山西大学学报(自然科学版)38(3):420－425,2015
Journal of Shanxi University(Nat. Sci. Ed.)

文章编号:0253-2395(2015)03-0420-06

基于分层方法的白细胞五分类算法

赵建伟[1],张敏淑[1],周正华[1],楚建军[2],曹飞龙[1]*

(1.中国计量学院 理学院,浙江 杭州 310018;2.嘉善加斯戴克医疗器械有限公司,浙江 嘉兴 314100)

摘　要:针对白细胞自动识别问题,在已分割好的细胞基础上结合五类白细胞(中性粒细胞、嗜酸性粒细胞、嗜碱性粒细胞、单核细胞和淋巴细胞)的具体特征,利用分层的思想对其分类。首先提取白细胞中细胞核的分叶特征和圆形度特征,对该类特征明显的细胞进行筛选;而对该类特征不明显的细胞,提取对偶旋转不变共生局部二进制(PRICoLBP)纹理特征作为判定标准,将它们分为颗粒细胞与无颗粒细胞;然后对颗粒细胞,利用 PRICoLBP 纹理特征区分出嗜碱性粒细胞、嗜酸性粒细胞和中性粒细胞;而对无颗粒细胞,则用圆度核质比区分出淋巴细胞和单核细胞。实验表明,本文所提的方法比已有的分层方法在总体识别率上提高了十几个百分点,并且各类细胞的分类精度都有所提高。

关键词:白细胞分类;分层方法;纹理特征;对偶旋转不变共生局部二值模式(PRICoLBP)

中图分类号:TP301.6　　**文献标志码**:A　　**DOI**:10.13451/j.cnki.shanxi.univ(nat.sci.).2015.03.006

A Classification Algorithm for Five Types of White Blood Cells Based on Hierarchical Method

ZHAO Jianwei[1],ZHANG Minshu[1],ZHOU Zhenghua[1],CHU Jianjun[2],CAO Feilong[1]

(1. *Department of Information and Mathematics Sciences,China Jiliang University,Hangzhou* 310018,*China*;
2. *Jiashan Jasdaq Medical Device Co., Ltd. Jiaxing* 314100,*China*)

Abstract: For the problem of automatic identification of white blood cells, this paper proposes a classification algorithm for its five types (neutrophils, eosinophils, basophils, monocytes and lymphocytes) using their specific features based on the idea of hierarchical method. Firstly, we extract the leaflet feature and the circularity feature of the nuclei of white blood cells to screen the cells with these obvious characteristics. While for the cells without those obvious characteristics, we use the pairwise rotation invariant co-occurrence local binary pattern (PRICoLBP) feature to divide them into granular cells and nongranular cells. Finally, we divide the granular cells into basophils, eosinophils and neutrophils using PRICoLBP feature, and divide the nongranular cells into lymphocytes and monocytes using the ratio of the texture features and circularity features. Some experiments illustrate that our proposed method gets better performance than the existing corresponding hierarchical method.

Key words: white blood cell; hierarchical method; texture feature; pairwise rotation invariant co-occurrence local binary pattern (PRICoLBP)

收稿日期:2015-05-26;**修回日期**:2015-06-10

基金项目:浙江省自然科学基金(No. LY14A010027);国家自然科学基金(No. 61272023;91330118)

作者简介:赵建伟(1977－),女,博士,教授,主要研究方向为计算机视觉和模式识别等;*通信作者:曹飞龙(1965－),E-mail:icteam@163.com

0 引言

众所周知,血液中的白细胞(White Blood Cell,WBC)对人体免疫功能起着重要的作用。血液中各类白细胞的数量及百分比在人类有疾病和无疾病的情况下是不同的,医生经常用这些基础数据作为判断疾病的种类和严重程度的标准。因此,研究白细胞的分类计数对医学诊断有着重要的意义和价值。

通常,血液学家利用白细胞细胞质的颗粒信息及形状信息将白细胞分为颗粒细胞:中性粒细胞(neutrophils)、嗜酸性粒细胞(eosinophils)、嗜碱性粒细胞(basophils)和无颗粒细胞:单核细胞(monocytes)和淋巴细胞(lymphocytes)五类(见图 1)。鉴于细胞在不同时期的形态有很大的差别,本文中只考虑成熟时期的白细胞。

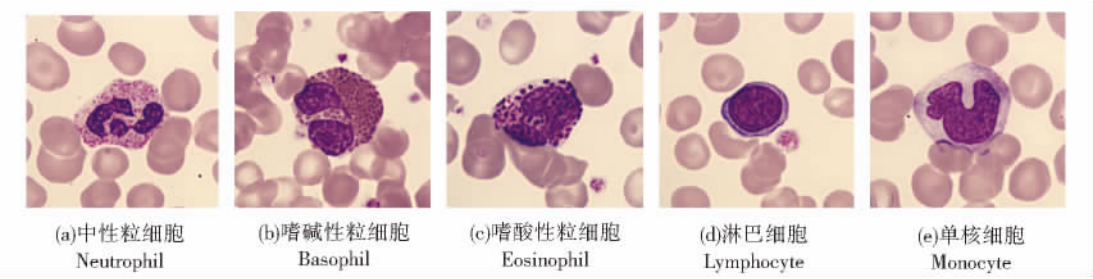

Fig. 1 Five types of WBC

图 1 五类白细胞示意图

传统的白细胞分类方法主要是医护人员在显微镜下根据经验进行白细胞计数。该方法存在工作量大、主观性强以及效率低等问题。随着计算机和人工智能技术的不断发展,基于图像分析技术的白细胞自动识别法已成为临床诊断的主要手段之一。它不仅可以解决上述人工计数方法存在的问题,而且具有图片可显示保存、以便于以后查验分类的优点。

通常,白细胞自动识别技术主要包括以下几个步骤:图像采集、白细胞图像的分割、特征提取和分类四个方面,其中特征提取和分类是白细胞自动识别技术的关键点和难点。因此,本文将在已分割好的白细胞图像上重点研究细胞的特征提取和分类问题。

白细胞的特征大致可分为三大类:几何特征(如面积、周长、圆形度和直径等)、纹理特征(如不变矩、对比度和熵等)和颜色特征(如颜色分布直方图等)。目前,许多文献利用以上特征和分类器的组合对白细胞进行分类。文献[1]利用白细胞的面积、核质比、模式谱及不变矩等特征并结合神经网络对白细胞进行了五分类,其分类精度达到了 75%。文献[2]利用已有的几何特征、纹理特征和颜色特征这三大类中的特征组成了 164 维的特征向量对白细胞进行了分类,取得了一定的效果。但是该方法需要提取大量的特征,这样会导致时间复杂性提高。文献[3]先是利用圆度特征区分颗粒细胞与无颗粒细胞,然后利用细胞质的颜色区分三类颗粒细胞以及利用核质比区分两类无颗粒细胞,其分类精度达到了 75.59%。该方法的缺点是在利用圆度特征区分颗粒细胞和无颗粒细胞时,没有充分利用颗粒信息及注意到单核细胞的细胞核形状不规则等问题,导致单核细胞和颗粒细胞分类效果不高的问题。

针对上述问题,本文基于分层的思想,重新设计分层路线。首先提取白细胞中细胞核的分叶特征和圆形度特征,分出部分细胞核较圆的淋巴细胞,对数据进行筛选以减少下面步骤错分的个数;其次对分叶特征和圆形度特征不明显的细胞,提取对偶旋转不变共生局部二值模式(PRICoLBP)纹理特征作为判定标准,将它们分为颗粒细胞与无颗粒细胞;然后对颗粒细胞,利用 PRICoLBP 纹理特征区分出嗜碱性粒细胞、嗜酸性粒细胞和中性粒细胞;而对无颗粒细胞,则用圆形度与核质比区分出淋巴细胞和单核细胞。实验表明,本文所提的方法比已有的分层方法在总体识别率上提高了十几个百分点,并且各类细胞的分类精度都有所提高。

1 本文所提的白细胞分类方法

本文根据五类白细胞的具体特征,基于分层的思想,重新设计了白细胞分类方法。该方法将白细胞的特征选择分为三层:顶层利用分叶数及圆形度对白细胞进行筛选;中层利用 PRICoLBP 特征作为颗粒信息将

白细胞分为颗粒细胞和无颗粒细胞;底层对于颗粒细胞利用其颗粒信息提取 PRICoLBP 特征对其进行三分类,而对于无颗粒细胞,提取其细胞核的圆度及核质比对其进行二分类。本文的分类器选择效果好的支持向量机(SVM)作为分类器。具体流程见图 2,其中绿色代表顶层,黄色代表中层,蓝色代表底层。

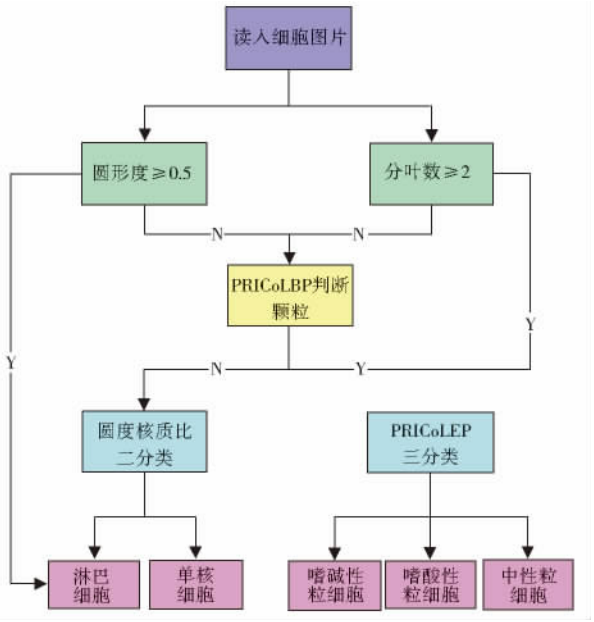

Fig. 2 Proposed classification scheme for WBC

图 2 本文所提的白细胞分类方案

1.1 特征提取

众所周知,选择区分度高、有代表性的特征是提高白细胞分类精度的关键。本小节将详细阐述本文所选取的白细胞重要特征。

1.1.1 顶层特征选取

因为部分淋巴细胞与其他细胞相比,体积小且其细胞核相对较圆,所以本文先利用圆形度特征来筛选出部分淋巴细胞,以减少下面步骤错分的个数。白细胞圆形度 $D_{\text{圆形度}}$ 计算公式如下:

$$D_{\text{圆形度}} = \frac{4\pi S_{\text{核}}}{C^2}, \tag{1}$$

其中 $S_{\text{核}}$ 是细胞核的面积,C 是细胞核的周长。显然,当细胞核为圆形时,其圆形度为 1。

众所周知,在基于分层思想的方法中,白细胞一旦在某层错分,则在后继的分类中必将错分,这样会影响最终的分类效果。根据血液学的实验统计[4]可知,血液中分叶数大于等于 2 的中性粒细胞占白细胞总数的 50%—70%。为了避免错分,本文先用分叶数特征筛选出部分白细胞直接进入底层分类。分叶数的确定主要是根据细胞的连通性,利用腐蚀的方法来做。若细胞核不是分叶的,则其连通个数为 1 或被腐蚀到 0;若细胞核是分叶的,则其连通数大于等于 2。选取其稳定的连通个数记为分叶数,其效果如图 3 所示。

Fig. 3 Erosion for leaflet

图 3 腐蚀分叶示意图

1.1.2 中层颗粒二分类

对于白细胞中的颗粒信息,目前方法基本上都是从全局提取颗粒特征,如图像的能量、方差、熵和平滑度等信息,从未考虑其局部特征。文献[5]验证了灰度共生矩阵比局部二值模式(LBP)[6]效果好,但计算灰度共生矩阵需要耗费时间多。基于此,本文引入文献[7]中的 PRICoLBP 特征描述颗粒信息,将细胞区分为颗粒细胞和无颗粒细胞。该特征不但体现空间结构性质,还对图片具有旋转不变性。

LBP 是一种描述图像纹理特征的算子,在点 A 及其 3×3 的窗口上,以窗口中心 A 的像素灰度值作为阈值,周围 8 个像素灰度值与其进行比较。若大于中心阈值,则其值置为 1,反之为 0。这样生成 8 个二进制数,然后转换成十进制作为该中心像素取得的函数值,以此反映该窗口区域的纹理信息。其数学公式描述如下:

$$LBP(A) = \sum_{i=0}^{n-1} s(g_i - g_c)2^i, \tag{2}$$

其中 g_i, g_c 分别代表第 i 个位置的像素值和中心点的像素值,且

$$s(x) = \begin{cases} 1, x \geqslant 0 \\ 0, x < 0 \end{cases}. \tag{3}$$

在 LBP 的基础上,文献[7]基于共生的思想提出了 PRICoLBP 特征,使该特征能更好地表示空间结构信息。其公式描述如下:

$$PRICoLBP(A,B) = [LBP^{ru}(A), LBP^{u}(B, i(A))]_{\omega}, \tag{4}$$

其中 $LBP^{ru}(A)$ 为旋转不变局部二值模式,$LBP^{u}(B,i(A))$ 为均匀局部二值模式,$i(A)=\underset{i\in\{0,1,2,\cdots,n-1\}}{\operatorname{argmax}}\{ROR(LBP(A),i)\}$ 是取点 A 的二值序列最大的下标作为点 B 的二值序列的起始点(如图 4 所示),从而保证共生 LBP 的旋转不变性。根据共生 LBP 的计算方法,以点 A 的梯度方向和法线方向分别作为 x 轴、y 轴,统计相应模式的个数直方图作为其特征,记为共生 LBP 特征。

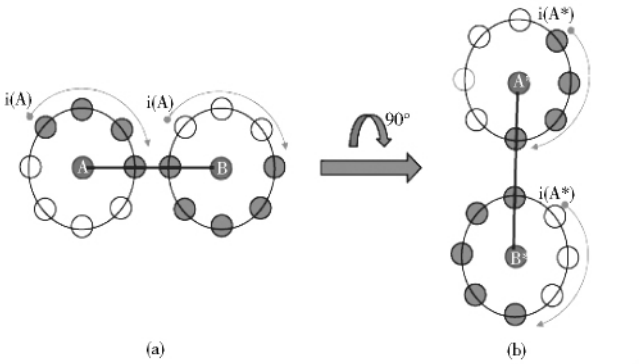

Fig. 4 Rotation invariance of PRICoLBP

图 4 PRICoLBP 的旋转不变性

直方图的相似性一般利用卡方距离来刻画,但由于我们只计算细胞本身(不包含背景)的共生 LBP 特征,而细胞的大小是不一样的,且由于细胞分割误差的出现,这就造成了在归一化的时候同一类细胞的直方图差异较大,为此我们引入了 BRD[8]。设 $p=[p_1,p_2,\cdots,p_n]$ 与 $q=[q_1,q_2,\cdots,q_n]$ 是直方图向量,则其 BRD 的计算公式如下:

$$d_{BRD}(p,q)=\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{p_iq_j-p_jq_i}{p_i+q_j}\right)^2. \tag{5}$$

至此,本文利用 PRICoLBP 特征将白细胞区分为颗粒细胞和无颗粒细胞。

1.1.3 底层细分类

无颗粒白细胞包括单核细胞和淋巴细胞,这两类细胞的体积分别是五类白细胞中最大和最小的。单核细胞的细胞质比淋巴细胞的细胞质要多的多,且单核细胞的细胞核通常是不规则的,而淋巴细胞的细胞核一般是类圆的。因此,本文选取圆度核质比区分单核细胞和淋巴细胞,其数学计算公式为:

$$D_{\text{核质比}}=\frac{S_{\text{核}}}{S_{\text{质}}}, \tag{6}$$

其中 $S_{\text{核}}$ 代表细胞核的面积,$S_{\text{质}}$ 代表细胞质的面积。

对颗粒细胞,利用 1.1.2 中的 PRICoLBP 纹理特征区分出嗜碱性粒细胞、嗜酸性粒细胞和中性粒细胞。

1.2 SVM 分类器

本小节将对 2.1 节中所提取的各个特征,选取支持向量机(SVM)[9-10]作为分类器进行分类。SVM 基于结构风险最小化原理,而非传统的经验风险最小化原理,从而能兼顾训练误差和泛化能力。并且 SVM 不存在过学习问题,得到的解是全局最优解,因此,它具有更好的泛化性能。SVM 分类器的具体操作如下:

给定一组训练样本集 $\{(x_i,y_i)\}_{i=1}^{n}\subseteq R^p\times\{-1,1\}$,SVM 的训练过程相当于求解下面的优化问题:

$$\begin{cases}\min\limits_{\alpha}\left\{\dfrac{1}{2}\alpha^T\boldsymbol{Q}\alpha-e^T\alpha\right\}\\ \text{s.t. } 0\leqslant\alpha_i\leqslant C,i=1,2,\cdots,n,\\ y^T\alpha=0,\end{cases} \tag{7}$$

其中 e 是分量全为 1 的向量,C 为常数,Q 是 n 阶矩阵,每个分量 $Q_{ij}=y_iy_jK(x_i,x_j)$,其中 $K(x_i,x_j)$ 为核函数,常用的核函数为高斯函数或 Sigmoid 函数。

上述 SVM 的优化过程可以转化为求解其 Lagrange 乘子的对偶问题,即

$$\begin{cases}\max\limits_{\alpha}\left\{\sum\limits_{i=1}^{n}\alpha_i-\dfrac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_i\alpha_jy_iy_jK(x_i,x_j)\right\},i=1,2,\cdots,n,\\ \sum\limits_{i=1}^{n}\alpha_iy_i=0,\\ 0\leqslant\alpha_i\leqslant C,\end{cases} \tag{8}$$

则确定的分类器为

$$f(x)=\sum_{i=1}^{n}\alpha_iy_iK(x_i,x). \tag{9}$$

本文实验中用的高斯核函数为

$$K(x,y)=\exp\left(-\frac{1}{A}d(x,y)\right), \tag{10}$$

其中常数 A 由直方图距离均值取得[11]。

2 实验

本实验中所用的数据来自瑞典 Cellavision 数据库[12]及嘉善加斯戴克公司提供的白细胞图片。图片尺寸均为 300 像素×300 像素,且每张图片只含一个白细胞(如图 1 所示)。其中加斯戴克数据库包含的中性粒细胞、嗜酸性粒细胞、嗜碱性粒细胞、淋巴细胞和单核细胞数量分别为 479、15、16、269、21 个;Cellavision 数据库中对应各类数量分别为 30、15、16、20、16 个。利用自适应阈值分割及 Grabcut 算法对图片提取细胞核及细胞质得到本文实验的数据库(如图 5 和图 6 所示)。本文的实验在 MATLAB 2014b 上运行。

Fig. 5 Cellavision database

图 5 Cellavision 数据库

Fig. 6 Jasdaq database

图 6 加斯戴克数据库

表 1 与表 2 分别是本文所提的方法在 Cellavision 数据库与加斯戴克数据库进行测试的结果。实验结果显示,本文的方法对于颗粒细胞的分类精度都在 90%以上,表明了本文所提的分类算法是合理且有效的。由于单核细胞的形状不规则,少量单核细胞的细胞核为类圆,所以使得其与淋巴细胞的区分度不是太明显。

表 1 本文所提的方法在 Cellavision 数据库上的测试结果

Table 1 Recognition of our proposed algorithm on Cellavision Database

	识别为嗜碱性细胞个数	识别为嗜酸性细胞个数	识别为淋巴细胞个数	识别为单核细胞个数	识别为中性粒细胞个数	识别精度
嗜碱性细胞(16 个)	15		1			93.75%
嗜酸性细胞(15 个)		15				100%
淋巴细胞(20 个)			19	1		95.00%
单核细胞(15 个)			4	11		80.00%
中性粒细胞(30 个)				1	29	96.67%

表 2 本文所提的方法在 Jasdaq 数据库上的测试结果

Table 2 Recognition of our proposed algorithm on Jasdaq Database

	识别为嗜碱性细胞个数	识别为嗜酸性细胞个数	识别为淋巴细胞个数	识别为单核细胞个数	识别为中性粒细胞个数	识别精度
嗜碱性细胞(16 个)	16					100%
嗜酸性细胞(15 个)		14	1			93.33%
淋巴细胞(269 个)			232	22	15	86.25%
单核细胞(21 个)		2	2	17		80.95%
中性粒细胞(479 个)		2	4	21	422	93.99%

表 3 和表 4 分别是本文所提的方法与文献[3]中的分层方法在 Cellavision 数据库以及加斯戴克数据库上的实验结果。从表中可以看出,本文所提的方法在每一类白细胞的分类效果上都要远远好于文献[3]中的结果,从而进一步验证了本文方法的优越性。

另外,在 Cellavision 数据上,本文的方法与文献[3]中的方法所花的测试时间分别为 13.19 s 和2.576 s,在 Jasdaq 数据库上所花的测试时间分别为 111.71 s 和 21.98 s。虽然本文所花的测试时间是文献[3]中方法的 6 倍,但是我们所提算法的精度是远远高于文献[3]的精度,并且在 2 min 之内测试将近 800 张图片也

是符合当前对白细胞检测的需求的。因此,我们所提的方法还是合理且有效的。

表 3 本文所提的方法与文献[3]中的 HSVM 方法在 Cellavision 数据库上的实验比较结果

Table 3 Recognition comparison of our proposed algorithm with the HSVM method in paper [3] on CellavisionDatabase

	嗜碱性细胞识别精度	嗜酸性细胞识别精度	淋巴细胞识别精度	单核细胞识别精度	中性粒细胞识别精度	总识别精度
文献[3]中 HSVM 方法	31.25%	13.33%	75.00%	40.00%	90.00%	57.29%
本文所提的方法	93.75%	100%	95.00%	80.00%	96.67%	92.71%

表 4 本文所提的方法与文献[3]中的 HSVM 方法在 Jasdaq 数据库上的实验比较结果

Table 4 Recognition comparison of our proposed algorithm with the HSVM method in paper [3] on Jasdaq Database

	嗜碱性细胞识别精度	嗜酸性细胞识别精度	淋巴细胞识别精度	单核细胞识别精度	中性粒细胞识别精度	总识别精度
文献[3]中 HSVM 方法	56.25%	80.00%	82.13%	40.00%	73.50%	75.59%
本文所提的方法	100%	93.33%	86.25%	80.95%	93.99%	91.04%

3 总结

本文从白细胞的具体特征出发,加入颗粒之间的结构与颜色信息以及细胞的形状特征,并利用分层的思想对其进行分类,实现白细胞的自动分类。在 Cellavision 数据库与嘉善加斯戴克公司数据库上的实验测试表明,与以往的白细胞分层方法相比,本文所提的分类方法的精度远远高于文献[3]中的结果。

参考文献:

[1] Nipon Theera-Umpon. Automatic White Blood cell Classification Using Biased-output Neural Networks with Morphological Features[J]. *Thammasat Int J Sc Tech*, 2003, **8**(1): 64-71.

[2] Osowski S, Siroic R, Markiewicz T, *et al*. Application of Support Vector Machine and Genetic Algorithm for Improved Blood Cell Recognition[J]. *Instrumentation and Measurement, IEEE Transactions on*, 2009, **58**(7): 2159-2168.

[3] Tai W L, Hu R M, Hsiao H C W, *et al*. Blood Cell Image Classification Based on Hierarchical SVM[C]//*Multimedia (ISM), 2011 IEEE International Symposium on*. *IEEE*, 2011: 129-136.

[4] Berk A, Zipursky S L. Molecular Cell Biology[M]. New York: WH Freeman, 2000.

[5] Rezatofighi S H, Soltanian-Zadeh H. Automatic Recognition of Five Types of White Blood Cells in Peripheral Blood[J]. *Computerized Medical Imaging and Graphics*, 2011, **35**(4): 333-343.

[6] Ojala T, Pietikainen M, Maenpaa T. Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns[J]. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2002, **24**(7): 971-987.

[7] Qi X, Xiao R, Li C G, *et al*. Pairwise Rotation Invariant Co-occurrence Local Binary Pattern[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2014, **11**: 2199-2213.

[8] Hu W, Xie N, Hu R, *et al*. Bin Ratio-based Histogram Distances and Their Application to Image Classification[J]. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2014, **12**: 2338-2352.

[9] Vapnik V N. An Overview of Statistical Learning Theory[J]. *IEEE Trans. Neural Networks*, 1999, **10**(5): 988-999.

[10] Chih-Chung Chang, Chih-Jen Lin. LIBSVM: a Library for Support Vector Machines[J]. *ACM Transactions on Intelligent Systems and Technology*, 2011, **2**(3): 1-27.

[11] Zhang J, Marszalek M, Lazebnik S, *et al*. Local Features and Kernels for Classification of Texture and Object Categories: A comprehensive study[J]. *Int J Comput Vis*, 2007, **73**(2): 213-238.

[12] Karin N. Cells in Peripheral Blood. CellaVision Inc, 2000, http://www.cellavision.com/?id=3651.