文章编号:1004-9037(2013)05-0614-06

EM 聚类和 SVM 自动学习的白细胞图像分割算法

郑 馨 王 勇 汪国有

(华中科技大学多谱信息处理技术国家级重点实验室,武汉,430074)

摘要:白细胞图像分割是白细胞自动识别的关键环节,其分割效果直接影响后续步骤。为提高光照、颜色不稳定情况下的分割精度,提出一种基于期望最大化(Expectation maximization, EM)聚类和支持向量机(Support vector machine, SVM)自动采样-学习的彩色白细胞图像分割方法。首先采用 EM 算法对 CIELUV 颜色空间的 L 分量聚类得到细胞核区域。在细胞浆分割阶段,先利用 EM 过分割和膨胀的细胞核区域获取正负样本候选区域;接着用基于 EM 的分层抽样得到正负样本;再提取颜色特征自动对正负样本训练获得 SVM 模型;最后利用 SVM 分类模型得到整个细胞区域。与传统的白细胞图像分割算法相比,本文方法更能适应图像光照和颜色的变化;与同类的分割算法相比,本文方法提高了分割精度。相关实验结果表明,本文算法具有良好的精度和鲁棒性。

关键词:彩色图像分割;期望最大化;支持向量机;白细胞

中图分类号:TP391.4 **文献标志码**:A

White Blood Cell Segmentation Using Expectation-Maximization and Automatic Support Vector Machine Learning

Zheng Xin, *Wang Yong*, *Wang Guoyou*

(National Key Laboratory of Science and Technology on Multi-spectral Information Processing, Huazhong University of Science & Technology, Wuhan, 430074, China)

Abstract: Leukocyte segmentation is the key step of an automatic leukocyte recognition system, and has a direct influence on subsequent processing steps. In order to improve the accuracy of leukocyte segmentation under the conditions with variant illumination and unstable staining, an innovative segmentation method for color leukocyte image, which is based on expectation maximization (EM) clustering and support vector machine (SVM), is proposed. Firstly, Gaussian mixture model (GMM) is adopted to segment the nuclei using the L channel of CIELUV color space, and EM algorithm is taken to estimate the parameters of GMM. Secondly, the over-segmentation results by EM algorithm in the first step are fully used to expand the nucleus regions and gain the positive and negative regions for SVM training. As for sampling in these regions, a stratified sampling technique based on EM algorithm is used. Then the SVM is trained online using the color features of sampling pixels. Thirdly, the whole leukocyte regions are segmented through classifying every pixel in the image using SVM. Different from traditional leukocyte segmentation methods, the proposed method can cope with illumination and variation very well. Compared with other similar methods, the proposed method improves the segmentation accuracy. Finally, experiments demonstrate the accuracy and robustness of the proposed method.

Key words: color image segmentation; expectation maximization; support vector machine (SVM); white blood cell

基金项目:国家高技术研究发展计划("八六三"计划)(2011AA02A118)资助项目。

收稿日期:2013-06-15;**修订日期**:2013-07-26

引　言

白细胞图像自动分析识别,可以降低全自动血液分析仪的成本[1],辅助临床诊断,因此具有重要的临床意义和广阔的发展前景。白细胞图像自动识别主要包含 4 个步骤:图像采集、细胞分割、特征提取和分类识别,细胞分割是最关键的一步[2],胞核、胞浆区域的分割精度直接影响后续步骤的准确性。然而,受白细胞图像复杂性的约束,目前的细胞分割算法仍不能适用于各类白细胞图像。其复杂性主要体现在两点:(1)对于不同类别的细胞,其胞核、胞浆的颜色和形态差异较大;白细胞和红细胞之间还存在粘连、重叠现象;(2)图像采集设备差异和染色制备条件不一致等导致白细胞图像存在光照和颜色不稳定的问题,使白细胞图像更加复杂。因此白细胞图像分割是医学图像处理领域中颇具挑战性的难题。

为了解决这一难题,国内外相关学者在这方面作了较广泛的研究。基于主动轮廓模型[3,4]的细胞分割算法可以使轮廓精确收敛。尽管该类方法具有对光照和噪声不敏感的优势[5],分割效果较好,然而其对初始轮廓的要求较高:若初始轮廓距离真实轮廓较远,则分割结果精度难以保证,且收敛耗时很长。印勇等[6]利用改进的 Meanshift 算法自适应获取图像饱和度 S 直方图和 G 通道直方图的分割阈值,分割速度快,对图像内与其他区域颜色差异明显的区域分割效果好、鲁棒性强;但没有充分利用图像的彩色信息,对于胞浆区域分割效果不好。文献[7]通过模糊 C 均值聚类对输入图像进行过分割得到若干小区域,再利用小区域的颜色均值和位置信息对其筛选和合并得到最终分割结果。该方法对细胞粘连情况处理较好,但对噪声敏感。文献[8]利用多光谱特征,通过人工标记获取训练样本训练支持向量机(Support vector machine, SVM)[9]分类器对细胞图像进行分割。虽然该方法分割效果好,但不能自动采样和学习,因此对光照、颜色变化的鲁棒性较差。文献[10]提出了利用 RGB 颜色特征在线采样-自动学习的 SVM 骨髓白细胞分割方法,对图像颜色变化、染色条件差异等表现出了较强的鲁棒性。但由于其正负样本采集方式的局限性,该方法常出现过分割或欠分割现象,例如分割结果出现孔洞或包含红细胞边缘。因此,目前基于 SVM 分类的白细胞分割算法尚不能满足精度需求。本文主要目的是提出一种对光照和颜色变化鲁棒的细胞分割算法,自动采集具有代表性的正负训练样本,训练出有效的分类模型对白细胞图像进行自动分割。

1　基于 EM 的细胞核区域分割

1.1　EM 算法

混合高斯模型(Gaussian mixture model, GMM)使用 K 个高斯模型来表征图像中各个像素点的特征。图像中的每个像素均被认为是通过 GMM 中某一个组成的密度函数计算得到。数据 $\boldsymbol{x}$ 的概率分布密度函数为

$$p(\boldsymbol{x}\mid\mu,\Sigma)=\sum_{k=1}^{K}\pi_k N(\boldsymbol{x}_i\mid\mu_k,\Sigma_k) \quad (1)$$

一旦确定每个组成的参数:混合权重 π_k,均值向量 μ_k 和协方差矩阵 Σ_k,GMM 就可以确定,从而可以判断每个输入样本的类别。通常采用期望最大化(Expectation maximization, EM)算法[11]对 GMM 的参数进行估计。EM 算法是一种以迭代的方式从"不完全数据"中求解 GMM 分布参数的极大似然估计的方法,其算法流程如下:

(1)初始化模型参数 π_k,μ_k 和 Σ_k;

(2)E 步骤:利用现有参数,计算隐藏变量的最大似然估计值

$$\gamma(z_{nk})=\frac{\pi_k N(\boldsymbol{x}_i\mid\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(\boldsymbol{x}_i\mid\mu_j,\Sigma_j)} \quad (2)$$

$$N_k=\sum_{n=1}^{N}\gamma(z_{nk}) \quad (3)$$

(3)M 步骤:最大化"E 步骤"得到的最大似然值来重新估计分布参数

$$\mu_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)\boldsymbol{x}_i \quad (4)$$

$$\Sigma_k=\frac{1}{N_k}\sum_{i=1}^{N}\gamma(i,k)(\boldsymbol{x}_i-\mu_k)(\boldsymbol{x}_i-\mu_k)^{\mathrm{T}} \quad (5)$$

$$\pi_k=\frac{N_k}{N} \quad (6)$$

(4)估计似然函数

$$\ln p(\boldsymbol{x}\mid\mu,\Sigma,\pi)=\sum_{n=1}^{N}\ln\{\sum_{k=1}^{K}\pi_k N(\boldsymbol{x}_i\mid\mu_k,\Sigma_k)\} \quad (7)$$

迭代收敛直至满足条件,否则转回 E 步骤。最终得到参数估计。

1.2　基于 EM 的细胞核区域分割步骤

基于 EM 的细胞核区域分割算法步骤如下:

(1)对图像进行平滑滤波,将滤波后图像的RGB颜色空间转换为CIELUV色彩空间,取 L 分量图并归一化。

$$L(i,j)=255\times(L(i,j)-L_{\min})/(L_{\max}-L_{\min}) \quad (8)$$

(2)用EM算法对 L 分量图估计 $K=3$ 的GMM的参数,为了解决EM算法收敛速度慢和对初始中心敏感的问题,本文采用 K 均值算法初始化参数,K 均值的3个类中心的灰度值分别设为0,128和255。

(3)根据GMM参数,取均值最小的一类为细胞核类。

理想情况下,当各类别大小较平均、类间颜色距离较远、类内颜色变化范围较小时,用EM聚类直接对RGB彩色空间估计 $K=4$ 的GMM能获得精确的细胞核区域,而实际情况中,不同类别白细胞的细胞浆颜色和大小变化范围较大,图像中的红细胞数量难以控制,红细胞与细胞浆颜色非常相近,这些因素会导致EM聚类结果不准确。而经过更接近视觉感知的CIELUV颜色空间转换后,细胞浆和红细胞在颜色上更接近,可以视作一类。因此仅需要估计 $K=3$ 的GMM参数就可以得到较精确的3个类别的类中心,可以得到精确核区域,同时计算量更少。

2 基于EM分层采样和SVM自动学习的细胞浆区域分割

细胞浆区域分割一直是细胞图像分析中最具挑战性的难题:(1)细胞浆颜色变化多:不同类别白细胞的细胞浆呈现不一样的颜色;单个粒细胞的细胞浆内含有染色颗粒,导致细胞浆颜色变化范围大;另外,细胞浆颜色与红细胞或背景颜色十分接近。(2)细胞浆形态差异大:不同类别白细胞的细胞浆的面积、形状差异大,且细胞浆和红细胞往往存在粘连现象。因此,细胞浆区域分割难度远远大于细胞核区域分割。本文结合EM聚类和SVM有监督分类的优势,自动获取精确的细胞浆区域。

2.1 基于EM的训练样本采集

目前主要有两种样本获取方法,一种是人工选取方法,另一种是自动选取方法。人工选取方法需要对细胞图像手工分割获取标记图像,一方面该方法需要足够大的图库并耗费大量的人力,另一方面该方法的鲁棒性差,若细胞图像与训练图库的颜色差异较大,则训练出的分类器分割效果较差。因此本文采用自动选取方法获取细胞图像中有效的训练样本。

由于EM算法直接对白细胞图像估计4类区域不准确,本文利用细胞浆包裹细胞核的先验知识对EM过分割结果进行聚合,预估候选细胞浆区域。先利用EM算法估计GMM参数($K>4$),将图像分割为若干个互不重叠的过分割区域;再利用形态学操作对上一步骤得到的核区进行膨胀,将与膨胀后的核区相重叠的过分割区域合并(去除与图像边界粘连的区域),标记为候选浆区,其他除核区外的区域标记为背景区域。接下来,分别对前景区域(候选浆区)和背景区域(主要包含背景和红细胞)进行采样获得训练样本。

训练样本的选择直接影响分类结果,其影响甚至大于分类器的选择。随机抽样是最常用的采样方法。随机抽样包含均匀采样、分层抽样、整群抽样等。为了获取有代表性的训练样本,防止周期性偏差,本文采用基于EM的分层抽样:根据不同的EM过分割区域占总体的比重,从预估浆区的每个过分割区域中随机抽取正样本,同时从背景区域的每个过分割中随机抽取负样本。基于EM的训练样本采集步骤如下:

(1)平滑图像,用EM算法估计 $K=7$ 的GMM模型,将图像分割为若干个过分割区域;

(2)膨胀核区,将与膨胀核区重叠的过分割区域标记为候选浆区。浆区内的过分割区域为 C_i,除核区外的其他区域标记为背景区域,其过分割区域为 B_j;

(3)每个 C_i 随机抽取 $\alpha N\times n_i/n_{\text{cyto}}$ 个像素作为正样本,每个 B_j 抽取 $(1-\alpha)N\times n_j/n_{\text{back}}$ 个点作为负样本。其中,n_{cyto} 和 n_{back} 分别为浆区和背景区域的像素个数,N 为样本总数,α 控制细胞区域与背景区域的采样比例。下面的实验中将分析不同 α 和 N 值对分类性能的影响,从而确定最优值。

2.2 基于SVM自动学习的细胞浆区域分割步骤

由于SVM具有良好的泛化能力和出色的小样本学习能力,目前基于SVM的图像分割算法已成功运用于医学图像分析领域[12]。SVM分类的原理是将样本空间映射到一个更高维的空间,在这个空间里建立方向合适的分隔超平面,使两个与之平行的超平面间的距离最大化。SVM以训练误差

为优化问题的约束条件，以置信范围值最小化为优化目标，因此推广能力优于传统学习算法，特别是在解决小样本、非线性及高维模式识别问题中表现出特有的优势[13]。

给定一个线性可分的训练样本集$(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})$，$\boldsymbol{x}\in\mathbf{R}^m$ 是 m 维特征，$\boldsymbol{y}\in\{1,-1\}$是分类标签。$i=1,2,\cdots,n$ 表示第 i 个样本。同时考虑最大分类间隔和最少错分样本，最优间隔分类器为

$$\begin{aligned}\min\quad & \frac{1}{2}\|\boldsymbol{w}\|^2+C\left(\sum\nolimits_{i=1}^{n}\xi_i\right)\\ \text{s.t}\quad & \boldsymbol{y}^{(i)}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(i)}+b)-1+\xi_i\geqslant 0\\ & \xi_i\geqslant 0\end{aligned}\tag{9}$$

式中：ξ_i 为松弛项，C 是对错分样本的惩罚常数。这是一个典型的二次凸规划问题，因此存在唯一全局最小解。应用 Lagrange 乘子并满足 KKT(Karush-Kuhn-Tucher)条件可得最优分类超平面的分类函数为

$$f(x)=\mathrm{sgn}\{\sum\nolimits_{i=1}^{n}\alpha_i^*\boldsymbol{y}^{(i)}K(\boldsymbol{x}^{(i)},\boldsymbol{x})+b^*\}\tag{10}$$

式中：α_i^*为 Lagrange 乘子，$K(\boldsymbol{x}^{(i)},\boldsymbol{x})$为支持向量 $\boldsymbol{x}^{(i)}$ 和未知向量 $\boldsymbol{x}$ 的核函数，b^* 为分类阈值。通过引入核函数可以避免在高维特征空间的复杂运算和“维数灾难”。

细胞浆区域分割步骤如下：

(1)获取核区临近的 EM 过分割区域作为候选浆区；

(2)利用基于 EM 的分层抽样获取正负训练样本；

(3)提取颜色特征作为每个像素点的特征向量；

(4)根据特征向量和样本标记训练 SVM 模型；

(5)利用训练好的 SVM 模型对像素点分类获取细胞区域和背景区域，并用形态学操作对细胞区域进行修正。

本文提取了具有代表性的训练样本，又利用了 SVM 良好的小样本学习和泛化能力，所以即使细胞浆颜色分布广泛，仍能得到准确的细胞区域。另外，由于大部分的红细胞被作为负训练样本，所以与细胞区域粘连的红细胞能被很好地分离开来。因此，细胞浆分割的两大难题都能得到解决。

本文所提出的基于 EM 聚类和 SVM 自动学习的白细胞图像分割完整流程图如图 1 所示。

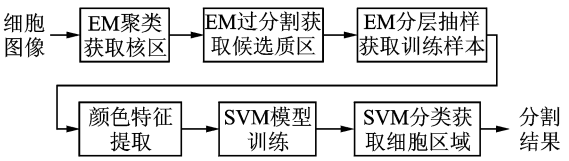

图 1　白细胞图像分割流程图

3　结果与讨论

对白细胞图像做分割实验，实验所采取的图像采集自 CellaVision 图像库(http://www.cellavision.com)，共包含 60 幅白细胞图像，每幅图含有一个白细胞，图像大小均为 360 像素×360 像素。实验环境为：Windows XP，VC6++，Pentium 2.6 GHz，2 GB。本文方法选用 RBF 函数作为 SVM 的核函数，参数固定，$\gamma=1$，$C=100$。

为了分析算法性能，由医生对每幅图像进行手动分割获取细胞核、浆和背景区域，并采用准确率(P)、召回率(R)和 F_1 值(M_{F1})分别评价核区和浆区的分割效果，并统计算法耗时

$$P=tp/(tp+fp)\tag{11}$$

$$R=tp/(tp+fn)\tag{12}$$

$$M_{F1}=\frac{2\times P\times R}{P+R}\tag{13}$$

3.1　不同训练样本总数和采样比例对 SVM 分割结果的影响

不同训练样本总数和采样比例会导致 SVM 分割结果不同，本文分别分析样本总数和采样比例对细胞区域分割结果的影响。首先固定采样比例 $\alpha=0.5$(即细胞区域与背景区域各占 1/2)，分析不同样本总数 N 对分割性能的影响。如表 1 所示，由于本文获取的候选样本区域比较精确，因此训练样本数对分割性能影响并不大，各项指标均在 95%左右。

接着，固定样本总数 $N=1\,000$，分析不同采样比例 α 对分割性能的影响。如表 2 所示，α 越大，则细胞区域采集样本数越多，像素更容易被分为细胞，因此细胞区域召回率越高，准确率越低。当 $\alpha=0.5$ 时，M_{F1} 最高，因此选择 $\alpha=0.5$ 作为采样比例。

表 1　不同训练样本总数对分割性能的影响

N	P/%	R/%	M_{F1}/%	t/s
500	95.65	94.32	94.52	1.89
1 000	95.29	95.62	95.03	2.47
1 500	94.45	95.67	94.59	2.72
2 000	94.35	95.78	94.63	3.11

表 2 不同比例对分割性能的影响

α	P/%	R/%	M_{F1}/%
0.1	97.87	81.09	86.47
0.2	98.11	87.38	91.45
0.3	97.81	91.39	94.07
0.4	96.20	94.51	94.50
0.5	95.29	95.62	95.03
0.6	93.95	96.51	94.75
0.7	90.87	97.45	93.12
0.8	87.89	98.55	91.70
0.9	78.98	99.12	85.88

3.2 不同方法分割结果对比

本文提出方法与 Ko 算法[4]和 Pan 算法[10]进行了实验对比，部分分割对比图见图 2，其中图 2(a)是医生手动分割的结果，图 2(b)是 Ko 算法的结果，图 2(c,d)分别是 Pan 算法和本文算法的分割结果。分割性能对比结果如图 3 所示，算法耗时对比见表 3。

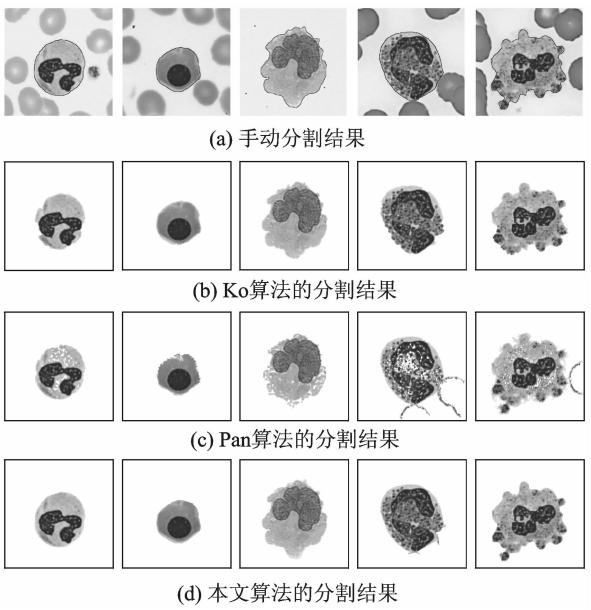

图 2 白细胞图像分割结果对比

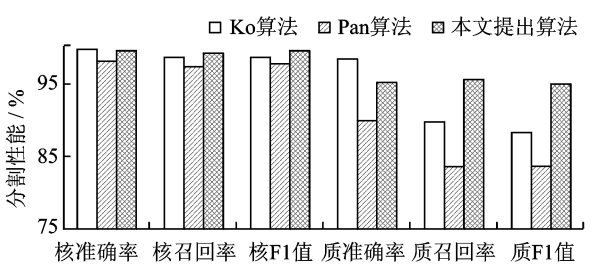

图 3 白细胞图像分割性能对比

表 3 不同分割方法耗时对比

方法	Ko 方法	Pan 方法	本文方法
耗时/s	56.8	2.02	2.47

3.3 讨 论

从图 2 可以看到，由于染色图像中细胞核颜色非常显著，3 种方法的细胞核分割精确度都较高，本文算法精确度略高。由于胞浆区域颜色变化范围大，还存在红细胞粘连的现象，因此分割难度比细胞核区域分割难度更大。基于主动轮廓模型的 Ko 算法的分割精度较高，但分割耗时过长，不满足临床诊断的实时性需求。基于 SVM 分类的 Pan 算法结果存在含有孔洞、欠分割和过分割的问题，有的孔洞过大，即使用形态学操作对其进行修正，仍不能获取完整的细胞区域。这主要是由候选正负样本不当引起的。本文算法采用了基于 EM 过分割的核区膨胀获取了更准确的候选浆区，同时利用基于 EM 的分层抽样采集样本，训练样本比 Pan 算法获取的样本更具代表性，因此本文算法的分割效果优于 Pan 的算法。

4 结束语

本文提出了一种对颜色变化鲁棒的白细胞图像分割算法，首先利用 EM 聚类获取细胞核区，再利用 SVM 自动采样学习分割细胞浆区域。在细胞核区域分割阶段，根据细胞核亮度低以及细胞浆和红细胞亮度接近的先验知识，通过 EM 聚类获取细胞核分布模型；细胞浆区域分割阶段，先根据细胞浆包裹细胞核的先验，利用基于 EM 过分割的细胞核区膨胀和分层抽样自动获取正负样本，再对像素点颜色特征进行 SVM 在线学习分类，获得细胞区域。将来的主要工作是，通过增加纹理特征进一步提高分割精度。

参考文献：

[1] 张时民. 五分类法血细胞分析仪测定原理和散点图特征[J]. 中国医疗器械信息，2008，14(12)：1-9.
Zhang Shimin. The Determination principles and scattergram characters of Leukocyte 5-part differential hematology analyzer[J]. China Medical Devices Information，2008，14(12)：1-9.

[2] Wang S T，Min W. A new detection algorithm (NDA) based on fuzzy cellular neural networks for white blood cell detection[J]. IEEE Transactions on Information Technology in Biomedicine，2006，10(1)：5-10.

[3] 蔡隽,鲍旭东,吴磊,等.基于活动轮廓模型的彩色白细胞图像自动分割方法研究[J].生物医学工程研究,2005,24(4):218-222.
Cai Jun, Bao Xudong, Wu Lei, et al. Automated segmentation of colored Leukocyte images based on the active contour model[J]. Journal of Biomedical Engineering Research, 2005,24(4):218-222.

[4] Ko B C, Gim J W, Nam J Y. Automatic white blood cell segmentation using stepwise merging rules and gradient vector flow snake[J]. Micron,2011,42(7):695-705.

[5] 黄敏,朱晓,朱启兵,等.基于主动轮廓模型的玉米种子高光谱图像分类[J].数据采集与处理,2013,28(3):289-293.
Huang Min, Zhu Xiao, Zhu Qibing, et al. Hyperspectral image classification of maize seeds based on active contour model[J]. Journal of Data Acquisition and Processing,2013,28(3):289-293.

[6] 印勇,王云,刘丹平.血细胞图像分割的改进 MEANSHIFT 方法[J].计算机工程与应用,2010(6):178-180.
Yin Yong, Wang Yun, Liu Danping. Improved MeanShift method for blood cell image segmentation [J]. Computer Engineering and Applications, 2010(6):178-180.

[7] Theera-Umpon N. White blood cell segmentation and classification in microscopic bone marrow images [M]. Berlin: Springer-Verlag Berlin, 2005, 3614: 787-796.

[8] Guo N N, Zeng L B, Wu Q S. A method based on multispectral imaging technique for white blood cell segmentation[J]. Computers in Biology and Medicine,2007, 37(1): 70-76.

[9] Burges C C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998,2(2):121-127.

[10] Pan C, Park D S, Yoon S, et al. Leukocyte image segmentation using simulated visual attention [J]. Expert Systems With Applications, 2012, 39 (8): 7479-7494.

[11] Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm [J]. Journal of the Royal Statistical Society: Series B,1977,39(1):31-38.

[12] Wang X Y, Zhang X J, Yang H Y, et al. A pixel-based color image segmentation using support vector machine and fuzzy C-means[J]. Neural Networks, 2012,33:148-159.

[13] 汪友生,胡百乐,张丽杰,等.基于支持向量机的动脉硬化斑块识别[J].数据采集与处理,2012,27(3):283-286.
Wang YouSheng, Hu Baile, Zhang Lijie, et al. Recognition of atherosclerotic plaque based on support vector machine[J]. Journal of Data Acquisition and Processing,2012,27(3):283-286.

作者简介:郑馨(1987-),女,博士研究生,研究方向:目标检测与识别、医学图像分析与处理,E-mail:zxaoyou@gmail.com;王勇(1989-),男,硕士研究生,研究方向:数字图像处理和模式识别;汪国有(1965-),男,博士,教授,研究方向:自动目标识别、图像与视频压缩。