第 34 卷 第 2 期 计 算 机 工 程 2008 年 1 月
Vol.34 No.2 Computer Engineering January 2008
·博士论文· 文章编号：1000—3428(2008)02—0014—03 文献标识码：A 中图分类号：TP18

基于形态特征和 SVM 的血液细胞核自动分析

曾 明[1]，孟庆浩[1]，张建勋[2]，鲍菁丹[1]
(1. 天津大学电气与自动化工程学院，天津 300072；2. 南开大学机器人与信息自动化研究所，天津 300071)

摘 要：以形态学分析和支持向量机为基础，构建了一套血细胞核显微图像自动分析与识别系统。在细胞核分割阶段，提出基于支持向量机的血液细胞核彩色图像分割算法。在特征提取环节中，除使用常规形态特征外，提出了一种新的能有效反映核分叶数差异的形态特征——腐蚀退化因子。采用“one-against-one”策略的多分类 SVM 方法对血细胞进行分类识别。实验测试表明，该系统具有较高的识别精度，平均识别率达 94.13%。

关键词：血液细胞核；图像分割；支持向量机；腐蚀退化因子

Automatic Analysis System of Blood Cell Nuclei Based on Morphological Features and Support Vector Machines

ZENG Ming[1], MENG Qing-hao[1], ZHANG Jian-xun[2], BAO Jing-dan[1]
(1. School of Electrical Engineering and Automation, Tianjin University, Tianjin 300072;
2. Institute of Robotics & Automatic Information System, Nankai University, Tianjin 300071)

【Abstract】 Based on morphological analysis and Support Vector Machines (SVM), a robust automatic analysis system of blood cell nuclei is developed. A novel algorithm for color image segmentation of blood cell nuclei based on the SVM is proposed. A new morphological feature named erosion degenerate factor is used to indicate the lobulated state of cell and achieves feature extraction by combining with traditional characteristics. One-against-one multi-class SVM is applied to classify the blood cell. Experimental results show that the proposed system yields better performance with the average recognition rate of 94.13%.

【Key words】 blood cell nuclei; image segmentation; Support Vector Machines(SVM); erosion degenerate factor

血细胞自动分析与识别是生物医学工程领域的研究热点，众多研究者倾力于此方面的研究，并取得了一些进展[1-2]。然而，大多数成果依然不能直接应用于临床，其主要原因有：(1)细胞显微图像易受染色、光照及噪声等因素的影响，已有的分割算法很难获得满意的分割效果，一般需要人工修复，才能用于后续分析；(2)同一类型细胞存在多态性，不同类型细胞之间的形态差异界定模糊，同时，缺乏针对特定细胞有效的特征描述算子。本文构建的系统将用于嗜中性粒细胞核像分析，即分析分叶核粒细胞(segmented granulocyte)和杆状核粒细胞(stab granulocyte)在血液中的比重变化。细胞核图像分割和形态特征提取是血液自动分析系统的 2 个关键环节。在分割阶段，本文提出一种基于向量机(Support Vector Machines, SVM)的血细胞核彩色图像分割算法。将SVM算法用于图像分割的优势在于：SVM是一类性能优良的分类器；便于引入分割相关的先验知识，如通过典型样本的选取和学习，可不断优化算法的性能，这是其他传统分割算法无法比拟的。在特征提取环节中，除使用常规形态特征外，本文还提出了一种新的能有效反映核分叶数差异的形态特征——腐蚀退化因子。

1 基于 SVM 的细胞核图像分割

1.1 支持向量机原理[3]

1.1.1 线性可分 SVM

考虑一个 2 类训练样本集的分类问题：

$D=\{(x_1,y_1),(x_2,y_2)\cdots,(x_n,y_n)\},\boldsymbol{x}\in R^n,y\in\{-1,1\}$

存在如下超平面：$\langle \boldsymbol{w},\boldsymbol{x}\rangle+b=0$，使得训练样本集完全正确分开，同时满足距离超平面最近的两类点间隔最大，称样本集被超平面最优划分。归一化超平面方程，使得对所有样本集满足如下约束条件：

$$y_i\left(\langle \boldsymbol{w},\boldsymbol{x}_i\rangle+b\right)\geq 1,\quad i=1,2,\cdots,n \tag{1}$$

此时分类间隔为 $2/\|\boldsymbol{w}\|$，最大间隔等价于使 $\|\boldsymbol{w}\|^2$ 最小。寻找最优分类超平面问题，可以转化如下二次规划问题求解：

$$\begin{cases}\min\ \dfrac{1}{2}\|\boldsymbol{w}\|^2\\ \text{s.t. } y_i\left(\langle \boldsymbol{w},\boldsymbol{x}_i\rangle+b\right)\geq 1,\quad i=1,2,\cdots,n\end{cases} \tag{2}$$

依据 Lagrange 对偶理论，将式(2)的二次规划问题变换为如下便于求解的对偶形式：

$$\begin{cases}Q(\boldsymbol{\alpha})=\sum\limits_{k=1}^{n}\alpha_k-\dfrac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_i\alpha_j y_i y_j\langle \boldsymbol{x}_i,\boldsymbol{x}_j\rangle\\ \text{s.t. }\sum\limits_{i=1}^{n}\alpha_i y_i=0,\quad 0\leq\alpha_i,\quad i=1,2,\cdots,n\end{cases} \tag{3}$$

其中，$\boldsymbol{\alpha}$为Lagrange乘子。求解以上优化问题，得最优解：$\boldsymbol{\alpha}^*=(\alpha_1^*,\alpha_2^*\cdots,\alpha_n^*)^{\mathrm{T}}$，进而计算得 $\boldsymbol{w}^*$ 和 b^*。依据KKT互补条件，其中只有少量最靠近超平面样本点的α_i值不为 0，Vapnik等人

基金项目：国家自然科学基金资助项目(60475028)
作者简介：曾 明(1973－)，男，博士，主研方向：生物医学图像处理，模式识别，智能信息处理；孟庆浩、张建勋，教授、博士生导师；鲍菁丹，硕士
收稿日期：2007-01-25 **E-mail：**yongshi@mail.nankai.edu.cn

称之为支持向量(SV)。最终决策函数为

$$f(\boldsymbol{x}) = \mathrm{sgn}\left(\sum_{\boldsymbol{x}\in SV} \alpha_i^* y_i \langle \boldsymbol{x} \cdot \boldsymbol{x}_i \rangle + b^* \right) \tag{4}$$

1.1.2 线性不可分 SVM——C-SVM

以上讨论仅限定在训练样本数据是线性可分的情况。然而，实际中存在大量线性不可分情况，1995 年 Cortes 和 Vapnik 提出在条件(1)中引入松弛项 $\xi_i \geq 0$，成为

$$y_i\left(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b\right) \geq 1 - \xi_i, \quad i = 1, 2, \cdots, n \tag{5}$$

将目标函数改为求

$$\Phi(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n} \xi_i \tag{6}$$

最小，折中考虑最少错分样本和最大分类间隔，得到广义最优超平面。其中，惩罚参数 C 作为综合这 2 个目标的权重。求解广义最优超平面的对偶问题与线性可分情况几乎完全相同，只是约束条件变为：$0 \leq \alpha_i \leq C$。最优决策函数的形式与式(4)一样。

由于对偶形式中只出现两向量的内积运算，Vapnik等人提出采用满足条件的核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ 来代替内积运算，即 $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \varphi(\boldsymbol{x}_i), \varphi(\boldsymbol{x}_j) \rangle$ 实现非线性软间隔分类。Mercer条件的核函数 $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$来代替内积运算，实现非线性软间隔分类。常用的核函数包括多项式核、径向基(RBF)核以及Sigmoid核等。

1.2 分割算法描述

分割算法实现步骤如下：

(1)截选具有代表性的白细胞核区域和非核区域(包括正常红细胞区域、白细胞胞浆区和空白区)。定义白细胞核区像素的彩色分量(如 RGB 分量)为 SVM 训练的正类样本，非核区域为负类样本。

(2)将标记好的样本采用序列最小优化方法(Sequential Minimal Optimization, SMO)对支持向量机进行训练，利用实验测试和网格搜索法(grid-search)确定最优 SVM 模型及模型参数，最终得到细胞分割决策函数。

利用训练好的分割决策函数，对待处理的图像中的像素进行分析，由决策函数的输出值确定图像中像素的所属类别(核区或非核区)，由此得到分割的二值图像。

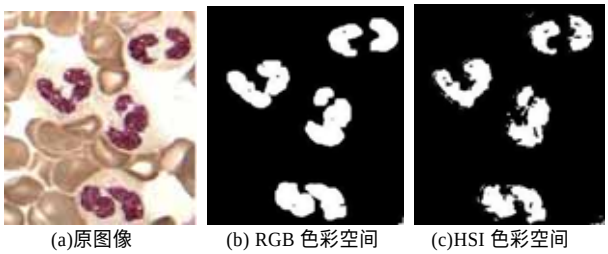
1.3 细胞核分割实验与分析

1.3.1 样本数据的采集

本文实验采用瑞氏(Wright)染色法处理的含有嗜中性分叶核细胞、嗜中性杆状细胞的慢性髓性白血病(CML)的血液涂片(由于在制片的过程中不可避免的会损伤部分细胞，因此在分析过程中还须对这类细胞进行处理)。血细胞图像经高倍光学显微镜(100×物镜，10×目镜)放大后，由彩色 CCD 摄像机采集，并传入到计算机，系统量化精度 24 bit，空间分辨率为 0.253 μm，白细胞尺寸大小为 5 μm~20 μm，采集的图像大小为 780×580，共采集了不同涂片上的 200 幅细胞图像，从中截取细胞核小图和非核区域小图各 60 幅，构成训练测试样本集。测试所用机器配置为 P4 2.4 GHz，512 MB 内存。

1.3.2 最佳色彩空间的选择

色彩空间的选择将直接影响彩色图像分割算法的性能，因此，本文将对 RGB 和 HSI(Hue- Saturation-Intensity)2 种色彩空间表达的细胞图像分割效果进行测试和比较。不同色彩空间和色彩分量的分割效果比较如图 1 所示。采用 RGB 和 HSI 色彩空间的分割结果比较见表 1，分析时间为实际大图 780×580 的平均处理时间，算法精度是 100 幅测试小图(正负类各 50 幅)的测试结果。表 1 的数据表明，采用 RGB 和 HSI 都可获得比较理想的分割结果，但 HSI 空间对细胞内染色不均分割过于灵敏，而出现少量的伪边界，见图 1(c)。从分析速度方面比较，采用 HSI 空间的分析速度是 RGB 的 2 倍~3 倍。图 1(b)和图 1(c)展示了 2 种色彩空间的一个分割实例效果。总地来说，采用 RGB 色彩空间，无论在分割精度和速度上都优于 HSI 空间。



(a)原图像　(b) RGB 色彩空间　(c)HSI 色彩空间

图 1　不同色彩空间和色彩分量的分割效果比较

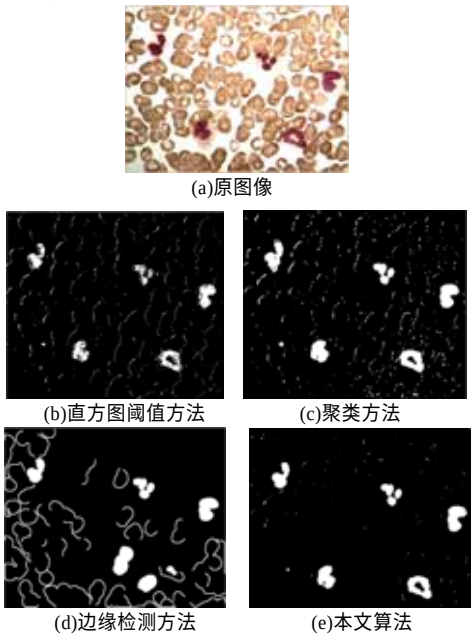
表 1　采用 RGB 和 HSI 色彩空间的分割结果比较

色彩空间	正类样本数量	负类样本数量	支持向量	测试分割正确率/(%)	平均分析时间/s
RGB	4 097	2 910	161	99.155	7.87
HSI	4 097	2 910	214	98.892	21.98

1.3.3 与其他算法的效果比较

为了评判本文算法的分割效果，这里将本文算法与 3 种经典分割算法作了比较和分析，它们是基于直方图阈值分割算法、K 均值聚类分割算法、边缘检测结合形态学混合分割算法。

不同算法的分割效果如图 2 所示，图 2(b)为直方图阈值分割算法的结果，由于目标区域小，因此无法通过 Otsu 和 Chow-Kaneko 等自适应阈值选取方法自动找到目标区域与背景区域的最佳分割阈值，只能通过累试方法人工参与完成，算法难以推广。此外，细胞核区域完整性较差。应用 K 均值聚类分割算法处理后细胞核区域完整(图 2(c))，特别是当目标区域较大(聚类 K 值较小)时效果较好，但算法实现过程中(当 K 值较大时)易出现局部极小问题，且最优聚类数 K 的确定与图像内容有关，易出现欠分割和过分割现象。此外，算法分析时间较长(平均约 60 s)。



(a)原图像

(b)直方图阈值方法　(c)聚类方法

(d)边缘检测方法　(e)本文算法

图 2　不同算法的分割效果

图 2(d)是边缘检测(采用 Canny 算子)结合形态学运算混合方法的分割结果。算法最大优势是速度较快，与本文算法的分析时间相当，但分割的不连续性问题和伪边界问题很难克服，易生成非细胞核区域，且最佳梯度阈值区间的选取也需要依据不同图像对比度进行调整。

本文算法结果如图 2(e)所示，细胞核区域完整，非细胞核区噪声较少。通过训练获得分割决策函数后，图像分析过程不需要人工参与，算法鲁棒性强、通用性好，且分割精度完全满足后续分析要求。

2 细胞核形态特征提取

2.1 常规血细胞核形态特征

常规血细胞核特征有多种，文献[4]提出对于粒系细胞，分析与识别，细胞核面积、核的圆形度、核的占凸比、平均胞浆色调及平均胞浆亮度为对分类最有效的 5 个参数。因此，本文选用了前 3 个作为核分析的常规特征，即面积特征 $F(1)$、圆形度 $F(2)$和占凸比 $F(3)$。

2.2 腐蚀退化因子

嗜中性粒细胞核像分析的关键是对核的分叶状态进行准确测定。部分分叶细胞核经分割处理后会分裂为多个子瓣，分叶状态测定较容易，但有些分叶核由于叶间有细丝相连(如图 3 所示)，经分割处理后子瓣之间并未分离，其形态与杆状核非常相似，现有的常规形态特征描述算子很难准确测定这部分分叶细胞的分叶状态，因此，有必要设计新的形态描述算子，区分子瓣粘连的分叶核和杆状核。通过观察发现，分叶核叶间虽有细丝相连，但连接关系与杆状核形态相比较弱，而受损细胞核形状相对杆状核更为规整(见图 3)。本文选择形态学腐蚀算子对各类细胞核核区图像进行处理，结果显示，各类细胞核经多次腐蚀后分化能力方面存在明显差异，通常情况有细丝相连的分叶核经 1 次~3 次腐蚀后就会分裂为多个子瓣，杆状核 4 次~8 次会分化，受损核经有限次腐蚀后一般不会分化。利用以上特性，本文提出了更具特异性的细胞核特征描述算子——腐蚀退化因子。

腐蚀退化因子的测算是通过对不同腐蚀次数后分化的细胞核的分叶数减去与腐蚀次数相对应的惩罚值(腐蚀次数多则惩罚值大)求得。该因子可作为一种通用的形态学分析算子，用于表征无规则形态物体结构元素的复杂程度。

腐蚀退化因子 $F(4)$测算公式如下：

$$F(4)=\begin{cases} N & N\geq 2 \\ \overline{N}-\phi\times C & \phi<top(n) \\ 1-top(n)\times C & \phi=top(n) \end{cases} \tag{7}$$

其中，N 为腐蚀前细胞的分叶数(细胞单连通域个数)；$\overline{N}$ 为经 ϕ 次腐蚀后细胞分化的分叶数；C 为惩罚因子(一般取 0.25)；$top(n)$为限定的腐蚀次数(一般取 10~12)。

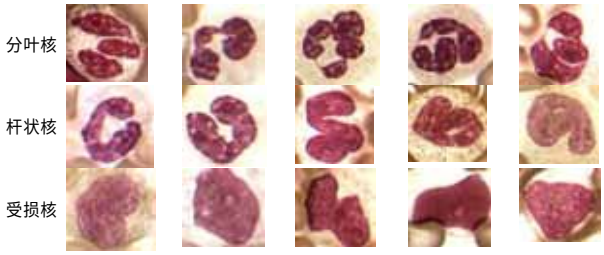
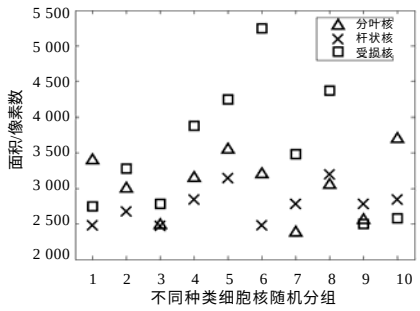
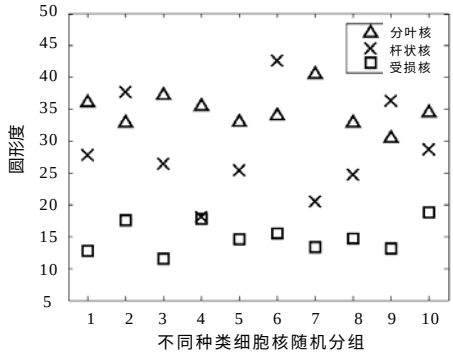


图 3 不同类型的细胞核图像

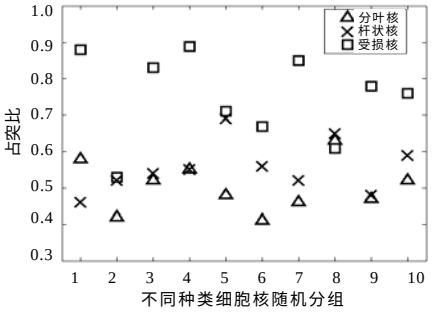
图 4 为随机抽取的 30 个不同种类细胞核各形态特征分布图。与常规特征相比，腐蚀退化因子能比较理想地区分 3 类细胞核，其数值的大小也基本反映了不同核的分叶数差异。



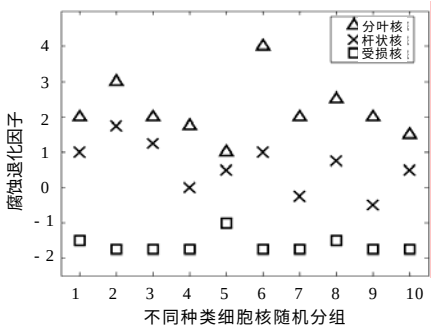
(a)各类细胞核面积特征的测算结果



(b)各类细胞核圆形度特征的测算结果



(c)各类细胞核占凸比特征的测算结果



(d)各类细胞核腐蚀退化因子的测算结果

图 4 不同类型细胞核的形态特征分布

3 基于多分类 SVM 的细胞核分类

细胞核分析的最后一个环节是对特征参数表达的细胞核进行分类识别。本文采集了 360 幅不同类型的细胞核图像(每一类 120 幅)进行分类测试，首先由血液医师预先进行分类，作为系统正确分类结果判定依据。分析系统采用one-against-one策略[5]的SVM多分类方法进行分类处理。不同核函数的分类效果如表 2 所示。可以看出，采用RBF核的效果最佳。表 3 给出了RBF核分类器各类别分类结果的混淆矩阵

(下转第 19 页)