# 改进 *k*-means 算法在图像标注和检索中的应用

潘 崇,朱红斌
PAN Chong,ZHU Hong-bin

丽水学院 计算机与信息工程学院,浙江 丽水 323000
Computer and Information Engineering College of Lishui University,Lishui,Zhejiang 323000,China
E-mail:lspanchong@126.com

**PAN Chong,ZHU Hong-bin.Application of improved *k*-means algorithm in image annotation and retrieval.Computer Engineering and Applications,2010,46(4):183-185.**

**Abstract**: A novel method of image annotation and retrieval based on improved *k*-means algorithm is proposed.First,image segmentation is applied to the training image;an improved *k*-means algorithm is proposed to cluster the segmented region.The improved *k*-means algorithm first uses the genetic clustering algorithm to determine the number of cluster *k*,and then chooses the clustering centers.At the time of image annotation,the correlation between image semantic concept and region cluster is established,and then it is used as a priori knowledge and is combined with the lower feature of the region to annotate the unannotated image.The system has been tested on an image database of about 1 000 images.The experiment results show that the proposed approach has a good retrieval performance.

**Key words**: image segmentation;genetic algorithm;*k*-means algorithm;image annotation;image retrieval

**摘 要**:提出一种基于改进的 *k*-means 算法的图像标注和检索方法。首先对训练图像进行分割,采用改进的 *k*-means 算法对分割后的区域进行聚类。改进的 *k*-means 算法首先采用遗传聚类算法确定聚类数 *k*,然后对聚类中心进行选择。在图像标注时,首先通过已标注的图像求出语义概念和聚类区域的关联度,用它作为待标注图像的先验知识,然后结合区域的低层特征,对未标注的图像进行标注。在一个包含 1 000 幅图像的图像库进行实验,采用标注的语义关键字进行检索,结果表明,提出的方法是有效的。

**关键词**:图像分割;遗传算法;*k*-means 算法;图像标注;图像检索
**DOI**:10.3778/j.issn.1002-8331.2010.04.058 **文章编号**:1002-8331(2010)04-0183-03 **文献标识码**:A **中图分类号**:TP391.41

## 1 引言

图像标注就是使用语义关键字对图像的语义内容进行描述,进而将图像检索转变成一个文本检索的问题。采用手工标注的方法在某些场合是有效的,但是采用手工方法用关键字标注图像语义,由于用户对图像的理解不同,这样会存在标注的主观性和不确定性,同时由于图像库越来越大时,完全使用手工标注,工作量太大[1-2]。

对于图像语义自动标注,目前主要的方法是基于视觉特征的分类方法,通常将具有相同视觉特征的区域归为一类,即使区域的语义完全不同,也用相同的关键字标注[3-5]。陈世亮[1]提出一种基于区域特征关联的图像语义标注方法,采用贝叶斯模型从已标注图像语义推导待标注图像的语义信息,在一定程度上缩小了低层特征和高层语义之间的语义鸿沟,但是它训练时没有对分割后的区域进行聚类,没有充分利用已经标注样本的先验知识,所以它标注的精确度也比较低。

提出一种改进的 *k*-means 算法的图像标注和检索方法,首先对训练图像进行分割,采用改进的 *k*-means 算法对分割后的区域进行聚类。改进的 *k*-means 算法对聚类中心进行选择和采用遗传聚类算法确定聚类数 *k*。在图像标注时,首先通过已标注的图像求出语义概念和聚类区域的关联度,用它作为待标注图像的先验知识;然后结合区域的低层特征,对未标注的图像进行标注;接着以标注的关键字对图像数据库进行检索。

## 2 基于目标区域的特征提取

为了获取有意义的目标区域,必须进行图像分割,采用文献[6]提出的一种基于颜色与空间的图像分割方法,该方法可以较好地实现图像中有意义的目标区域的提取,通过图像分割得到一些有意义的目标区域。假设通过图像分割,得到 $n$ 个有意义的图像区域,用 $\{r_i\}_{i=1}^{n}$ 表示,分别提取归一化的颜色、位置、形状特征。

### 2.1 颜色特征

采用 CIE L*a*b 的颜色空间的 3 个亮度分量作为颜色特征即 $\boldsymbol{I}(\boldsymbol{p})=[I_L(\boldsymbol{p}),I_a(\boldsymbol{p}),I_b(\boldsymbol{p})]^{\mathrm{T}}$,提取区域 $r_i$ 的 $L$、$a$、$b$ 分量的特征,分别用 $f_{i1}$、$f_{i2}$、$f_{i3}$ 表示如下:

$$f_{i1}=\frac{1}{100\cdot A_i}\sum_{\boldsymbol{p}\in r_i}I_L(\boldsymbol{p}),\ f_{i2}=\frac{\dfrac{1}{A_i}\sum_{\boldsymbol{p}\in r_i}I_a(\boldsymbol{p})+80}{200},$$

**作者简介**:潘崇(1978-),硕士,研究方向:图形图像,嵌入式系统;朱红斌(1979-),男,硕士,研究方向:图像处理和模式识别。
**收稿日期**:2008-08-12 **修回日期**:2009-08-06

$$f_{i3}=\frac{\frac{1}{A_i}\sum_{\boldsymbol{p}\in r_i} I_b(\boldsymbol{p})+80}{200} \tag{1}$$

其中，$\boldsymbol{p}=[p_x,p_y]^{\mathrm{T}}$ 表示像素点的位置坐标，$p_x=1,\cdots,x_{\max}$，$p_y=1,\cdots,y_{\max}$，$A_i$ 为区域 $r_i$ 中像素的个数，$x_{\max}$、$y_{\max}$ 为图像区域的尺寸。

### 2.2 位置特征

提取归一化的区域空间中心作为区域图像的位置特征，分别用 $f_{i4}$、$f_{i5}$ 表示。

$$f_{i4}=\frac{1}{A_i\cdot x_{\max}}\sum_{\boldsymbol{p}\in r_i} p_x,\ f_{i5}=\frac{1}{A_i\cdot y_{\max}}\sum_{\boldsymbol{p}\in r_i} p_y \tag{2}$$

### 2.3 形状特征

提取归一化的形状大小和离心率作为区域图像的形状特征，分别用 $f_{i6}$ 和 $f_{i7}$ 表示。

$$f_{i6}=\frac{A_i}{x_{\max}\cdot y_{\max}} \tag{3}$$

为了计算区域 $r_i$ 的离心率，定义协方差矩阵 $\boldsymbol{C}_i$：

$$\boldsymbol{C}_i=\frac{1}{A_i}\sum_{\boldsymbol{p}\in r_i}(\boldsymbol{p}-\boldsymbol{S}(r_i))(\boldsymbol{p}-\boldsymbol{S}(r_i))^{\mathrm{T}} \tag{4}$$

其中，$\boldsymbol{S}(r_i)=\frac{1}{A_i}\sum_{\boldsymbol{p}\in r_i}\boldsymbol{p}$ 为区域的中心。定义 $\rho_k$、$\boldsymbol{u}_k(k=1,2)$分别为 $\boldsymbol{C}_i$ 的特征值和特征向量，即：$\boldsymbol{C}_i\boldsymbol{u}_k=\rho_k\boldsymbol{u}_k(\rho_1\geqslant\rho_2)$，特征向量 $\boldsymbol{u}_1$ 为区域的方向，而 $\boldsymbol{u}_2$ 与 $\boldsymbol{u}_1$ 正交，区域的离心率 $\varepsilon_i=1-\frac{\rho_1}{\rho_2}$，对它归一化后得到离心率 $f_{i7}$：

$$f_{i7}=\mathrm{e}^{\varepsilon_i} \tag{5}$$

得到区域 $r_i$ 的特征向量为 $\boldsymbol{f}_i=(f_{i1},f_{i2},\cdots,f_{i7})$。

## 3 基于改进 $k$-means 算法的区域聚类

传统的 $k$-means 算法的缺陷有：(1)过分依赖于初始的聚类中心的选择；(2)最佳的聚类数 $k$ 的确定，$k$ 为多少时聚类效果最好；(3)目标函数的选择，如果目标函数选得不好，可能聚类效果很差。这里对前 2 个缺点进行改进，采用一种新的聚类中心选择算法对 $k$-means 算法的第一个缺陷进行改进，采用遗传聚类算法确定聚类数 $k$。

### 3.1 初始聚类中心的选择

对图像区域进行聚类，最根本的目的是使同一个聚类中的区域之间的相似度高，而不同聚类中的区域之间的相似度低。如果采用距离表示区域之间的相似度，相似的区域之间的距离比不相似的区域之间的距离要小。由于传统的 $k$-means 聚类算法的性能过分依聚类中心的选择，初始聚类中心的选择直接影响聚类的效果，不同的初始聚类中心会产生不同的聚类结果，初始聚类中心选择得不好，聚类算法可能不收敛。文中采用一种新的初始聚类中心的计算方法，采取下列步骤计算初始聚类中心：

设 $\chi=\{r_i\}_{i=1}^{n}$ 表示分割后得到的 $n$ 个图像区域，首先计算每个图像区域两两之间的距离；找出距离最近的两个区域，形成一个区域集合 $A_1$，并将它们从总的区域集合 $\chi$ 中删除；然后计算 $A_1$ 中每一个区域与集合 $\chi$ 中每一个样本的距离，找出在 $\chi$ 中与 $A_1$ 中最近的区域，将它并入集合 $A_1$ 并从 $\chi$ 中删除，直到 $A_1$ 中的区域个数达到一定阈值；再从 $\chi$ 中找到两两间距离最近的两个区域构成 $A_2$，重复上面的过程，直到形成 $k$ 个聚类集合；最后对 $k$ 个区域集合分别进行算术平均，形成 $k$ 个初始聚类中心。详细的算法如下：

(1)计算任意两个图像区域间的距离 $d(x,y)$，找到集合 $\chi$ 中距离最近的两个图像区域，形成集合 $A_m(1\leqslant m\leqslant k)$，并从集合 $\chi$ 中删除这两个区域；

(2)在 $\chi$ 中找到距离集合 $A_m$ 最近的区域，将其加入集合 $A_m$，并从集合 $\chi$ 中删除该区域；

(3)重复第 2 步直到集合中的区域对象个数大于等于 $\alpha\times n/k$ $(0<\alpha\leqslant 1)$；

(4)如果 $m<k$，则 $m\leftarrow m+1$，再从集合 $\chi$ 中找到距离最近的两个图像区域，形成新的集合 $A_m(1\leqslant m\leqslant k)$，并从集合 $\chi$ 中删除这两个区域，返回第 2 步执行；

(5)将最终形成的 $k$ 个集合中的区域分别进行算术平均，从而形成 $k$ 个初始聚类中心；

(6)从这 $k$ 个初始聚类中心出发，应用 $k$-均值聚类算法形成最终聚类。

### 3.2 遗传聚类算法

聚类数 $k$ 的选择也是影响聚类质量的一个重要参数。当样本数量很小的时，利用穷举法可以寻找最佳的聚类数，当样本数量很大时，穷举几乎是不可能的。当样本数目很大的时，聚类数的确定是随机的。为了获得高精度的聚类结果，采用改进的遗传聚类混合算法，该方法使用遗传算法中的特殊适应度函数不仅能自动搜索到最佳聚类数，而且能对 $k$-means 算法得到的聚类进一步优化。实验结果表明采用遗传算法和初始聚类中心选择的有效结合使聚类精度获得明显改善。

(1)编码：对聚类数 $k$ 编码。$k$ 是介于 1 和最大类别数 *Max-Classnum* 之间的一个整数，这个整数可以用二进制表示，即为染色体。文献[7]已经证明，当 $Max\text{-}Classnum\leqslant\sqrt{n}$（$n$ 为样本个数）时，空间聚类达到优化。例如样本数量 $n$=3 600，则 $Max\text{-}Classnum<60$，此时染色体长度为 8，前 6 位基因是类别数的二进制表示，第 7 位是类别数对应的十进制，最后一位是个体对应的适应度。

(2)初始化种群：随机生成含 $P$ 个染色体的种群，$P$ 取值一般为[30,150]，交叉概率 $P_c$=0.65，变异概率 $P_m$=0.08。

(3)聚类：将群体中的每个染色体编码，得到对应的类别数 $k$。对于每个个体，应用 $k$-means 聚类算法。

(4)适应度函数的计算

$$f(k)=\frac{1}{k}\times\frac{1}{E_k}\times D_k \tag{6}$$

其中，$E_k=\sum_{j=1}^{k}\sum_{i\in I_j}\|z_i-z_j\|^2$ 表示类内距，$k$ 表示类别数，$I_j$ 为 $j$ 类的样本集合，$z_j$ 为 $j$ 类的类中心。$D_k=\max_{i,j=1}^{k}\|z_i-z_j\|^2$ 表示类间距，其中 $z_i$、$z_j$ 分别表示 $i$ 类和 $j$ 类的类中心。

(5)遗传操作：选择算法采用轮盘赌算法，按一定概率选择那些适应度值高的个体进入下一代群体，交叉算子采用单点交叉，变异算子采用单点变异[7]。在交叉变异后判断个体对应的类别数是否小于等于 *Max-Classnum*，淘汰那些类别数大于 *Max-Classnum* 的个体。

(6)终止准则：如果种群中最优适应度值不再发生变化，则停止，否则转入步骤(3)迭代。

## 4 图像语义标注和检索

由于图像的语义大都通过图像区域来体现，考虑训练图像

的先验知识和区域的低层特征，首先通过已标注的图像求出语义概念和聚类区域的关联度，用它作为待标注图像的先验知识，在图像标注时将待标注图像的区域特征与聚类区域的特征进行比较，实现待标注图像语义的自动标注，然后以标注的关键字进行检索。

### 4.1　语义概念与聚类区域关联度

通过图像分割和区域聚类后，得到 $k$ 个聚类 $R=\{R_1,R_2,\cdots,R_k\}$，定义 $C=\{c_1,c_2,\cdots,c_W\}$表示图像标注的关键字的集合。为了确定语义关键字和聚类之间的联系，首先建立一个概率表[8]，假设图像库中有 $N$ 幅训练图像，总共有 $W$ 个语义关键字，$k$ 个图像聚类区域$(R_1,R_2,\cdots,R_k)$。

然后用两个矩阵 $\boldsymbol{M}_{N\times W}$ 和 $\boldsymbol{M}_{N\times k}$ 表示数据集，在矩阵 $\boldsymbol{M}_{N\times W}$ 中，行 $N$ 表示图像的个数，$W$ 列表示 $W$ 个语义关键字，$\boldsymbol{M}_{N\times W}[i,j]$ 表示第 $j$ 个语义关键字出现在第 $i$ 幅图像中的权值。在矩阵 $\boldsymbol{M}_{N\times k}$ 中，行 $N$ 表示图像的个数，$k$ 列表示 $k$ 个聚类。$\boldsymbol{M}_{N\times k}[i,j]$为第 $j$ 个聚类出现在第 $i$ 幅图像中的权值。

下面的问题是确定 $\boldsymbol{M}_{N\times W}[i,j]$和 $\boldsymbol{M}_{N\times k}[i,j]$中每一项的值，由于这两个矩阵中的每一项的值的计算方法类似，下面只针对 $\boldsymbol{M}_{N\times k}[i,j]$进行计算。假设 $w_{il}$ 为图像 $l$ 中聚类 $R_i$ 的权值，设 $N$ 为总的图像的数目，$n_i$ 为在聚类 $R_i$ 中出现的图像的数目。定义归一化频率 $tf_{il}$：

$$tf_{il}=\frac{freq_{il}}{\max_h freq_{hl}} \tag{7}$$

其中，$freq_{il}$ 是聚类 $R_i$ 在图像 $l$ 中出现的次数，$\max_h freq_{hl}$ 是聚类 $\{R_i\}(i=1,2,\cdots,k)$在图像 $l$ 中出现的最大次数。假设当一个聚类出现在大多数图像中，则认为这个聚类是不重要的。下面为 $R_i$ 定义一个倒排文档频率 $idf_i$：

$$idf_i=\log\frac{N}{n_i} \tag{8}$$

为了平衡上面的两个因素，得到聚类区域的权值 $w_{il}$：

$$w_{il}=tf_{il}\times idf_i=tf_{il}\times\log\frac{N}{n_i} \tag{9}$$

应用相似的方法确定 $\boldsymbol{M}_{N\times W}[i,j]$每一项的权值。然后将 $\boldsymbol{M}_{N\times W}[i,j]$的转置矩阵和 $\boldsymbol{M}_{N\times k}[i,j]$相乘，即 $\boldsymbol{M}_{N\times W}[i,j]^{\mathrm{T}}\times\boldsymbol{M}_{N\times k}[i,j]$，得到一个 $W\times k$ 维的概率矩阵，然后对矩阵中的每一列都归一化，这样就得到一个概率表 $P_{coor}$，它表示关键字和聚类区域的关联度。$P_{corr}[i,j]$是给定聚类区域 $R_j$ 的条件下，关键字 $c_i$ 的条件概率，即 $p(c_i|R_j)$。

### 4.2　图像标注与检索

首先对待标注的图像 $I'$进行分割，得到区域的集合 $r'=\{r_1',r_2',\cdots,r_k'\}$，然后对所有的 $r_j'\in r'$，计算条件概率 $p(c_i|r_j')(i=1,2,\cdots,W)$，以 $p(c_i|r_j')$最大时的 $c_i$ 作为图像区域的语义标注。

给定区域 $r_j'$和候选语义关键字 $c_i\in C$，条件概率 $p(c_i|r_j')$计算如下：

$$p(c_i|r_j')=\sum_{k=1}^{K}(sim(r_j',R_k)\times p(c_i|R_k)) \tag{10}$$

其中，先验概率 $p(c_i|R_k)$是 4.1 节中已经求出的语义概念和聚类区域的关联程度，$sim(r_j',R_k)$表示待标注区域 $r_j'$与已标注聚类 $R_k$ 低层特征的相似性。选择$\arg\max_{c_i} p(c_i|r_j')$的关键字 $c_i$ 标注该图像区域的语义。

## 5　实验结果

该系统在 Intel P4-2.8 GHz CPU 的微机，Windows XP 平台下用 Visual C++6.0 作为开发环境，图像库中收集了近 1 000 幅图片，包括有风景、汽车、动物、花等 10 类主要的图片，每类有 100 幅左右。选择每类中的 50 幅图像作为训练样本，剩下的 50 幅图像作为测试样本。对已标注的训练样本进行分割，对分割得到的 2 012 个区域采用改进的 $k$-means 算法进行聚类，对未标注的样本采用 4.2 节中的算法进行标注，整个数据集共有 22 个不同的关键字。在实验中，将该文提出的方法和文献[3,5]进行对比实验，结果如图 1 所示。从图 1 中可见提出的方法比文献[3,5]的方法标注的正确率高。

| image |  |  |  |  |
|---|---|---|---|---|
| 文献[3] | flowers grass leaf | horse trees grass | grass water building | elephant leaf field grass |
| 文献[5] | flowers grass trees | horse trees grass elephant | grass water mountains | elephant leaf water grass |
| 该文方法 | flowers leaf | horse leaf grass | sky building | elephant trees field grass |

**图 1　不同标注方法的结果比较**

对检索算法的性能评价的重要指标是查准率(Precision)、查全率(Recall)和 $F1$ 测度值。$F1$ 测度值是综合考虑了查全率和查准率的性能评价指标，$F1$ 测度值用公式表示如下：

$$F1=\frac{2\times recall\times precision}{recall+precision} \tag{11}$$

下面对传统的 $k$-means 算法和改进的 $k$-means 算法的标注性能进行比较，同样以标注的关键字进行检索，以查全率、查准率和 $F1$ 测度值来衡量系统的检索性能，实验结果如表 1 所示。从表 1 中可以看出，采用改进的 $k$-means 算法后，查全率从 0.34 上升到 0.43，上升率为 26.47%，查准率从 0.36 上升到 0.40，上升率为 11.11%，$F1$ 测度值从 0.350 上升到 0.414，上升率为 18.29%。实验结果表明对 $k$-means 算法进行改进后，标注更加精确，检索性能得到了提高。

**表 1　不同聚类算法的标注性能比较**

| 不同的聚类方法 | 查全率 | 查准率 | $F1$ 测度值 |
|---|---|---|---|
| 传统的 $k$-means 算法 | 0.34 | 0.36 | 0.350 |
| 改进的 $k$-means 算法 | 0.43 | 0.40 | 0.414 |

## 6　结论

提出基于 $k$-means 算法的图像自动标注与检索方法，采用初始聚类中心选择和遗传算法对 $k$-means 算法进行改进，实验结果该文方法具有更高的精确度，和传统的 $k$-means 算法比较，改进的算法具有更好的检索性能，系统具有较好的鲁棒性。

## 参考文献：

[1] 陈世亮，李战怀，袁柳.一种基于区域特征关联的图像语义标注方法[J].计算机工程与应用，2007，43(2)：53-55.

[2] 芮晓光，袁平波，何芳.一种新的基于语义聚类和图算法的自动图像标注方法[J].中国图象图形学报，2007，12(2)：239-245.

[3] 路晶，马少平.基于概念索引的图像自动标注[J].计算机研究与发展，2007，44(3)：452-459.

(下转 188 页)