚类数目. 但是由于 ORB 描述子产生的是二进制描述向量, 无法直接采用传统的基于欧氏距离的 k-means 方法进行聚类, 因此, 本文采用 Hamming 距离计算各个特征之间的距离, 使用 k-majority 算法[7]来求二进制描述向量的聚类中心. 具体算法流程如下:

假设从图像中提取到的 ORB 特征描述子集合 D.

步骤 1. 随机生成 k 个二进制聚类中心记为集合 C.

步骤 2. 计算 D 中各描述子到各个聚类中心的距离, 并划分到个类中.

步骤 3. 重新计算各类的聚类中心.

重复步骤 2、3, 当各个聚类中心不再改变时算法结束.

其中步骤 3 中聚类中心的计算方法如下:

假设某一具有 n 个特征描述子的集合 D

$$d_i=b_{i1}b_{i2}\mathrm{L}\ b_{i256} \tag{12}$$

其聚类中心为 $c=c_1c_2\mathrm{L}\ c_j\mathrm{L}\ c_{256}$, 其中

$$c_j=\begin{cases}0,\sum_{1\le i\le 256}d_{ij}<\dfrac{n}{2}\\ 1,\sum_{1\le i\le 256}d_{ij}\ge\dfrac{n}{2}\end{cases} \tag{13}$$

即对于集合中所有特征描述子的每一个 bit, 统计所有特征的对应 bit 上的 0、1 的数量, 并取高者作为该 bit 的值. 这样得到的聚类中心向量也是二进制表示, 在进行距离计算时可以利用汉明距离进行快速计算.

通过聚类最终得到的 k 个聚类中心即为所求的视觉单词. 图像特征聚类过程如图 2 所示.

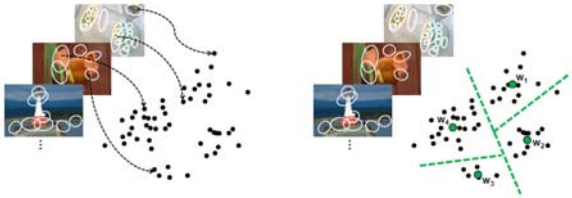

图 2 图像特征聚类

### 2.3 图片的向量表示

对于训练集中每一个图像，累计图片中的特征在单词表中的每一个单词 $F_i(1 \le i \le t)$ 当中出现的频率 $m_i$，t 为视觉单词总数. 由于在训练阶段已得到该单词的权值，即 $\lg\frac{N}{n_i}$，同样根据 TF-IDF 的原理，计算出该图像在单词 $F_i$ 维度上的值:

$$w_i = m_i \times \lg\frac{N}{n_i}, (1 \le i \le t) \tag{14}$$

最终，每一副图像 $d_j$ 都可以用关于单词的权值向量表示:

$$d_j = (w_{1,j}, w_{2,j}, \mathrm{L}\ , w_{t,j}) \tag{15}$$

### 2.4 图片间的相似度测量

训练集中图像 $d_j = (w_{1,j}, w_{2,j}, \mathrm{L}\ , w_{t,j})$，待查询图像也转换为向量 $q = (w_1, w_2, \mathrm{L}\ , w_t)$ 表示.

定义训练图像与查询图像之间的差异程度为:

$$S(d_i, q) = \left\| d_i - q \right\|_p \tag{16}$$

这里采用的是 2-范数. 比较查询图像与训练图像之间的差异程度 $S(d_i, q), (1 \le i \le N)$，选取差异程度最小的前 n 个作为查询结果返回.

## 3　实验结果与分析

为了验证本文提出的方法的图像检索效果，我们选取标准 Corel 库中 1000 张图片和 Caltech101 库中部分图片共 2400 多张图片作为图像检索库，图像大小为 384 x 256 像素，部分样图如图 3 所示.

图 3　部分训练集图像

待检索图像直接从图像库中选取，随机从图像数据库中选取不同类别的图像，每次在查询结果中将按相似度排序后前 10 幅图像作为检索结果. 检索结果示例如图 4 所示，每行为一次检索结果，每行 10 幅图像均为检索结果，由于待检索图像直接从图像库中选取，且检索结果按相似度排序，所以检索结果中的第一幅图像就是原待检索图像本身，从左到右按图片与待检索图像的相似度由高到低排序.

图 4　图像检索结果示例

首先对单个图像分别进行 SIFT 特征跟 ORB 特征的提取，每种特征分别提取 300 和 500 个特征点，由表 1 可以看出，在特征提取速度方面 ORB 算法的速度是明显比 SIFT 快得多.

表 1　特征提取时间对比

| 特征提取算法 | 提取特征数量 | 平均提取时间/ms |
|---|---|---|
| SIFT | 300 | 475.32 |
| SIFT | 500 | 577.84 |
| ORB | 300 | 35.69 |
| ORB | 500 | 39.54 |

为了验证本文算法的检索效果，实验以查准率作为评价标准，即检索结果中用户满意的图像数目与检索结果返回中所有图像数目之比. 同时为了准确衡量本文算法的检索效率，分别使用不同数量的视觉单词进行图像检索实验，最后计算平均查准率并计算平均检索时间. 实验结果数据如表 2 所示. 当视觉单词数量取值为 400 时，按类图像平均查准率如表 3 所示.

表 2　图像检索实验结果

| 视觉单词数量 | 平均查准率/% | 平均检索时间/ms |
|---|---|---|
| 100 | 64.25 | 179.13 |
| 200 | 65.01 | 256.93 |
| 300 | 69.46 | 337.95 |
| 400 | 68.34 | 465.72 |
| 500 | 72.58 | 560.28 |
| 600 | 73.95 | 655.80 |

表 3 按类别检索结果统计

| 视觉单词数量 | 平均查准率/% | 平均检索时间/ms |
|---|---|---|
| 100 | 64.25 | 179.13 |
| 200 | 65.01 | 256.93 |
| 300 | 69.46 | 337.95 |
| 400 | 68.34 | 465.72 |
| 500 | 72.58 | 560.28 |
| 600 | 73.95 | 655.80 |

由表 2 可以看出, 随着视觉单词数量的增加, 平均查准率越来越高, 但是平均检索时间也呈线性增长趋势. 结合表 1 中的实验结果数据可以看出, 仅仅是 SIFT 的特性提取阶段的耗时已相当于本文方法的平均检索时间. 虽然图像检索的查准率偏低, 但是检索时间快, 能够满足系统实时性的要求.

## 4 结语

本文提出了一种使用 ORB 特征的视觉词袋模型的快速图像检索的方法, 利用 ORB 特征替代 SIFT 对图像提取局部特征后进行聚类, 生成得到一个视觉单词“字典”, 然后对于每幅图像, 统计图像特征中各个视觉词汇出现的频数, 得到一个图像的描述向量, 并对向量进行归一化处理, 用该一维向量来表示图像, 其维数为视觉单词的数目. 进行图像检索时, 对待检索图像 ORB 特征, 经过视觉词袋的映射之后, 待检索图像也会用一个向量来表示, 通过计算该向量与图像库中的图像向量的欧式距离, 求取距离最小的图像, 即是与查询图像最相似的结果.

实验结果表明, 本文提出的方法在保持了传统视觉词袋模型算法的鲁棒性的同时, 由于采用了更加快速的二进制特征 ORB, 因此很大程度地缩短了图像检索时间, 提高了图像检索效率. 本文只是将 ORB 特征应用到视觉词袋模型中, 没有考虑图像的颜色特征, 在未来的工作中可以与图像的颜色特征相结合, 进一步提高图像检索的准确率.

## 参考文献


1 Sivic J. Video Google: A text retrieval approach to object matching in videos. Proc. of the International Conf. on Computer Vision. Nice, France. IEEE Press. 2003.

2 Lowe D. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004: 91–110.

3 Rublee E, Rabaud V, Konolige K, et al. ORB: An efficient alternative to SIFT or SURF. IEEE International Conference on Computer Vision(ICCV), 2011. IEEE. 2011. 2564–2571.

4 David L. Naive(Bayes) at forty: The independence assumption in information retrieval. European Conference on Machine Learning, 1998: 4–15.

5 Rosten E, Drummond T. Machine learning for high-speed corner detection. Computer Vision-ECCV 2006. Springer Berlin Heidelberg, 2006. 430–443.

6 Calonder M, Lepetit V, Strecha C, et al. Brief: Binary robust independent elementary features. Computer Vision-ECCV 2010, 2010: 778–792.

7 Grana C, Borghesani D, Manfredi M, et al. A fast approach for integrating ORB descriptors in the bag of words model. IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2013: 866709–866709-8.

8 Mansoori NS, Nejati M, Razzaghi P, et al. Bag of visual words approach for image retrieval using color information. 2013 21st Iranian Conference on Electrical Engineering (ICEE). IEEE. 2013. 1–6.

9 黄超,刘利强,周卫东.改进的二进制特征图像检索算法.计算机工程与应用,2015,14:23–27.

10 霍华,赵刚.基于改进视觉词袋模型的图像标注方法.计算机工程,2012,22:276–278,282.

11 Mansoori NS, Nejati M, Razzaghi P, et al. Bag of visual words approach for image retrieval using color information. 2013 21st Iranian Conference on Electrical Engineering (ICEE). IEEE. 2013. 1–6.

12 董坤,王倪传.基于视觉词袋模型的人耳识别.计算机系统应用,2014,23(12):176–181.

13 Zhu L, Jin H, Zheng R, et al. Weighting scheme for image retrieval based on bag-of-visual-words. Image Processing, IET, 2014, 8(9): 509–518.