一种更高效的尿沉渣自动识别算法

张灿龙[1],唐艳平[2],王 强[1],韦春荣[3]
ZHANG Can-long[1],TANG Yan-ping[2],WANG Qiang[1],WEI Chun-rong[3]

1.广西师范大学 计算机科学与信息工程学院,广西 桂林 541004
2.桂林电子科技大学 信息材料科学与工程系,广西 桂林 541004
3.广西师范大学 物理与电子工程学院,广西 桂林 541004
1.College of Computer Science and Information Engineering,Guangxi Normal University,Guilin,Guangxi 541004,China
2.Department of Information Material Science and Engineering,Guilin University of Electronic Technology,Guilin,Guangxi 541004,China
3.College of Physics and Electronic Engineering,Guangxi Normal University,Guilin,Guangxi 541004,China
E-mail:clzhang@mailbox.gxnu.edu.cn

ZHANG Can-long,TANG Yan-ping,WANG Qiang,et al.A more effective algorithm of automatic recognition urinary sediment.Computer Engineering and Applications,2010,46(3):232-235.

Abstract: An algorithm based on Support Vector Machine(SVM) and template matching is designed to classify the urinary sediment.Firstly,visible compositions in urinary sediment are classified roughly into big object and small object following their area,the former includes epithelial cells and pipe type,the latter includes crystallization,leukocyte and erythrocyte.Secondly,the template matching method is used to identify the crystallization,and by constructing an optimal SVM classifier,leukocyte and erythrocyte are different in the most extent.Finally,these big objects are classified into epithelial cells and pipe type following narrow extent.Experiment shows that the proposed method not only can gain better recognition rate(96.7%),but also reduces 22.4% computing time.

Key words: urinary sediment;Support Vector Machine(SVM);template matching;narrow extent

摘 要:提出了一种综合支持向量机(Support Vector Machine,SVM)和模板匹配的尿沉渣识别算法。首先根据面积特征将有形成分粗分成大目标类和小目标类,然后对小目标类中的草酸钙结晶以模板匹配法识别,而红、白细胞采用SVM的方法进行分类,最后对大目标类中的上皮细胞和管型则根据其狭长度加以区分。实验表明,该算法在将尿沉渣识别率提高到96.7%的同时还节约了22.4%的识别时间。

关键词:尿沉渣;持向量机;模板匹配;狭长度
DOI:10.3778/j.issn.1002-8331.2010.03.071 **文章编号:**1002-8331(2010)03-0232-04 **文献标识码:**A **中图分类号:**TP391.41

尿常规检查是指通过分析尿沉渣中的有形成分来完成泌尿系统疾病的诊断。由于尿沉渣成分的复杂性,让计算机代替人来完成各有形成分的自动分析是一个非常困难而又富有挑战性意义的课题。其关键技术在于对沉渣有形成分的分割和分类。文献[1-3]提出过一些分类方法,但这些分类方法的形式过于单一,缺乏区别对待,故未能取得理想的分类效果。

尿沉渣有形成分在形态、色态和面积上的不同是分类算法选择的重要依据。从总体上来说,红细胞、白细胞和草酸钙结晶面积较小,而管型和上皮细胞面积较大;红、白细胞呈圆盘形,而草酸钙结晶具有某种特定的模式,管型呈长条形,上皮细胞为鼓状凸多边形;红细胞颜色偏白,白细胞颜色偏黑,上皮细胞颜色较浅。在设计分类算法时,应该分阶段、有区别地对待,例如,首先根据面积大小将红、白细胞和草酸钙结晶归为小目标类、上皮细胞和管型细胞归为大目标类;然后采用模板匹配法将小目标类中的草酸钙结晶的识别出来,而红、白细胞的区分则可采用SVM的方法, 最后对大目标类中管型和上皮细胞则可根据其外形的狭长性加以分类。该文就是利用尿沉渣形态、色态和面积的不同,综合运用支持向量机和模板匹配等方法进行分类识别。

1 尿沉渣分割与特征表示

尿沉渣显微图像中, 有形成分和背景之间的区分度较低,故分割结果的好坏直接影响有形成分的特征提取和分类。论文采用了一种融合Snake模型和Canny边缘定位法的尿沉渣图

基金项目:广西省教育厅资助科研项目(the Research Project of Department of Education of Guangxi,China under Grant No.200508107);广西师范大学青年基金(No.师政科技[2006]5号)。
作者简介:张灿龙(1975-),男,讲师,研究领域为图像处理与模式识别;唐艳平(1977-),女,讲师,研究领域为计算机应用;王强(1952-),男,博士,教授,研究领域为人工智能与模式识别;韦春荣(1975-),女,讲师,研究领域为图像处理与模式识别。
收稿日期:2008-08-05 **修回日期:**2008-10-23

像分割算法,在文献[4]中对该算法有详细描述,并证明了该算法具有使用灵活、检测到的轮廓准确、完整等优点。图 1 展示了这一算法的分割结果(图(b)中的黑线代表了分割结果的轮廓曲线)。

分类识别主要依据有形成分的特征,所以提取最有效的特征就成为关键。通过 K-L 变换并结合专家意见从众多特征中遴选出以下一些最具分类信息的特征:面积(S)、圆形度(Shape)、光密度(IOD)、光滑度(V)、核浆面积比(HJA)、核浆灰度比(HJG)、狭长度(E)以及纹理特征中的熵(EN),具体计算公式见后续内容。

(a)尿沉渣图像

(b)闭合轮廓

图 1　尿沉渣分割

(从左到右依次为红细胞、白细胞、结晶体、上皮细胞和管型)

2　更高效的尿沉渣自动识别法

2.1　基于面积特征的粗分类

通过对大量尿沉渣样本进行分析后发现:红、白细胞和草酸钙结晶的面积一般较小,而管型细胞和上皮细胞面积相对较大。根据这一线性可分的特点,在进行有形成分的归类时,先以面积作为依据将有形成分粗分为小目标和大目标两类,其中红、白细胞和草酸钙结晶作为小目标类,上皮细胞和管型细胞作为大目标类。

2.2　快速模板匹配法用于识别草酸钙结晶

从图 1 中可以看出,草酸钙结晶基本上呈正方形,内部包含一个星形,且与正方形的对角线吻合,星形的灰度值较高,而星形之外的灰度值较低,这些特征是小目标类中其他成分所不具备的。一种快速模板匹配法被设计,用于将结晶体从小目标类中识别出来。

设已知目标对象的图像模板为 T,待考察的图像 I,匹配的过程是设法把模板 T 叠加在图像 I 上,并比较 T 与它覆盖下的 I 的子图像的差别,若差别小于某事先设定的阈值则认为 T 在该处与 I 的子图像有较好的匹配,即找到了目标对象。对图像 I 按逐象素扫描并实施上述操作,即可确定其中是否存在模板 T 所确定的目标对象。上述算法为传统匹配算法,存在计算量大、匹配效率不高以及方位旋转和尺度缩放的不适应性等缺点。针对其缺点,文献[5]提出了一种基于旋转、平移、缩放不变(RST 不变)中心矩描述的快速模板匹配算法。下面给出该算法中归一化的不变一阶中心矩:

$$\varphi=\frac{\left(\sum_{i\in R}\sum_{j\in R}(i-\frac{m_{10}}{m_{00}})^2 f(i,j)+\sum_{i\in R}\sum_{j\in R}(j-\frac{m_{01}}{m_{00}})^2 f(i,j)\right)^2}{\left(\sum_{i\in R}\sum_{j\in R}f(i,j)\right)} \quad (1)$$

其中,$f(i,j)$是坐标为(i,j)的象素的灰度值,R 为模板区域或目标区域。

$$m_{00}=\sum_{i\in R}\sum_{j\in R}f(i,j)\quad m_{10}=\sum_{i\in R}\sum_{j\in R}i\times f(i,j)\quad m_{01}=\sum_{i\in R}\sum_{j\in R}j\times f(i,j) \quad (2)$$

针对本案例,建立图 2(b)所示模板 T 和图 2(c)所示的待识别目标(阴影所示)方位示意图。设模板和目标所示正方形的面积均为 S,大正方形的边长为 a,在实际的计算过程中,S 取有形成分区的象素数,a 取有形成分区上下跨度和左右跨度的算术均值。假设模板 T 中心点的坐标值为(0,0),则模板在点(x,y)处的灰度值 $T(x,y)$按下面的算法来确定:

(1)当$\sqrt{S/2}\times 2/3\leqslant|x|+|y|\leqslant\sqrt{S/2}$ 时,$T(x,y)=0$,0 代表黑色。

(2)当$|x|+|y|<\sqrt{S/2}\times 2/3$ 时,如果$|x|+5|y|\leqslant\sqrt{S/2}\times 2/3$ 或者$5|x|+|y|\leqslant\sqrt{S/2}\times 2/3$,则 $T(x,y)=1$,1 代表白色;其余点的灰度值 $T(x,y)=0$。

观察图 2(c),由几何关系可得方位角 θ 的计算公式如下:

$$\theta=\arcsin\frac{a-\sqrt{2S-a^2}}{2\sqrt{S}} \quad (3)$$

由于模板 T 的初始角度为 45°,所以将其旋转到图 2(c)所示位置要经过的旋转角度为 $\beta=45°-\theta$。

(a)结晶体二值图　(b)模板 T　(c)方位示意图

图 2　模板匹配识别结晶体

结晶体快速模板匹配算法的具体步骤如下:

(1)建立初始模板,选择合适的阈值对小目标类中待识别的有形成分进行二值化。按照式(1)、(2)求取二值图像的归一化不变一阶中心矩。

(2)将有形成分与模板 T 的中心矩比较,并取相差值小于某一域值的即可初步认为是模板所表示的有形成分。然后按式(3)计算有形成分的方位角 θ,并将模板的形心与有形成分的形心坐标重合。

(3)将模板进行旋转和缩放以使其与有形成分的方位和尺寸一致。缩放和旋转采用的算法分别为:

$$\begin{cases}x_2=N(x_1-x_0)+x_0\\ y_2=N(y_1-y_0)+y_0\end{cases}\quad\begin{cases}x_2=(x_1-x_0)\cos\beta+(y_1-y_0)\sin\beta+x_0\\ y_2=(y_1-y_0)\cos\beta+(x_1-x_0)\sin\beta+y_0\end{cases}$$

式中(x_2,y_2)为象素的新坐标,(x_1,y_1)为原坐标,(x_0,y_0)为模板的形心,$T(x_2,y_2)=T(x_1,y_1)$。$N=\sqrt{S_{模}/S_{物}}$ 为缩放系数,值为模板与有形成分面积比的方根,β 为旋转角度。

(4)对上述两幅已经对准的二值图像进行“异或”操作,以异点百分比的大小来判断是否为结晶体。

2.3　红细胞和白细胞的 SVM 分类

对于红细胞和白细胞,仅靠单一特征是无法区别的,需采用多特征分类,但它们在许多方面很相似,反映在特征空间中是类间距较短,交叉域较大,且样本数目有限,无法用线性分类器来进行分类。对这种小样本高维非线性两类问题,用 SVM 来解决是比较好的[6]。SVM 的基本思路是:首先用某种非线性映射将输入向量映射到高维特征空间,然后在这个高维空间中构造最优超平面,使两类之间的间隔最大,同时保证样本的分类误差尽可能小,详细的数学描述请参考文献[6]。下面给出最优分类函数:

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n}\alpha_i^* y_i K(x_i,x)+b^*\right) \quad (4)$$

其中 a^* 为最优解，b^* 是分类阈值，可以用任意一个支持向量求得，$K(x_i,x)$为核函数，它有多种形式，研究表明在细胞识别领域以高斯函数为核函数的 SVM 有很强的学习能力和较好的分类效果，所以该文采用高斯核函数 $K(x,x_i)=\exp(-\gamma\|x-x_i\|^2)$。式(4)的求和实际上只对支持向量进行。

高斯核 SVM 的主要参数有核参数 γ，误差惩罚因子 C。文献[6-7]的研究表明核参数 γ 和惩罚因子 C 的选择在很大程度上决定了 SVM 分类器的性能，所以该文采用 k-折交叉验证法对两参数进行优选。具体步骤如下：

(1)将训练样本集分成 k 个子集 $T_1,T_2,\cdots T_i,\cdots T_k$，各个子集包含大致相同数目的训练样本。

(2)在确定参数 γ 和 C 取值范围的基础上，对其进行均匀取值；设 γ 取 m 个值，C 取 n 个值，则构成了 $m\times n$ 个参数对。

(3)对第 j 个参数对(γ,C)，训练 k 个 SVM 分类器，其中第 i 个分类器(i=1,2,…,k)是以 T_i 中的样本作为测试集，而其余的样本(即$\overline{T_i}$)作为训练集而得到的，并得到其识别率 r_i。k 个分类器共 k 个识别率，取 $R_j=(r_1+r_2+\cdots+r_k)/k$ 为第 j 个参数对所对应的识别率。

(4)重复第(3)步共 $m\times n$ 次。对 $R_1,R_2,\cdots,R_{m\times n}$ 进行排序，选取其中最大者所对应的参数对(γ^*,C^*)作为高斯核 SVM 的最佳参数。

实际检测中 k=5，$\gamma\in\{2^{-14},2^{-12},2^{-10},\cdots,2^{-2},2^0,2^2\}$，$C\in\{2^{-5},2^{-3},2^{-1},\cdots,2^{11},2^{13},2^{15}\}$，$m$=9，$n$=16。通过训练可得出支持向量、$a^*$、$b^*$的值、最优参数对$(\gamma^*,C^*)$，然后以这些训练结果所构成的 model 对未知样本进行识别。

2.4 基于狭长度描述的大目标分类

大目标类中的上皮细胞外形较鼓、管型外形狭长(如图 1 所示)，故采用形状判据可以将两者区分开来。描述形状的方法有多种：一类是基于目标边界点的形状描述，如链码、曲率等[8-9]；另一类是基于目标内部区域的形状描述，如区域主轴、中轴变换等[10-11]。文章提出了一种基于区域狭长度描述的上皮细胞和管型分类算法。所谓狭长度是指一个区域长度与厚度之比，长度是一个不容易确定的因子，但是厚度确定之后，它的面积与长度有大致的正比关系，因此可采用面积来度量。这样，区域的狭长度可定义为：

$$E=\frac{S}{H^2} \tag{5}$$

其中，S 为有形成分面积，H 为有形成分厚度。分类算法具体步骤如下：

(1)对有形成分区用黑点进行区域填充(假设轮廓线以外区域为白色)，初始化 h=0。

(2)用结构元素为 3×3 的黑点块对有形成分进行一次腐蚀运算，h+1≥h。

(3)判断有形成分区是否还存在黑点，若存在则转向(2)，否则 2h≥H。

(4)用式(5)计算狭长度，若 E 大于事先设定的阈值则认定其为管型，否则为上皮细胞。

3 实验及结果分析

3.1 实验设计

为了测试所提出的分类算法的性能，文章从尿沉渣图像库中遴选出典型的结晶体，管型和上皮细胞样本各 50 个，红细胞和白细胞样本各 200 个组成训练集；再另选出同样数目和比例的尿沉渣样本组成测试集；设计三类测试实验：(1)采用该文所提出的综合法进行分类的实验；(2)仅用 SVM 进行分类的实验；(3)用文献[3]报道的方法进行分类的实验。实验所选用的特征变量及其计算公式和平均计算时间见表 1。

表 1 中 n 为有形成分的像素点个数，L 为轮廓的长度，$f(x_i,y_i)$为区内第 i 个像素点的灰度，$\overline{gray}$ 代表平均灰度，下标 H 和 J 分别代表细胞核和细胞浆，m_{ij} 为灰度共生矩阵中非 0 元素值，N_g 为灰度级数。

实验(1)中的 SVM 方法只适用于红细胞和白细胞组成的训练集，每个样本为 7 维特征向量，并将训练集平均分成 5 个子集，每个子集包括红细胞和白细胞样本各 40 个，用 5-折交叉验证法进行参数优选。实验(2)中每个样本为 9 维特征向量(在实验(1)的基础上增加了一阶中心矩特征和狭长度特征)，也将训练集平均分成 5 个子集，每个子集包括结晶体、管型和上皮细胞样本各 10 个，红、白细胞样本各 40 个，其他同实验(1)。SVM 训练机采用通用的 LibSVM，按一类对余类法进行分类训练。实验(3)是采用文献[3]中的方法，其基本思想是：在完成尿沉渣特征提取与选择的基础上，采用 BP 神经网络的方法进行训练和分类，此方法实际上可以看成是 SVM 方法中核函数为 Sigmoid 的一种回归。实验环境是 Intel 奔腾® D CPU2.8 GHz，内存 512 MB，算法用 VC 编程实现。

3.2 实验结果及分析

实验(1)中计算所得的一阶中心矩、狭长度以及面积值见表 2。评估的性能指标有各算法的识别时间、识别率、所得到的分类器个数以及 SVM 方法中的训练时间、支持向量数等。这里的识别时间是指计算单个样本特征的时间和用分类器对其归

表 1 特征变量及其计算公式和计算时间

Feature name	Formula	Computer time/ms	Feature name	Formula	Computer time/ms
S	$S=n$	10	IOD	$IOD=\frac{1}{n}\sum_{i=1}^{n}f(x_i,y_i)$	8
Shape	$Shape=\frac{4\pi S}{L^2}$	8	V	$V=\frac{1}{n}\sqrt{\sum_{i=1}^{n}[f(x_i,y_i)-IOD]^2}$	10
HJA	$HJG=\frac{\overline{gray_H}}{\overline{gray_J}}$	15	EN	$EN=-\sum_{i=1}^{N_g}\sum_{j=1}^{N_g}m_{ij}\log m_{ij}$	25
HJG	$HJA=\frac{S_H}{S_J}$	15			

表 2 一阶中心矩、狭长度和面积

一阶中心矩 φ				狭长度 E		面积 S	
模板	结晶体	红细胞	白细胞	上皮细胞	管型	大目标类	小目标类
0.026 72	0.026 356±0.000 5	0.163 56±0.002 5	0.162 69±0.002 5	≤2.36	≥10.53	≥550	≤450

表 3　实验(1)、(2)、(3)的实验结果

实验代号	训练时间/ms	支持向量数	(γ^*, C^*)	识别时间/ms				分类器个数	识别率/(%)
				算法 2.1	算法 2.2	算法 2.3	算法 2.4		
(1)	125	34	$(2^{-12}, 2^{15})$	12	28	100	15	4	96.37
(2)	253	69	$(2^{-1}, 2^{3})$			215		5	88.54
(3)	…	…	…			274		1	89.23

类的时间的总和，由于识别时间的长短会因样本不同而不同，所以最终结果取大量样本识别时间的期望。考虑到实验的随机性误差等因素，所有实验结果都是独立运行 20 次之后的平均值。实验结果如表 3 所示。

观察表 3 数据可以发现,实验(1)所获得的识别率和识别所耗的时间均明显优于其他两个实验,这说明该文所提出的分类算法是高效的。产生这样的结果,主要有以下几个原因:充分利用先验知识,减少不必要的特征计算。识别是以特征计算为前提,按实验(2)和(3)的算法进行识别的过程中,对每个样本都必须计算同样多的特征,但实际上某些特征的计算对一些样本类而言是必要的，而对另外一些样本类而言是完全没有必要,并且引入这些特征反而会加大分类器的错分率。譬如核浆面积比、核浆灰度比对区分红细胞和白细胞是重要的特征,但对区分结晶体、上皮细胞和管型却不起任何作用,因为它们根本不存在这样的特征,同样狭长度对区分红、白细胞不起作用,但对上皮细胞和管型的区分却至关重要。经统计发现一般红、白细胞和结晶体的面积比管型和上皮细胞要小许多,故仅依据面积特征就可将有形成分快速归入某一大类;管型和上皮细胞在外形上的显著区别使得通过狭长度就可将两者有效地区分开来;结晶体所独有的模式有助于其快速识别。文章所提出的算法就是综合利用先验知识,分阶段、有针对性地采用不同分类算法,达到缩短识别时间、提高识别率的目的;从全局出发,降低系统的统计识别时间。所谓统计识别时间是指一个检测系统在概率统计意义上的执行时间，其计算公式为 $T=\sum p_i \times t_i$,其中 i 为类别号,p_i 和 t_i 分别为第 i 类样本出现的概率和平均识别时间。一般而言在发生病变的尿液样本中红细胞、白细胞、结晶体、管型和上皮细胞出现的概率比为 0.35:0.35:0.1:0.1:0.1,则实验(1)的 $T=1\times12+0.1\times28+(0.35+0.35)\times100+(0.1+0.1)\times15=85.1$ ms,实验(2)的 $T=1\times(10+8+15+15+8+10+25+5+15)=111$ ms,可见实验(1)比(2)节约了 22.4%的时间。

4　结论

文章在分析了尿沉渣数据和现有尿沉渣识别文献的基础上,提出了一种尿沉渣自动识别算法。该算法是按照自顶向下、逐步细化的思想,分阶段、有针对性地采用模板匹配、SVM 和单一特征识别等分类算法。文章详细描述了各阶段的分类算法,并给出了详实的实验数据并进行了分析。实验结果表明,所提出的分类识别算法在将尿沉渣识别率提高到 96.7%的同时还节约了 22.4%的识别时间。该文的研究具有较强的实际应用价值,对其他从事相关领域研究的工作者有一定的参考价值。

参考文献:

[1] Linko S,Kouri T T,Toivonen E,et al.Analytical performance of the Iris iQ200 automated urine microscopy analyzer [J].Clin Chim Acta,2006,372:54-64.

[2] Chien T I,Kao J T.Urine sediment examination:A comparison of automated urinalysis systems and manual micros-copy [J].Clinica Chimica Acta,2007,384(1/2):28-34.

[3] 李勇明.尿沉渣图像自动识别算法的研究[D].重庆:重庆大学,2007:113-122.

[4] 唐艳平,张灿龙.基于形态学和 Snake 模型的尿沉渣提取[J].桂林电子科技大学学报,2007,27(6).

[5] 王强,宋京民,胡建平,等.一种快速模板匹配目标识别算法[J].计算机工程与应用,2000,36(6):42-43.

[6] Burges C J C.A tutorial on support vector machines for pattern recognition[J].Data Mining and Knowledge Discovery,1998,2(2):121-167.

[7] 张小云,刘允才.高斯核支撑向量机的性能分析[J].计算机工程,2003,29(8):22-25.

[8] Osada R.Shape Distributions[J].ACM Trans on Graphics,2002,21(4):807-832.

[9] Wolf son H J.On curve matching[J].IEEE Trans on Pattern Analysis and Machine Intelligence,1990,12(5):483-489.

[10] Synder W E,Qi Hairong.Machine vision[M].BeiJing:China Machine Press,2005.

[11] 梁光明,孙即祥.基于区域弦分布直方图的形状匹配算法及仿真[J].计算机工程与科学,2008,30(1):56-59.

(上接 227 页)

可以做到有效防范信用骗取。在将来的工作中,诸如建立电子商务信用骗取的预警模型,以及如何针对不同形式、不同影响的信用骗取行为完善信用评价体系等问题值得关注并进行系统性研究。

参考文献:

[1] Jones K,Leonard L N K.Trust in consumer-to-consumer electronic commerce[J].Information and Management,2008,45(2):88-95.

[2] 李征.基于ESS 均衡的电子商务信任模型[J].计算机应用,2008,28(8):2173-2176.

[3] 吴亮,文静.C2C 电子商务中的诚信问题研究[J].商场现代化,2007(4):139.

[4] 刘凤鸣,丁永生.基于生态网络的 P2P 环境信任博弈进化模型[J].计算机工程与应用,2007,43(23):24-27.

[5] 姜启源.数学模型[M].2 版.北京:高等教育出版社,1993.

[6] 金雪军,毛捷,袁佳.商业银行客户经理制有效性研究[J].南大商学评论,2004(3):98-121.