中图法分类号：TP391.41 文献标志码：A 文章编号：1006-8961(2011)04-0566-06
论文索引信息：朱峰，张晓娜，陈健美，刘哲. 基于多阶抽样的高斯混合模型彩色图像分割[J]. 中国图象图形学报,2011,16(4):566-571

基于多阶抽样的高斯混合模型彩色图像分割

朱峰，张晓娜，陈健美，刘哲
（江苏大学理学院，镇江 212013）

摘 要：针对传统高斯混合模型应用于彩色图像分割时计算复杂度高等问题，提出一种多阶抽样的高斯混合模型的彩色图像分割算法。首先，给出采样数定理及其证明，并推导出与聚类类别数和最小聚类相关的最小采样数目；其次，设计一罚函数判断抽样优劣，消除抽样对聚类模型影响，根据最小采样数数目，对像素点进行均匀采样，并利用高斯混合模型对采样像素点进行聚类；最后，定义像素点和类之间的距离，对剩余的像素点按距离最近原则进行划分。实验结果表明算法具有有效性。

关键词：图像分割；高斯混合模型；多阶抽样

Color image Segmentation based on Gaussian mixture model with muti-sampling

Zhu Feng, Zhang Xiaona, Chen Jianmei, Liu Zhe
(*Faculty of Science, Jiangsu University, Zhenjiang* 212013 *China*)

Abstract: The application of classical Gaussian mixture model to image segmentation has highly computational complexity. A image segmentation method based on Gaussian mixture model with multi-sampling is proposed in order to solve this problem. First, the sampling theorem is given and proved, and the minimum sample size is derived according to the smallest cluster and cluster number. Second, a penalty function, which is to judge the good sample, has been designed to eliminate the error of clustering model, and image pixels are sampled based on the minimum sample size to be clustered according to Gaussian mixture model. Finally, by the means of the definition on the distance between a pixel point and the categories, the remaining points is assigned respective cluster depending on the principles of the nearest distance. The experimental results show the effectiveness of the algorithm.

Keywords: image segmentation; Gaussian mixture model; muti-sampling

0 引 言

图像分割是将图像分成具有不同特征的不相交的连通子区域的过程，其结果对图像分析和图像理解具有重要意义。聚类是一种常用的图像分割技术，多年来已经提出了许多基于聚类的图像分割算法，例如 K 均值[1]、模糊 C 均值[2]、核密度聚类[3]、有限混合模型[4]等。其中有限混合模型是一种基于模型的统计聚类方法，由于它属于一种半参数的密度估计方法，融合了参数估计和非参数估计的优点，而且不局限于特定的概率密度函数的形式，因而，它在许多领域得到了广泛应用[5-7]。

如果在高斯混合模型中选取多元高斯分布作为分量密度函数，它可直接用于 3 通道彩色图像的分割，而不需要将彩色图像转变为相应的灰度图像进行分割或者将彩色图像空间变换成 3 个相互独立的分量，分别进行处理，但是模型的复杂度与所聚类的

收稿日期：2009-12-10；修回日期：2010-01-11
基金项目：国家自然科学基金项目(60841003)。
第一作者简介：朱峰(1977—)，男，讲师。主要研究领域为图像处理与识别。E-mail：zhufe@ujs.edu.cn。

对象像素及其维数相关,一幅 512×512 彩色图像,时间复杂度将达到 O(512×512×3)。据此,本文提出基于多阶抽样的高斯混合模型彩色图像分割法。通过设计一罚函数判断抽样优劣,选择最具有代表性的样本进行聚类训练,得出最终的高斯混合模型参数近似估计,解决了高斯混合应用于彩色图像分割时计算机复杂度高的问题,而且对图像噪声也有一定的鲁棒性。

1　传统高斯混合模型

设 x_i 为一幅图像第 i 个像素的观察值,如灰度、纹理等,并且 $x_i(i=1,2,\cdots,N)$ 满足独立同分布(IID)的特性。另设该图像含有 K 个目标类,每个目标类都服从已知概率分布,记为 $\phi_j(x|\theta_j)$,该分布由参数集 θ_j 确定。给出所有类的参数集,可以得到每个像素由各个类合成概率分布式为

$$p(x_i|\boldsymbol{\Theta})=\sum_{j=1}^{K}\pi_j\phi_j(x_i|\theta_j) \tag{1}$$

式中,π_j 为混合系数,该表达式所含参数集为 $\boldsymbol{\Theta}=(\pi_1,\cdots,\pi_K,\theta_1,\cdots,\theta_K)$,同时满足限制条件 $\sum_{j=1}^{K}\pi_j=1$。一般情况下,$\phi_j(x|\theta_j)$ 取均值为 μ_j,协方差为 $\boldsymbol{\Sigma}_j$ 的高斯分布,其分布概率表达式为

$$\phi_j(x|\theta_j)=\phi_j(x|\mu_j,\boldsymbol{\Sigma}_j)=\frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}_j|}}\mathrm{e}^{-(x-\mu_j)^{\mathrm{T}}\boldsymbol{\Sigma}_j^{-1}(x-\mu_j)} \tag{2}$$

如果式(1)中的 $\phi_j(x|\theta_j)$ 服从高斯分布,那么式(1)即为高斯混合分布模型(GMM),该模型完全有其参数确定,被广泛用于数据分类。最大似然估计(ML)是常用求解概率分布参数的方法,上述观测数据 X 的对数似然表达式为

$$\ln(L(\boldsymbol{\Theta}|X))=\ln\prod_{i=1}^{N}p(x_i|\boldsymbol{\Theta})=\sum_{i=1}^{N}\ln\left(\sum_{j=1}^{K}\pi_j\phi_j(x_i|\theta_j)\right) \tag{3}$$

目前直接利用最大似然估计法求解式(3)是很困难的,因为其表达式含有对数和。Dempster 等人[8]提出的 EM 算法是解决该问题的一个很好的方法。EM 算法求解参数第 $t+1$ 次迭代公式如下

$$\pi_k^{(t+1)}=\frac{1}{N}\sum_{i=1}^{N}p(k|x_i,\boldsymbol{\Theta}^{(t)}) \tag{4}$$

$$\mu_k^{(t+1)}=\frac{\sum_{i=1}^{N}x_i\cdot p(k|x_i,\boldsymbol{\Theta}^{(t)})}{\sum_{i=1}^{N}p(k|x_i,\boldsymbol{\Theta}^{(t)})} \tag{5}$$

$$\boldsymbol{\Sigma}_k^{(t+1)}=\frac{\sum_{i=1}^{N}p(k|x_i,\boldsymbol{\Theta}^{(t)})(x_i-\mu_k^{(t+1)})(x_i-\mu_k^{(t+1)})^{\mathrm{T}}}{\sum_{i=1}^{N}p(k|x_i,\boldsymbol{\Theta}^{(t)})} \tag{6}$$

式中,混合系数 π_k 作为先验概率,像素 x_i 属于每一个类的概率 $p(k|x_i,\boldsymbol{\Theta}^{(t)})$ 可由贝叶斯公式计算如下

$$p(k|x_i,\boldsymbol{\Theta}^{(t)})=\frac{\pi_k^{(t)}\phi_k(x_i|\theta_k^{(t)})}{p(x|\boldsymbol{\Theta}^{t})}=\frac{\pi_k^{(t)}\phi_k(x_i|\theta_k^{(t)})}{\sum_{j=1}^{K}\pi_j^{(t)}\phi_j(x_i|\theta_j^{(t)})} \tag{7}$$

2　基于多阶抽样的高斯混合模型图像分割

2.1　相关概念

定义 1　(点与类之间距离)设像素点 $x\in\boldsymbol{V}$,类 $A\subset\boldsymbol{V}$,定义 x 与 A 之间的距离为

$$D(x,A)=\sqrt{(x-\mu_A)^{\mathrm{T}}\boldsymbol{\Sigma}_A^{-1}(x-\mu_A)} \tag{8}$$

式中,μ_A 和 $\boldsymbol{\Sigma}_A$ 为类 A 的均值与协方差矩阵。

定义 2　(类间与类内离差平方和)设样本总体 Ω 可分为 K 类,则它的类间离差平方和 D_A 和类内离差平方和 D_E 定义为

$$D_A=\sum_{j=1}^{K}\sum_{i=1}^{n_j}(\bar{x}_j-\bar{x})^{\mathrm{T}}(\bar{x}_j-\bar{x}) \tag{9}$$

$$D_E=\sum_{j=1}^{K}\sum_{i=1}^{n_j}(x_{ij}-\bar{x}_j)^{\mathrm{T}}(x_{ij}-\bar{x}_j) \tag{10}$$

式中,$\bar{x}_j$ 为第 j 类的均值,$\bar{x}$ 为总体均值,x_{ij} 为第 j 类第 i 个元素,n_j 为第 j 类的样本总数。

定理　对于一个聚类 u,如果取样本大小 s 满足

$$s\geqslant fN+\frac{N}{|u|}\ln\left(\frac{1}{\delta}\right)+\frac{N}{|u|}\sqrt{\left(\ln\left(\frac{1}{\delta}\right)\right)^2+2f|u|\ln\left(\frac{1}{\delta}\right)} \tag{11}$$

那么样本中属于聚类 u 的点的个数小于 f。$|u|$ 的概率小于 δ,其中 $0<\delta<1$,$0<f<1$,$|u|$ 为聚类 u 的大小,N 为全体数据集的大小。

证明　对 $P(X<f|u|)<\delta$ 进行变形得

$$P\left(X < \left(1 - \left(1 - \frac{f|u|}{\mu}\right)\right)\mu\right) < \delta$$

由 Chernoff 界[9]

$$P(X < (1-\varepsilon)\mu) < \mathrm{e}^{-\frac{\mu\varepsilon^2}{2}}$$

结论可知,上式成立的条件是

$$\mathrm{e}^{\frac{-\mu\left(1-\frac{f|u|}{\mu}\right)^2}{2}} \leqslant \delta$$

令 $\mu = \frac{s|u|}{N}$,上式关于 s 的解即得证(证毕)。

从定理可以看出,样本至少有 $f|u|$ 点以很大概率属于聚类 u。如果假设 $u_{\min}$ 是所有聚类中最小的一个聚类,$s_{\min}$ 是式(11)的右边以 $|u_{\min}|$ 替代 $|u|$ 的结果,显然,在式(11)中,当 $u = u_{\min}$ 时,$|u| \geqslant |u_{\min}|$。所以当样本大小为 $s_{\min}$ 时,可以保证以 $1-\delta$ 的概率使得任意聚类 u 包含 $f|u|$ 个点。然而,式(11)中最小采样数并不是随着 N 增加而增加,而是随着聚类个数 k 增加而增加,这是因为那些小的聚类不是我们需要的,而可以看作孤立点。因此,假设最小聚类为平均聚类数据的某个比例,即 $|u_{\min}| = \frac{N}{k\rho}, \rho > 1$,另设最小聚类至少有 ξ 个数据点,则有 $f = \frac{\xi}{|u_{\min}|}$,可得定理的推论。

推论 设数据点数目 N 大于聚类数 k 且最小聚类至少有 ξ 个数据点,那么独立于原始数据点个数的最小取样数目为

$$s_{\min} = \xi k\rho + k\rho\ln\left(\frac{1}{\delta}\right) + \sqrt{\ln\left(\frac{1}{\delta}\right)^2 + 2\xi\ln\left(\frac{1}{\delta}\right)\rho} > 1 \quad (12)$$

2.2 算法描述

初始条件对高斯混合聚类算法非常重要,设计一罚函数作为采样好坏的评价标准,选择最具有代表性的样本进行聚类训练,该罚函数具体定义如下

$$F = \frac{D_E/(N-K)}{D_A/K} \quad (13)$$

式中,D_E 为类内离差平方和(式(10)),D_A 为类间离差平方和(式(9)),N 为样本容量,K 为聚类类别数。理想的情况是,类内离差平方和 D_E 越小越好,类间离差平方和 D_A 越大越好,所以 F 值应该越小越好。

依据提出基于多阶抽样高斯混合模型,实现图像分割的算法描述如下:

输入 像素观测值 $x_i(i=1,2,\cdots,N)$,分割数 k;

输出 分类标签 $x_i^j(i=1,2,\cdots,N; j=1,2,\cdots,K)$。

1)利用式(12)确定最小抽样数 $s_{\min}$,从图像 I 中不放回地随机抽取样本 I_s;

2)对样本 I_s 使用 K 均值方法初始化参数 $\Theta^{(0)}$;

3)计算 F 值,如果 $F < Thresh$,转到 4),否则转到 1);

4)根据式(4)—(6)计算新的参数迭代更新;

5)令 $t = t+1$,重复步骤 4)—5),直到 $|\ln(L(\Theta^{(t+1)}|X) - \ln(L(\Theta^{(t)}|X) < \varepsilon$ 中止;

6)据定义 1 对剩余的像素计算点与各聚类类别距离,按距离最近的原则将该像素划分到已知聚类中。

3 实验结果与分析

为了验证论文提出的基于多阶抽样高斯混合模型彩色图像分割法,采用抽样与非抽样两种方式的高斯混合模型及其 FCM 算法进行分割比较实验,实验中设置相同的参数与初始条件。模拟数据计算最小采样数的式(12)中的参数设定:$\delta = 0.1, \rho = 2, \xi = 10$,真实图像数据计算最小采样数的式(12)中的参数设定:$\delta = 0.1, \rho = 2, \xi = 20$。设定参数 $\varepsilon = 0.1$,参数 *Thresh* 通过重复实验获取。本实验使用 Dell 公司图像专用工作站,CPU 为 Intel(2)nl(TM),速度 1.5 GHz,内存 1 GB,硬盘 120 GB,MATLAB 7.0 作为实验开发工具。

3.1 抽样阈值 *Thresh* 的选取及抽样分析

在基于多阶抽样的高斯混合模型图像分割算法中,判断抽样数据优劣的阈值 *Thresh* 的确定非常重要,因为它直接影响算法的效率。根据式(12)计算最小采样数,然后进行抽样,使用 K 均值方法对上述两中数据进行分类,并根据公式(13)计算 F 值,重复进行实验 100 次。对两种重叠与不重叠模拟数据[10]进行重复实验,通过实验结果确定阈值 *Thresh* 的取值。

图 1(a)是具有 3 类的不重叠的模拟数据,数据大小为 1 800,(b)是具有 4 类的不重叠的模拟数据,数据大小为 2 000。(c)(d)为实验结果。从(c)(d)可知,F 值主要集中在 0 ~ 0.01 内,而 F 值大于 0.01 的出现的次数较少,这种现象与抽样原理是相符的。根据实验结果,选取抽样阈值 *Thresh* = 0.01。

选取美国加州大学伯克利分校的 BSDB 标准测试图像数据库[11]的两幅真实图像,图像标号为55 067,118 035,分别进行 100 次样本选取重复实验。图 2 为利用上面确定的阈值进行样本选取的迭代次数柱状图,从图 2 可以看出,大部分情况下,只需要 1—2 次的初始样本选取,就可以满足设定条件。样本的选取并未占用太多的时间,可以达到效率与质量的兼顾。

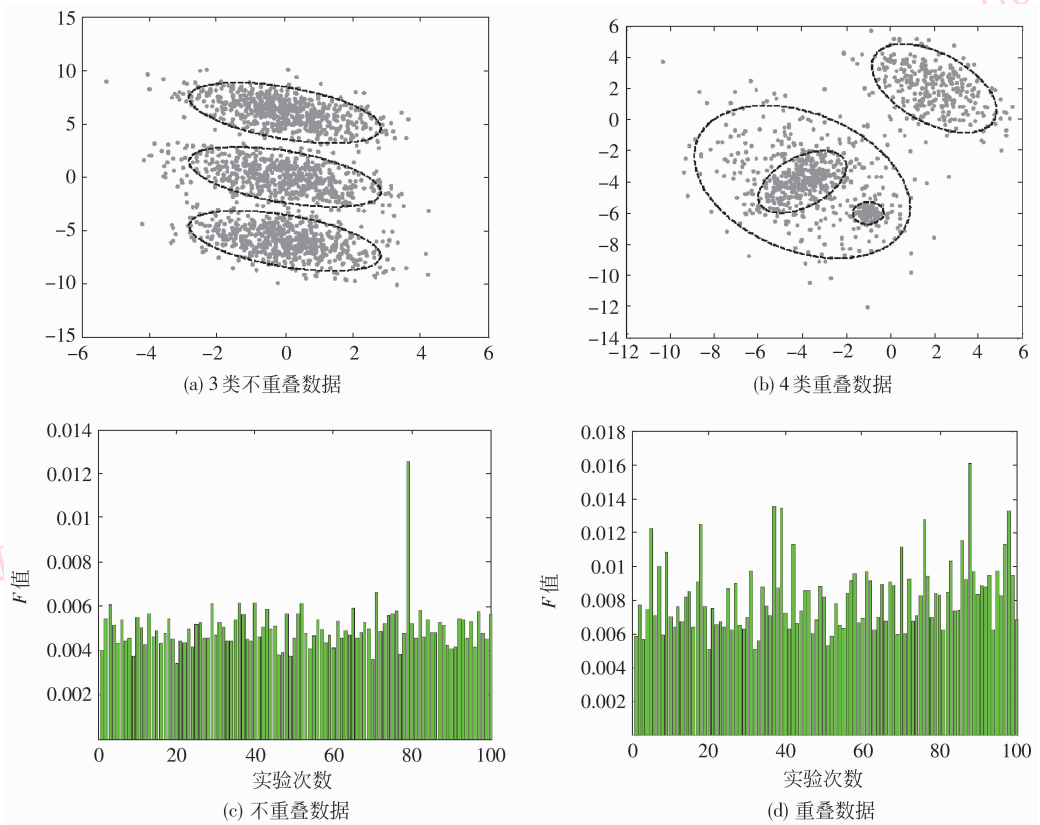

图 1　模拟数据

Fig. 1　Simulated data

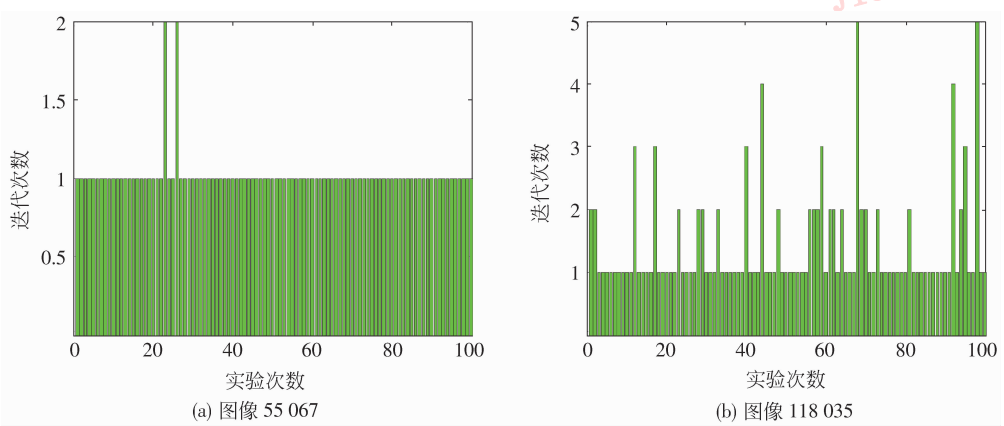

图 2　样本选取迭代次数

Fig. 2　The iterative steps of selecting sample

3.2　抽样与非抽样高斯混合模型分割结果比较

实验数据取自美国加州大学伯克利分校的 BSDB 标准测试图像数据库真实彩色图像。分别采用抽样与非抽样两种方式的高斯混合模型及 FCM 算法进行分割比较实验,实验中设置相同的参数与初始条件。分割类别数都是 $K=4$,图 3 为在抽样

与非抽样两种情况下高斯混合模型及 FCM 算法分割结果,图 3 中第 1 行为原图,第 2 行为 FCM 算法分割结果,第 3 行为非抽样高斯混合模型分割结果,最后一行为采用抽样的分割结果。从图 3 可以看出,后两种高斯混合模型方法要好于 FCM 算法,这是由于高斯混合模型既考虑聚类族的均值,又考虑它的协方差,而 FCM 方法只考虑聚类族的均值。从图 3 最后两行的结果看,采用抽样技术后,并没有降低图像的分割质量,而其中部分图像的分割结果甚至要好于非抽样的分割结果。同时采用抽样的高斯混和模型图像分割方法,模型参数迭代时间平均不超过 1 s。而图 2 的结果表明,样本数据大部分情况下只需 1—2 次的迭代就可获取,并不需要太大的时间代价,可以大大提高分割效率。

(a) 35 010 图 (b) 232 038 图 (c) 124 084 图

图 3 3 种不同方法图像分割结果对比

Fig. 3 The comparison of three different methods

为了量化评估各种算法的分割效果,采用文献[12]提出的彩色图像分割质量评价标准如下

$$V(I) = \frac{1}{1\,000(M \times N)} \sqrt{K} \sum_{i=1}^{K} \frac{e_i^2}{\sqrt{A_i}} \tag{14}$$

式中,I 为分割后图像,$M \times N$ 为图像大小,K 为图像分割的类别数,A_i 为第 i 个区域的面积即像素总个数,e_i 为第 i 个区域的平均颜色错误数,其定义为

$$e_i = \sum_{i=1}^{A_i} \sqrt{(C_{1x_1}^j - C_{2x_1}^j)^2 + (C_{1x_2}^j - (C_{2x_2}^j)^2 + (C_{1x_3}^j - C_{2x_3}^j)^2} \tag{15}$$

式中,C_1 和 C_2 分别表示分割前图像和分割后图像,x_1, x_2, x_3 分别表示彩色图像 R、G、B 分量。V 值越小,分割质量越好。

选择 BSDB 标准测试图像数据库的 10 幅图像

分别用两种方法做 20 次分割实验,分割结果的 V 值的平均值如图 4 所示。从图 4 可以看出,两种方法分割结果的 V 值比较接近,甚至有近一半的图像抽样分割结果的 V 值要优于非抽样分割结果的 V 值,表明采用抽样的分割技术并没有降低图像的分割质量。这是由于传统的高斯混合模型是局部收敛的,参数初始化对结果影响较大,而采用抽样技术,可以获取比较好的初始分割参数。

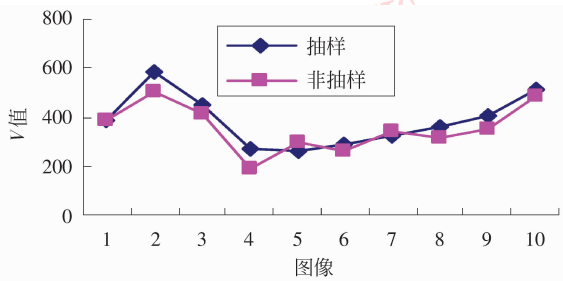

图 4　抽样与非抽样分割结果的 V 值比较

Fig. 4　The comparison of V value with sample and non-sample segmentation

4　结　论

提出基于多阶抽样的高斯混合模型彩色图像分割法。该方法根据 Chernoff 界给出了最小采样数公式,选择最具有代表性的样本进行聚类训练,得出最终的模型参数近似估计,并对采样点利用进行聚类,对剩余的点按距离最近原则进行聚类。针对数据采样具有随机性和不稳定性,设计一罚函数作为采样好坏的评价标准,消除抽样对高斯混合聚类模型的稳定性影响。通过真实彩色图像的对比实验,验证了算法具有满意的分割效果和较高的分类 V 值。算法中分割类别数需要人工预先设定、图像处理中分割类数需要自动确定的问题将是下一步的研究目标。

参考文献(References)

[1] Tapas Kanungo, Mount D M, Netanyahu N S, et al. An efficient k-means clustering algorithm: analysis and implementation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 881-892.

[2] Mark J L, Michael K N, Cheung, Y M, et al. Agglomerative fuzzy k-means clustering algorithm with selection of number of cluster [J]. IEEE Transactions on Knowledge and Data Engineering, 2008, 20 (11): 1519-1534.

[3] Song Yuqing, Xie Conghua, Zhu Yuquan, et al. Research on medical image clustering based on approximate density function [J]. Journal of Computer Research and Development, 2006, 43(11): 1947-1952. [宋余庆,谢从华,朱玉全,等. 基于近似密度函数的医学图像聚类分析研究[J]. 计算机研究与发展, 2006,43(11):1947-1952.]

[4] Zhang Jun, Tian Hao, Huang Yingjun. A novel CFAR algorithm for detecting targets in SAR images using Gaussian mixture model [J]. Tournal of Image and Graphics, 2009, (1): 19-24. [张军, 田昊, 黄英君. 利用高斯混合模型的 SAR 图像目标 CFAR 检测新方法[J]. 中国图象图形学报, 2009, (1): 19-24.]

[5] Greenspan H, Goldberger J. Constrained Gaussian mixture model framework for automatic segmentation of MR brain image [J]. IEEE Transactions on Medical Imaging, 2006, 25 (9): 1233-1245.

[6] Athanasiadis E I, Cavouras D A, Spyridonos P P. Complementary DNA microarray image processing based on the fuzzy Gaussian mixture model [J]. IEEE Transactions on Information Technology in Biomedicine, 2009, 13(4): 419-425.

[7] Agiomyrgiannakis Y, Stylianou Y. Wrapped Gaussian mixture models for modeling and high-rate quantization of phase data of speech[J]. IEEE Transactions on Audio, Speech and Language Processing, 2009, 17(4): 775-786.

[8] Dempster N M, Rubin D B. Maximum likelihood from incomplete data via EM algorithm [J]. Roy. Statist. Soc., 1977, 39(1): 1-38.

[9] Torben H, Christine R. Guided tour of Chernoff bounds [J]. Information Processing Letters, 1990, 33(6): 305-308.

[10] Figueiredo M, Jain A K. Unsupervised learning of finite mixture models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(3): 381-396.

[11] Fowlkes C. The Berkeley segmentation dataset and benchmark [DB/OL](2003-10-31)[2009-12-1]. www. cs. berkeley. edu/projects/vision/grouping/segbench/.

[12] Hui Zhang, Jason E, Fritts B, et al. Image segmentation evaluation: a survey of unsupervised methods [J]. Computer Vision and Image Understanding, 2008, 110(2): 260-280.