文章编号: 1000−8152(2008)04−0608−05

基于自适应带宽的快速动态高斯核均值漂移算法

周芳芳[1]，樊晓平[1]，叶　榛[2]

(1. 中南大学 信息科学与工程学院, 湖南 长沙 410075;

2. 清华大学 智能技术与系统国家重点实验室, 北京 100084)

摘要: 由核密度估计推导获得的高斯核均值漂移算法因收敛速度慢在应用中效率不高. 本文提出基于自适应带宽的动态更新改进方法. 首先采用空间离散方法对数据集化简, 然后引入动态更新机制, 每次迭代后将数据集更新到均值点, 并将聚集在一起的数据点用一个收敛点表示, 同时根据数据集直径的变化, 自适应地计算各向异性的带宽参数. 实验表明, 该方法提高了算法的收敛速度, 降低了计算复杂度.

关键词: 均值漂移; 高斯核; 核密度估计; 自适应带宽

中图分类号: TP273　　**文献标识码**: A

Fast dynamic Gaussian mean-shift algorithm based on adaptive bandwidth

ZHOU Fang-fang[1]，FAN Xiao-ping[1]，YE Zhen[2]

(1. College of Information Science and Engineering, Central South University, Changsha Hunan 410075, China;

2. State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing 100084, China)

Abstract: The Gaussian kernel mean-shift algorithm which is deduced from kernel density estimation has not been widely employed in applications because of its low convergence rate. We propose a dynamic mean-shift algorithm based on adaptive bandwidth. The number of data sets is reduced by adaptive space discretization; the convergence rate is improved by dynamically updating the data set, and the efficiency is promoted by replacing the overlapping points with a special point in the iterations. The anisotropic bandwidth is updated according to the diameter of the data set. Experiments validate the improvement of the convergence rate of Gaussian mean-shift with lower complexity in computation.

Key words: mean shift; Gaussian kernel; kernel density estimation; adaptive bandwidth

1　引言(Introduction)

均值漂移算法是一种高效的统计迭代算法, 它的收敛性由核函数决定[1~3]. 常用的均匀核、Epanechnikov核计算简单、收敛速度快, 但计算精度不高. 高斯核函数计算精度高、收敛路径平滑, 但收敛速度慢, 因此在处理大规模数据时, 应用较少. 为了提高高斯核均值漂移算法的计算效率, Fashing证明了当核函数是连续函数时, 均值漂移算法是一种二次边界优化算法[4], Shen在此基础上提出over-relaxed策略[5], Cheng在均值漂移算法中引入模糊机制[1]. Zhang证明了该改进算法以超线性收敛[6], Carreira提出用直方图的熵来作为迭代的停止规则[7]. 上述的改进算法均采用固定带宽, 且当数据量很大时, 算法的速度和准确性仍需进一步提高.

本文提出基于自适应带宽的快速动态均值漂移算法, 通过减少迭代次数和降低每次迭代的计算量的方法来提高算法的效率.

2　高斯核均值漂移(Gaussian mean shift)

高斯核均值漂移算法由核密度估计的梯度推导获得. 给定d维欧拉空间$\mathbb{R}$中的n个采样点$S = \{x_i, 1 \leqslant i \leqslant n\}$, 利用高斯核函数$K_N(x) = (2\pi)^{-\frac{d}{2}} \exp(-\|x\|^2)$及正定的$d \times d$的带宽矩阵$H_i$, 核密度估计公式为

$$\hat{f}(x) = (2\pi)^{-\frac{d}{2}} \sum_{i=1}^{n} w_i |H_i|^{-\frac{1}{2}} \exp(-\frac{1}{2}\|x-x_i\|^2_{H_i}). \quad (1)$$

其中$w_i \geqslant 0$表示采样点x_i的权重, 满足$\sum w_i = 1$. 高斯核定义了采样点x_i与核中心x之间的相似性, 带宽矩阵H_i决定了核函数的影响范围, $\|x - x_i\|_{H_i^2} = (x - x_i)H_i^{-1}(x - x_i)$, 称作马式距离. 核密度估计

收稿日期: 2006−12−21; 收修改稿日期: 2007−09−17.
基金项目: 国家自然科学基金资助项目(69975003).

$\hat{f}(x)$是每个采样点处的高斯核加权求和的结果. 由于密度极值点是密度梯度为零的采样点, 即$\nabla f = 0$, 因此计算密度梯度的估计:

$$\nabla \hat{f}(x) = (2\pi)^{-\frac{d}{2}} \sum_{i=1}^{n} w_i |H_i|^{-\frac{1}{2}} H_i^{-1}(x - x_i) \times \exp(-\frac{1}{2}\|x - x_i\|_{H_i}^2) = \hat{f}(x) \sum_{i=1}^{n} H^{-1}(x - x_i). \tag{2}$$

定义参数H_x:

$$H_x = \frac{\sum_{i=1}^{n} |H_i|^{-\frac{1}{2}} \exp(-\frac{1}{2}\|x - x_i\|_{H_i}^2)}{\sum_{i=1}^{n} |H_i|^{-\frac{1}{2}} H_i^{-1} \exp(-\frac{1}{2}\|x - x_i\|_{H_i}^2)}. \tag{3}$$

用H_x乘以式(2)可得$H_x \nabla \hat{f}(x) = \hat{f}(x) M(x)$, 因此

$$M(x) = H_x \frac{\nabla \hat{f}(x)}{\hat{f}(x)}. \tag{4}$$

其中$M(x) = m(x) - x$, 称作均值漂移向量$m(x)$:

$$m(x) = \frac{\sum_{i=1}^{n} w_i |H_i|^{-\frac{1}{2}} H_i^{-1} \exp(-\frac{1}{2}\|x - x_i\|_{H_i}^2) x_i}{\sum_{i=1}^{n} w_i |H_i|^{-\frac{1}{2}} H_i^{-1} \exp(-\frac{1}{2}\|x - x_i\|_{H_i}^2)} \tag{5}$$

为均值漂移迭代公式, 表示采样点的加权平均值. 由公式(4)可知均值漂移向量$M(x)$总是指向密度大的方向, 因此算法收敛到密度极大值点. 假设带宽参数H_i固定且各向分布一致, 带宽记做$H = h^2 I$. 此时的均值漂移向量和均值漂移公式为

$$M(x) = m(x) - x = H \frac{\nabla \hat{f}(x)}{\hat{f}(x)}, \tag{6}$$

$$m(x) = \frac{\sum_{i=1}^{n} w_i \exp(-\frac{1}{2}\|\frac{x - x_i}{h}\|^2) x_i}{\sum_{i=1}^{n} w_i \exp(-\frac{1}{2}\|\frac{x - x_i}{h}\|^2)}. \tag{7}$$

3 自适应带宽的快速动态高斯核均值漂移算法(Fast dynamic Gaussian mean shift based on adaptive bandwidth)

为了提高高斯核均值漂移算法的收敛速度, 本文引入数据集的动态更新机制[6]. 在标准的均值漂移算法中, 数据集S固定不变, 路径点T不断的更新, 算法如图1(a)所示. 而动态的均值漂移算法对数据集S和路径T同时更新, 每次迭代计算每个数据点的均值漂移向量, 并将数据点更新到均值点, 下一次迭代在新的数据集上进行, 算法的过程如图1(b)所示. 然而上述的方法均采用固定带宽h, 如果带宽过大, 可能会合并某些极值点. 因此本文提出自适应带宽的数据集动态更新的均值漂移算法.

for $x_i \in S,\ i = 1, \cdots, N$
　$t_i = x_i$
　repeat
　　$m(t_i) = \dfrac{\sum_{j=1}^{n} w_j \exp(-\frac{1}{2}\|\frac{t_i - x_j}{h}\|^2) x_j}{\sum_{j=1}^{n} w_j \exp(-\frac{1}{2}\|\frac{t_i - x_j}{h}\|^2)}$
　　distance $= \|m(t_i) - t_i\|_2$
　　$t_i \leftarrow m(t_i)$
　until distance $<$ tolerance
end

(a) 标准的均值漂移算法

repeat
　for $x_i \in S,\ i = 1, \cdots, N$
　　$t_i = x_i$
　　$m(t_i) = \dfrac{\sum_{j=1}^{n} w_j \exp(-\frac{1}{2}\|\frac{t_i - x_j}{h}\|^2) x_j}{\sum_{j=1}^{n} w_j \exp(-\frac{1}{2}\|\frac{t_i - x_j}{h}\|^2)}$
　end
　$\forall x_i = m(t_i)$
until stop rule
Combination$(x_i, \min_\text{distance})$

(b) 动态的均值漂移算法

图 1　算法伪码图

Fig. 1　Pseudo code of the algorithm

3.1 自适应带宽的计算(Computing adaptive bandwidth)

定理 1　设密度函数f服从正态分布$N(\mu, \sigma^2)$, 用带宽为h的高斯核进行均值漂移, 当$h = \sigma$, 带宽正规化的均值漂移向量的模取最大值.

证　由于f满足正态分布, 方差为σ^2, 由中心极限定理可得密度估计$\hat{f}(x)$的均值也为正态分布: $\mathrm{E}[\hat{f}(x)] = \Phi(x; \sigma^2 + h^2)$, 同理梯度估计的均值为: $\mathrm{E}[\nabla \hat{f}(x)] = \nabla \Phi(x; \mu^2 + h^2)$. 当采样点的数量足够大时, 由大数定理可得[8]

$$p\lim M(x) = H \frac{\mathrm{E}[\nabla \hat{f}(x)]}{\mathrm{E}[\hat{f}(x)]} = H \frac{\nabla \Phi(x; \sigma^2 + h^2)}{\Phi(x; \sigma^2 + h^2))} = -\frac{h^2}{\sigma^2 + h^2}(x - \mu). \tag{8}$$

其中$p\lim$表示概率极限. 带宽正规化的均值漂移向量的模为

$$M(x; h) = \|\frac{p\lim M(x)}{h}\| = \frac{h}{\sigma^2 + h^2}\|x - \mu\|. \tag{9}$$

若$h = \sigma$, $M(x; \sigma)^2 - M(x; h)^2 \geqslant 0$, $M(x, \sigma)$为最大值:

$$M(x; \sigma)^2 - M(x; h)^2 =$$

$$(\frac{1}{4\sigma^2}-\frac{h^2}{\sigma^2+h^2})\|x-\mu\|^2=\frac{(\sigma^2-h^2)^2}{4\sigma^2(\sigma^2+h^2)}\|x-\mu\|^2. \tag{10}$$

当$h=\sigma$, 上式取等号, 即$M(x;\sigma)^2$为最大值.

证毕.

定理1证明了高斯分布数据集的属性, 其他结构的数据集具有类似的特点.因此可以获得自适应带宽的计算方法. 在第t次迭代过程中, 数据集中每个数据点的最优带宽的计算步骤为

1) 设第t次迭代, 初始的带宽为$h_{\rm prev}^{(t)}$;

2) 分别用较大的带宽$h_{\rm prev}^{(t)}+\Delta h(t)$和较小的带宽$h_{\rm prev}^{(t)}-\Delta h(t)$计算带宽正规化的均值漂向量$M=\{M_-,M,M_+\}$;

3) 计算正规化的均值漂移向量的模$\{\|M_-\|,\|M\|,\|M_+\|\}$, 将模取最大值的带宽作为数据点的最优带宽$h_{\rm opt}^{(t)}$. Shen证明了更新后的步长$M$小于2倍初始步长时, 算法一定收敛. 因此当更新后的步长大于2倍初始步长, 取初始步长进行计算;

4) 更新数据场.

动态更新后的数据点位置改变, 数据集的结构也发生了变化, 因此下一次迭代的初始带宽$h_{\rm prev}^{(t+1)}$则需要做相应的更新. 可以证明如果数据集满足高斯分布, 变化后的数据集仍然保持高斯分布.

定理 2 设t次迭代后的数据集为$S^{(t)}=\{x_i^{(t)},i=1,\cdots,n\}$服从正态分布$N(\mu,(\sigma^{(t)})^2)$, 则$t+1$次高斯核均值漂移迭代后, 更新后的数据集$S^{(t+1)}=\{x_i^{(t+1)},i=1,\cdots,n\}$仍然服从正态分布$N(\mu,(\sigma^{(t+1)})^2)$, 其中均值$\mu$保持不变, 而方差收缩为$(\sigma^{(t+1)})^2=(\alpha^{(t)})^2(\sigma^{(t)})^2$, 其中$\alpha^{(t)}=(\sigma^{(t)})^2/((\sigma^{(t)})^2+h^2)$, 且$\lim\limits_{t\to\infty}(\sigma^{(t+1)})^2=0$.

证 数据集中任何一点$x_i^{(t)}\in S^{(t)}$根据公式(4)和(8), $t+1$次迭代后数据更新为

$$\begin{aligned}&x_i^{(t+1)}=x_i^{(t)}+M(x_i^{(t)})=\\&x_i^{(t)}-\frac{h^2}{h^2+\sigma^2}(x_i^{(t)}-\mu)=\\&\frac{\sigma^2}{h^2+\sigma^2}x_i^{(t)}+\frac{h^2}{h^2+\sigma^2}\mu=\alpha x_i^{(t)}+c.\end{aligned} \tag{11}$$

容易证明: 当$x_i^{(t)}=\mu$时, 代入式(11)得$x_i^{(t+1)}=\mu$, 即均值保持不变; 而由于$x_i^{(t+1)}$与$x_i^{(t)}$之间存在线性关系, 因此$S^{(t+1)}$方差为$(\alpha^{(t)})^2(\mu^{(t)})^2$. $\alpha<1$总成立, 因此$\lim\limits_{t\to\infty}(\alpha^{(t)})^2(\mu^{(t)})^2=\lim\limits_{t\to\infty}(\mu^{(t+1)})^2=0$.

证毕.

由定理2可知, 高斯分布的数据集动态更新后仍满足高斯分布, 均值不变, 方差减少, 数据集收缩. 因此, $t+1$次迭代时的带宽h应该根据$\alpha^{(t)}$减小. 而当数据集不满足高斯分布时, 数据点亦会收缩, 但是主方向上的数据收缩速度慢, 非主方向上的收缩速度快. 因此可以根据数据点的结构来计算每次迭代的初始带宽参数$H_{\rm prev}^{(t)}$. 本文利用数据点各个方向的数据的直径来度量各个方向的带宽的收缩比例.

定义 1 给定d维欧拉空间R^d中的n个采样点$S=\{x_i,1\leqslant i\leqslant n\}$, $x_i=\{x_{i1},\cdots,x_{id}\}$, 计算$\rho=\{\rho_j=\max(x_{ij})-\min(x_{ij}),\ i=1,\cdots,n,\ j=1,\cdots,d\}$, ρ为$d\times d$的矩阵, ρ_j为对角线上的元素, 称ρ为数据集的直径.

t次迭代后的带宽为$H^{(t)}$, 则$t+1$次迭代的带宽$H^{(t+1)}$的收缩比例$\alpha^{(t+1)}=\rho^{(t+1)}/\rho^{(t)}$, 此时的带宽$H^{(t+1)}=\alpha^{(t+1)}H^{(t)}$. $H^{(t+1)}$为$d\times d$的对角矩阵, 对角线上的元素表示各项异性的带宽参数. 因此, 根据自适应带宽的计算方法可知, 对于初始的数据集, 设置各向同性的带宽参数, 随着数据的迭代, 带宽根据数据集的结构变化为各向异性的带宽参数.

3.2 快速的动态均值漂移算法(Fast dynamic mean shift)

动态均值漂移算法虽然使数据集逐渐地收缩到极值点, 但是数据集的数目始终保持不变, 因此数据点逐渐地重叠在一起. 本文将重叠的数据点用一个收敛点表示, 该点的权值是所有重叠点的权值之和. 用收敛点计算均值漂移算法不影响算法的准确性, 并且能显著降低了每次迭代的计算量.

定义 2 当数据点$G=\{x_i,i=1,\cdots,l\}$重叠在一起时, 即, $\|x_i-x_j\|^2=0,(i,j=1,\cdots,l$且$i\neq j)$, 可以用一个点$c=(w_c,numc)$来表示, 其中点$c$的权值为$w_c=\sum w_{xi}$, $numc=l$, 记录重叠的数据点的数目, 因此称c为G的收敛点.

快速动态均值漂移算法的停止规则是比较连续两次迭代后数据集中数据量的个数, 当数据量不变时可以结束迭代. 通过实验发现, 数据点常常会收敛到鞍值点, 降低了算法的准确性. 但通常收敛到鞍值点的数据点个数很少, 可以通过判断收敛点c中的$numc$将鞍值点中的数据点归并到离峰值点最近的一个类中. 快速的动态均值漂移算法在不影响算法的准确性的情况下, 通过减少数据集的数据量来降低了算法的计算量.

3.3 空间的离散化方法(Spatial discretization)

虽然快速的动态均值漂移算法在多次迭代后, 数据量会显著下降.但初始迭代时, 若数据量大, 算法总计算时间仍然很长. 本文提出了均匀的空间离散化和kd-tree的自适应的空间离散化两种方法.

均匀的空间离散化方法首先将整个空间平均的分成m块, 用每一块中的均值点参加快速动态均值漂移算法, 最后均值点所属的类为数据块所属的类.该离散化方法简单, 适合对图像数据进行离散化. 另一种离散化方法是采用kd-tree算法对空间自适应

划分. 计算每个块m中数据点的均值和方差, 如果方差太大, 则将该块继续划分, 如果方差较小, 则停止划分. 最后自适应带宽的快速动态均值漂移算法为

1) 根据数据集所含数目的大小及分布选择均匀的空间离散化方法或kd-tree自适应的方法化简数据集, 同时计算数据点的初始带宽;

2) 根据定理 1 选择每个数据点计算最优带宽, 并根据均值漂移公式计算均值;

3) 将数据点更新到均值处, 用收敛点表示相互重叠的数据点, 并计算新数据集各个维度的直径;

4) 计算连续两次的数据集直径的比值, 更新带宽;

5) 比较连续两次数据集中数据的数目, 若相等则停止迭代; 否则循环执行2).

3.4 收敛速度和计算复杂度的分析(Speed of convergence and complexity)

由定理2可知, 当数据集满足正态分布时, 自适应带宽的快速动态均值漂移算法使数据集的方差收敛渐近地收敛到0, 因此可以计算它的收敛速度为

$$\lim_{t\to\infty}\frac{|\sigma^{t+1}-0|}{|\sigma^{t}-0|}=\lim_{t\to\infty}\frac{|\alpha^{t}\sigma^{t}|}{|\sigma^{t}|}=\lim_{t\to\infty}\alpha^{t}=0. \tag{12}$$

因此, 自适应带宽的快速动态均值漂移也是超线性收敛. 当数据集不是高斯分布, 数据集仍为超线性收敛. 另外, 当数据集满足正态分布时, 可进一步计算出收敛的阶数:

$$\lim_{t\to\infty}\frac{|\sigma^{t+1}|}{|\sigma^{t}|^{p}}=\lim_{t\to\infty}\frac{|\alpha^{t}\sigma^{t}|}{|\sigma^{t}|^{p}}=\lim_{t\to\infty}\frac{(\sigma^{t})^{3-p}}{(\sigma^{t})^{2}+(h^{t})^{2}}. \tag{13}$$

当$p=3$时, 数据集的方差3次收敛, 相应的数据集也3次收敛, 即动态的高斯核均值漂移算法3次收敛, 而静态的高斯核均值漂移算法的为1次收敛.

本文通过分析算法中数据点访问次数来计算算法的复杂度, 如表1所示. 在标准的均值漂移算法(MS)中, 数据点的访问次数约为t_1dn^2, t_1为平均迭代次数, d为维数, n为数据点的个数. 动态的均值漂移算法(DMS)为t_2dn^2, t_2为DMS的迭代次数, 因为它以超线性的速度收敛, 因此$t_2<t_1$.快速的动态均值漂移算法(ADMS)的访问次数约为$t_2d\sum_{t=1}^{t_2}(n^{(t)})^2$, $n^{(t)}$为每次迭代数据点的个数, 其中$n^{(t)}>n^{(t+1)}, t=1,\cdots,t_2$.

表 1　算法的计算复杂度比较

Table 1　Comparison of the complexity

	MS	DMS	ADMS
复杂度	t_1dn^2	t_2dn^2	$t_2d\sum_{t=1}^{t_2}(n^{(t)})^2$

4 实验结果(Experiment results)

本文采用MATLAB实现算法. 首先对人工合成的数据进行实验, 图2是合成的非线性数据集, 数据点数目为32640. 由于数据量很大计算时间很长, 采用kd tree算法对空间进行自适应的离散化, 获得3300个数据点, 如图2(a)实心点所示. 自适应动态均值漂移的过程如图2(a)(b)(c)所示, 最后的分类结果如图2(d)所示, 算法正确的将数据集分为3类. 对于化简后的数据集, 自适应带宽的快速动态均值漂移算法比固定带宽的动态均值漂移算法提高了51.2%, 比标准的均值漂移算法提高了92.12%. 3种算法的计算时间、迭代次数以及数据点访问次数的比较如表2所示.

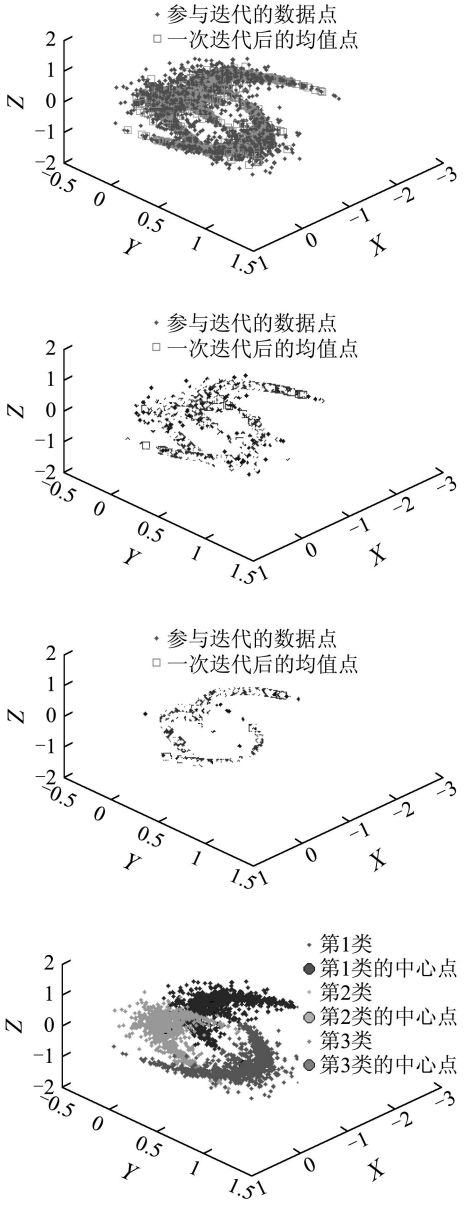

图 2　非线性分布数据集的聚类过程及结果.

Fig. 2　The process and results of clustering for non-linear distributed data

表 2 算法的结果比较

Table 2 Comparison of the performance

	MS	DMS	ADMS
时间/s	100.157	16.17	7.89
迭代次数	22	10	10
访问次数	73260	33000	14298

均值漂移算法常用来对图像进行分割, 因此本文将该算法应用到图像数据. 首先对图像采用均匀的空间离散化方法, 将数据集化简, 然后利用自适应带宽的快速动态均值漂移算法, 分割的效果如图3所示. 图3中的第1个图是298×234的牙齿切片图像. 平均的空间离散化后为60×47, 初始带宽为20, 迭代次数为5次后分类结果为图3中的第2图. 通过观察发现, 第2图将牙齿切片数据正确的分类为牙釉质、牙本质、牙髓和背景4类, 并用不同的颜色表示聚类的结果. 实验表明, 本文的方法在保证了分类效果的同时, 提高了计算效率.

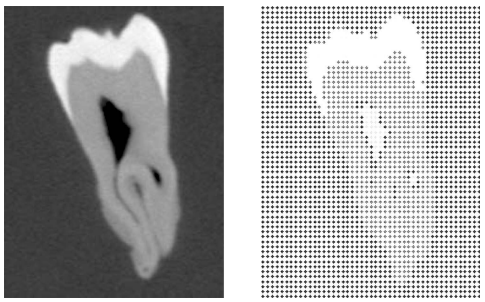

图 3 tooth切片分类结果图

Fig. 3 Result of segmentation of tooth slice

5 总结(Conclusion)

本文提出了高斯核均值漂移加速算法. 动态更新机制使数据集超线性收敛到极值点. 收敛点和离散化方法降低了每次迭代的计算量. 根据数据集的直径自适应更新带宽, 提高了算法的准确性. 未来的研究方向是利用快速高斯变换来化简计算, 进一步提高计算效率.

参考文献(References):

[1] CHENG Y Z. Mean shift, mode seeking, and clustering[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995, 17(8): 790 – 799.

[2] COMANICIU D, MEER P. Mean shift: A robust approach toward feature space analysis[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(5): 603 – 619.

[3] 李乡儒, 吴福朝, 胡占义. 均值漂移算法的收敛性[J]. 软件学报, 2005, 16(3): 365 – 374.
(LI Xiangru, WU fuchao, HU Zhanyi. Convergence of a mean shift algorithm[J]. *Journal of Software*, 2005, 16(3): 365 – 374.)

[4] FASHING M, TOMASI C. Mean shift is a bound optimization[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, 27(3): 471 – 474.

[5] SHEN C, BROOKS M J. Adaptive over-relaxed mean shift[C]//*Proceedings of the 8th International Symposium on Signal Processing and Its Applications*. New York: IEEE CS Press, 2005: 28 – 31.

[6] ZHANG K, KWOK J T, TANG M. Accelerated convergence using dynamic mean shift[C]//*Proceedings of the 9th European Conference on Computer Vision*. New York: Springer, 2006: 257 – 268.

[7] CARREIRA-PERPINAN M A. Acceleration strategies for Gaussian mean-shift image segmentation[C]//*Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. New York: IEEE CS Press, 2006: 543 – 549.

[8] COMANICIU D. An algorithm for data-driven bandwidth selection[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, 25(2): 281 – 288.

作者简介:

周芳芳 (1980—), 女, 博士研究生, 研究方向为虚拟现实技术、科学计算可视化等, E-mail: zff@mail.csu.edu.cn;

樊晓平 (1961—), 男, 教授, 博士生导师, 研究方向包括智能控制、智能机器人、虚拟现实技术、智能交通系统等;

叶 榛 (1945—), 女, 教授, 研究方向为系统仿真、临场感及虚拟现实技术等.