



(12)发明专利申请

(10)申请公布号 CN 108564114 A

(43)申请公布日 2018.09.21

(21)申请号 201810262889.9

(22)申请日 2018.03.28

(71)申请人 电子科技大学

地址 611731 四川省成都市高新区(西区)
西源大道2006号

申请人 宁波摩米创新工场电子科技有限公司

(72)发明人 刘娟秀 王祥舟 申志杰 邓鼎文
杜晓辉 赵家喜 张静 倪光明
郝如茜 刘霖 刘永 周辉

(74)专利代理机构 电子科技大学专利中心
51203

代理人 陈一鑫

(51)Int.Cl.

G06K 9/62(2006.01)

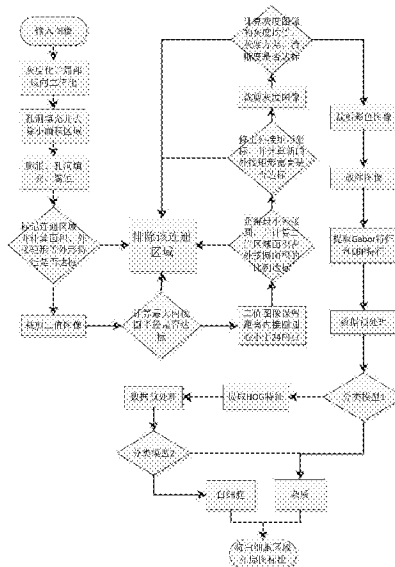
权利要求书3页 说明书5页 附图3页

(54)发明名称

一种基于机器学习的人体粪便白细胞自动识别方法

(57)摘要

本发明公开了一种基于机器学习的人体粪便白细胞自动识别方法,涉及基于数字图像处理技术,特别是机器学习技术对人体粪便中白细胞自动识别检测方法。本发明采用对样本图像进行特征提取,采用提取的特征对SVM支持向量机进行训练,然后再将训练好的向量机应用于白细胞的识别,本发明方法具有识别准确,速度快,效率高,计算量小的优点。



1. 一种基于机器学习的人体粪便白细胞自动识别方法,该方法包括以下步骤:

步骤1:对人体粪便进行稀释、搅拌、沉淀的预处理,使用生物显微镜采集人体粪便样本的显微图像;

步骤2:对步骤1中的图像进行灰度化处理,转为灰度图像;

步骤3:对步骤2中的灰度图像进行局部反向二值化处理,获得二值化图像;

步骤4:对步骤3中的二值图像进行孔洞填充操作,填充面积小于 S_1 的区域,其中 S_1 的取值范围为180~230个像素;

步骤5:对步骤4中的二值图像进行形态学膨胀运算,采用半径为 R_1 的圆形模板;对膨胀后的二值图像进行孔洞填充;对孔洞填充后的二值图像进行形态学腐蚀运算,采用半径为 R_2 的圆形模板; R_1 的取值范围为2~5个像素, R_2 的取值范围为1~3个像素,且 R_1 大于 R_2 ;

步骤6:将步骤5的二值图像连通区域标记,并计算连通区域的面积特征和外接矩形;保留面积大于 S_2 ,且占外接矩形面积比例大于 D_1 ;外接矩形宽高大于 H_1 ,小于 H_2 的区域; S_2 的取值范围为1400~1550个像素, D_1 的取值范围为60%~80%, H_1 的取值范围为30~50个像素, H_2 的取值范围为100~120个像素;

步骤7:对步骤6保留的区域,根据外接矩形裁剪二值图像;对裁剪的二值图像,首先计算二值图像的最大内接圆的圆心和半径,仅保留内接圆半径大于 R_3 的区域;对二值图像进一步操作,只保留距离圆心小于 L_1 的点; R_3 的取值范围为15~20个像素, L_1 的取值范围为20~30个像素;

步骤8:对步骤7的二值图像,计算二值图像的最小外接圆,保留二值区域面积占外接圆面积 D_2 以上的二值图像; D_2 的取值范围为72%~90%

步骤9:对步骤8的二值图像重新计算并修正外接矩形坐标;保留外接矩形宽、高大于 H_3 ,宽高绝对值之差小于 H_4 的区域; H_3 的取值范围为50~57个像素, H_4 的取值范围为12~18个像素;

步骤10:对步骤9保留的区域,根据新计算的外接矩形坐标,裁剪对应的灰度图和彩图;计算灰度图的方差、灰度均值和清晰度;保留方差大于 F_1 ,灰度均值大于 G_1 ,清晰度大于 Q_1 的区域; F_1 的取值范围为295~308, G_1 的取值范围为55~70, Q_1 的取值范围为11~20;

以下分为训练分类器和实际检测两个过程;

步骤11~步骤16为分类器训练过程;步骤17~步骤20为实际检测过程;

步骤11:将步骤10中保留的区域坐标与对应的人工框选白细胞区域坐标对比,去除真实的白细胞,并截取剩余区域的彩图作为负样本集的杂质;训练使用的正样本集采用人工框选的白细胞;步骤4~步骤10所有采用的参数,均可保证所有样本中的白细胞进入步骤10后的被保留区域;

步骤12:对所有的样本集,将样本图像进行归一化处理,然后分别提取三个颜色通道的Gabor特征和LBP特征;

步骤13:将步骤12中样本集提取的特征进行预处理;

步骤13-1:首先对特征向量进行PCA主成分分析,仅保留99%的主要成分;

步骤13-2:对步骤13-1得到的特征进行'L2'规范化;

步骤13-3:对步骤13-2得到的特征进行标准化;

步骤14:使用k-折交叉验证的方法,将样本集等分为 N 份, $N>3$,每次使用其中一份或两

份作为训练集,其余的作为测试集;

步骤15:对步骤14中的训练集,使用SVM支持向量机模型进行训练,获得分类模型1;

步骤16:根据步骤15中获得的分类模型1,训练分类模型2;

步骤17:裁剪步骤10中被保留区域对应的彩色图像,提取Gabor和LBP特征并组合,按步骤13的方法进行数据预处理,获得待测的特征向量;

步骤18:将步骤17中的待测特征向量输入至分类器1,获得分类结果;

步骤19:提取步骤18中结果为白细胞的彩色图像HOG特征,按步骤13的方法进行数据预处理,并将处理后的特征向量送入分类器2中,获得最终的分类结果;

步骤20:将步骤19中所有检测到的白细胞在步骤1的彩色显微图像上标记出来。

2.如权利要求1所述的一种基于机器学习的人体粪便白细胞自动识别方法,其特征在于所述步骤3中局部区域二值化的方法为:处理局部的区域为方形,边长为11个像素,阈值选取方法为灰度均值,阈值放缩系数为0.98。

3.如权利要求1所述的一种基于机器学习的人体粪便白细胞自动识别方法,其特征在于所述步骤10中的清晰度计算采用二次模糊的方法,具体步骤为:

步骤10-1:计算灰度图像每个像素点的梯度值,获得原梯度图。

步骤10-2:对灰度图进行高斯模糊,模板大小为9*9,方差取1.5。

步骤10-3:对经过10-2步骤模糊后的灰度图像计算每个像素点的梯度值,获得模糊梯度图。

步骤10-4:计算步骤10-1得到的原梯度图和步骤10-3得到的模糊梯度图的平均绝对差值,即获得清晰度值。

4.如权利要求1所述的一种基于机器学习的人体粪便白细胞自动识别方法,其特征在于所述步骤12的具体方法为:

将图像缩放至80*80大小,然后分别提取三个颜色通道的Gabor特征和LBP特征;其中Gabor特征尺度选择15;方向选择 $\{0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ\}$ 8个方向;波长取 $\pi/4$;空间纵横比取1.0;标准差取1.5;相位偏移为0;同时采用8倍降采样降低特征维度;其中LBP特征采用LBP等价模式,组合Gabor特征和LBP特征。

5.如权利要求1所述的一种基于机器学习的人体粪便白细胞自动识别方法,其特征在于所述步骤15中:

使用SVM支持向量机模型进行训练时,选择RBF核函数,gamma为自动选取,惩罚系数C取值为2,正负样本权重比值为2:1。

6.如权利要求1所述的一种基于机器学习的人体粪便白细胞自动识别方法,其特征在于所述步骤16中的具体步骤为:

步骤16-1:将步骤15中模型1所有误检的负样本作为模型2的负样本集。正样本集为放缩后的人工框选白细胞。

步骤16-2:对样本集的灰度图提取HOG (Histogram of Oriented Gradient) 方向梯度直方图特征。Cell大小取8*8,梯度取9个方向;Block大小取2*2;步长取8。获得3600维特征向量。

步骤16-3:对样本集特征采用步骤13中数据预处理方式,获得2320维特征向量。

步骤16-4:采用与模型1相同训练方法,使用SVM模型,获得分类模型2,达到在保留白细

胞的基础上进一步去除误检杂质的目的。最后选择RBF核函数, γ 为自动选取; 惩罚系数 C 取1; 正负样本权重比值为6:1, 其他均为默认参数。

一种基于机器学习的人体粪便白细胞自动识别方法

技术领域

[0001] 本发明是涉及基于数字图像处理技术,特别是机器学习技术对人体粪便中白细胞自动识别检测方法。

背景技术

[0002] 人体粪便中的白细胞常在粘液及脓血便中检出,且多为中性分叶核粒细胞。由于细胞退变,细胞多胀大,结构不清。如数量较多,成堆存在,且细胞膜不完整或已破碎,此时亦称为脓细胞,表示感染严重。不同于人体的血液与尿液,人体的粪便中含有大量食物残渣等杂质,导致图像背景复杂;部分白细胞细胞膜不再完整,且与部分杂质粘连。以上原因均导致白细胞不易分割。由于细胞的退变,白细胞的形态多变,不易使用基于形态学的方法识别。

[0003] 本申请人14年申请过专利一种针对支气管肺泡灌洗涂片中白细胞的自动检测方法(专利号为2014103696359)该发明公开了通过对显微镜采集支气管肺泡灌洗液涂片的显微图像进行灰度化、二值化处理,同时利用白细胞的外形特征与内部特征进行筛选,最终识别出白细胞,采用纯图像处理的方法检测样本中的白细胞,实际应用时计算量大。

发明内容

[0004] 本发明的目的是针对粪便显微图像背景复杂及白细胞形态变化多样性,设计了一种针对人体粪便中白细胞的自动识别方法,该方法基于数字图像处理技术和机器学习技术,从人体粪便显微图像中自动识别白细胞,从而达到检测准确率高、漏检率低的目的,满足临床检测的需要。

[0005] 本发明提供的技术方案为一种基于机器学习的人体粪便白细胞自动识别方法,该方法包括以下步骤:

[0006] 步骤1:对人体粪便进行稀释、搅拌、沉淀的预处理,使用生物显微镜采集人体粪便样本的显微图像;

[0007] 步骤2:对步骤1中的图像进行灰度化处理,转为灰度图像;

[0008] 步骤3:对步骤2中的灰度图像进行局部反向二值化处理,获得二值化图像;

[0009] 步骤4:对步骤3中的二值图像进行孔洞填充操作,填充面积小于 S_1 的区域,其中 S_1 的取值范围为180~230个像素;

[0010] 步骤5:对步骤4中的二值图像进行形态学膨胀运算,采用半径为 R_1 的圆形模板;对膨胀后的二值图像进行孔洞填充;对孔洞填充后的二值图像进行形态学腐蚀运算,采用半径为 R_2 的圆形模板; R_1 的取值范围为2~5个像素, R_2 的取值范围为1~3个像素,且 R_1 大于 R_2 ;

[0011] 步骤6:将步骤5的二值图像连通区域标记,并计算连通区域的面积特征和外接矩形;保留面积大于 S_2 ,且占外接矩形面积比例大于 D_1 ;外接矩形宽高大于 H_1 ,小于 H_2 的区域; S_2 的取值范围为1400~1550个像素, D_1 的取值范围为60%~80%, H_1 的取值范围为30~50

个像素, H_2 的取值范围为100~120个像素;

[0012] 步骤7:对步骤6保留的区域,根据外接矩形裁剪二值图像;对裁剪的二值图像,首先计算二值图像的最大内接圆的圆心和半径,仅保留内接圆半径大于 R_3 的区域;对二值图像进一步操作,只保留距离圆心小于 L_1 的点; R_3 的取值范围为15~20个像素, L_1 的取值范围为20~30个像素;

[0013] 步骤8:对步骤7的二值图像,计算二值图像的最小外接圆,保留二值区域面积占外接圆面积 D_2 以上的二值图像; D_2 的取值范围为72%~90%

[0014] 步骤9:对步骤8的二值图像重新计算并修正外接矩形坐标;保留外接矩形宽、高大于 H_3 ,宽高绝对值之差小于 H_4 的区域; H_3 的取值范围为50~57个像素, H_4 的取值范围为12~18个像素;

[0015] 步骤10:对步骤9保留的区域,根据新计算的外接矩形坐标,裁剪对应的灰度图和彩图;计算灰度图的方差、灰度均值和清晰度;保留方差大于 F_1 ,灰度均值大于 G_1 ,清晰度大于 Q_1 的区域; F_1 的取值范围为295~308, G_1 的取值范围为55~70, Q_1 的取值范围为11~20;

[0016] 以下分为训练分类器和实际检测两个过程;

[0017] 步骤11~步骤16为分类器训练过程;步骤17~步骤20为实际检测过程;

[0018] 步骤11:将步骤10中保留的区域坐标与对应的人工框选白细胞区域坐标对比,去除真实的白细胞,并截取剩余区域的彩图作为负样本集的杂质;训练使用的正样本集采用人工框选的白细胞;步骤4~步骤10所有采用的参数,均可保证所有样本中的白细胞进入步骤10后的被保留区域;

[0019] 步骤12:对所有的样本集,将样本图像进行归一化处理,然后分别提取三个颜色通道的Gabor特征和LBP特征;

[0020] 步骤13:将步骤12中样本集提取的特征进行预处理;

[0021] 步骤13-1:首先对特征向量进行PCA (Principal Component Analysis) 主成分分析,仅保留99%的主要成分;

[0022] 步骤13-2:对步骤13-1得到的特征进行'L2'规范化;

[0023] 步骤13-3:对步骤13-2得到的特征进行标准化;

[0024] 步骤14:使用k-折交叉验证的方法,将样本集等分为 N 份, $N > 3$,每次使用其中一份或两份作为训练集,其余的作为测试集;

[0025] 步骤15:对步骤14中的训练集,使用SVM (Support Vector Machine) 支持向量机模型进行训练,获得分类模型1;

[0026] 步骤16:根据步骤15中获得的分类模型1,训练分类模型2;

[0027] 步骤17:裁剪步骤10中被保留区域对应的彩色图像,提取Gabor和LBP特征并组合,按步骤13的方法进行数据预处理,获得待测的特征向量;

[0028] 步骤18:将步骤17中的待测特征向量输入至分类器1,获得分类结果;

[0029] 步骤19:提取步骤18中结果为白细胞的彩色图像HOG (Histogram of Oriented Gradient) 特征,按步骤13的方法进行数据预处理,并将处理后的特征向量送入分类器2中,获得最终的分类结果;

[0030] 步骤20:将步骤19中所有检测到的白细胞在步骤1的彩色显微图像上标记出来。

[0031] 进一步的,所述步骤3中局部区域二值化的方法为:处理局部的区域为方形,边长

为11个像素,阈值选取方法为灰度均值,阈值放缩系数为0.98。

[0032] 进一步的,所述步骤10中的清晰度计算采用二次模糊的方法,具体步骤为:

[0033] 步骤10-1:计算灰度图像每个像素点的梯度值,获得原梯度图。

[0034] 步骤10-2:对灰度图进行高斯模糊,模板大小为9*9,方差取1.5。

[0035] 步骤10-3:对经过10-2步骤模糊后的灰度图像计算每个像素点的梯度值,获得模糊梯度图。

[0036] 步骤10-4:计算步骤10-1得到的原梯度图和步骤10-3得到的模糊梯度图的平均绝对差值,即获得清晰度值。

[0037] 进一步的,所述步骤12的具体方法为:

[0038] 将图像缩放至80*80大小,然后分别提取三个颜色通道的Gabor特征和LBP特征;其中Gabor特征尺度选择15;方向选择 $\{0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ\}$ 8个方向;波长取 $\pi/4$;空间纵横比取1.0;标准差取1.5;相位偏移为0;同时采用8倍降采样降低特征维度;其中LBP (Local Binary Pattern) 特征采用LBP等价模式 (Uniform Pattern), 组合Gabor特征和LBP特征。

[0039] 进一步的,所述步骤15中:

[0040] 使用SVM支持向量机模型进行训练时,选择RBF核函数, gamma为自动选取,惩罚系数C取值为2,正负样本权重比值为2:1;

[0041] 所述步骤16中的具体步骤为:

[0042] 步骤16-1:将步骤15中模型1所有误检的负样本作为模型2的负样本集。正样本集为放缩后的人工框选白细胞。

[0043] 步骤16-2:对样本集的灰度图提取HOG (Histogram of Oriented Gradient) 方向梯度直方图特征。Cell大小取8*8,梯度取9个方向;Block大小取2*2;步长取8。获得3600维特征向量。

[0044] 步骤16-3:对样本集特征采用步骤13中数据预处理方式,获得2320维特征向量。

[0045] 步骤16-4:采用与模型1相同训练方法,使用SVM模型,获得分类模型2,达到在保留白细胞的基础上进一步去除误检杂质的目的。最后选择RBF核函数, gamma为自动选取;惩罚系数C取1;正负样本权重比值为6:1,其他均为默认参数。

[0046] 本发明采用对样本图像进行特征提取,采用提取的特征对SVM支持向量机进行训练,然后再将训练好的向量机应用于白细胞的识别,本发明方法具有识别准确,速度快,效率高,计算量小的优点。

附图说明

[0047] 图1是本发明一种针对人体粪便中白细胞的自动识别方法检测流程图。

[0048] 图2是对应的训练分类模型的流程图。

[0049] 图3是最终的自动识别结果图。

具体实施方式

[0050] 一种针对人体粪便白细胞的检测方法,该方法包括以下步骤:

[0051] 步骤1:对人体粪便进行稀释、搅拌、沉淀等预处理操作,使用生物显微镜采集人体

粪便样本的显微图像。

[0052] 步骤2:对步骤1中的图像进行灰度化处理,转为灰度图像。

[0053] 步骤3:对步骤2中的灰度图像进行局部反向二值化处理,获得二值化图像。处理局部的区域为方形,半径为11;阈值选取方法为灰度均值,阈值放缩系数为0.98。

[0054] 步骤4:对步骤3中的二值图像进行孔洞填充操作,然后去除面积小于200的区域。

[0055] 步骤5:对步骤4中的二值图像进行形态学膨胀运算,采用半径为4的圆形模板。对膨胀后的二值图像进行孔洞填充。对孔洞填充后的二值图像进行形态学腐蚀运算,采用半径为2的圆形模板。

[0056] 步骤6:将步骤5的二值图像连通区域标记,并计算连通区域的面积特征和外接矩形。保留面积大于1500,且占外接矩形面积比例大于70%;外接矩形宽高大于35,小于110的区域。

[0057] 步骤7:对步骤6保留的区域,根据外接矩形裁剪二值图像。对裁剪的二值图像,首先计算二值图像的最大内接圆的圆心和半径,仅保留内接圆半径大于18的区域。对二值图像进一步操作,只保留距离圆心小于24的点。

[0058] 步骤8:对步骤7的二值图像,计算二值图像的最小外接圆,保留二值区域面积占外接圆面积84%以上的二值图像。

[0059] 步骤9:对步骤8的二值图像重新计算并修正外接矩形坐标。保留外接矩形宽高大于53,宽高绝对值之差小于15的区域。

[0060] 步骤10:对步骤9保留的区域,根据新计算的外接矩形坐标,裁剪对应的灰度图和彩图。计算灰度图的方差、灰度均值和清晰度。保留方差大于300,灰度均值大于60,清晰度大于15的区域。

[0061] 以下分为训练分类器和实际检测两个过程。步骤11~步骤16为分类器训练过程。步骤17~步骤20为实际检测过程。

[0062] 步骤11:将步骤10中保留的区域坐标与对应的人工框选白细胞区域坐标对比,去除真实的白细胞,并截取剩余区域的彩图作为负样本集的杂质。训练使用的正样本集采用人工框选的白细胞。步骤4~步骤10所有采用的参数,均可保证所有样本中的白细胞进入步骤10后的被保留区域。

[0063] 步骤12:对所有的样本集,将图像缩放至80*80大小,然后分别提取三个颜色通道的Gabor特征和LBP特征。其中Gabor特征尺度选择15;方向选择 $\{0^\circ, 45^\circ, 90^\circ, 135^\circ, 180^\circ, 225^\circ, 270^\circ, 315^\circ\}$ 8个方向;波长取 $\pi/4$;空间纵横比取1.0;标准差取1.5;相位偏移为0。同时采用8倍降采样降低特征维度。共获得2400维特征。其中LBP(Local Binary Pattern)特征采用LBP等价模式(Uniform Pattern),共提取78维特征。组合Gabor特征和LBP特征,获得2478维特征向量。

[0064] 步骤13:将步骤12中样本集提取的特征进行预处理,获得1342维特征向量。

[0065] 步骤14:使用k-折交叉验证的方法,将样本集等分为5份,每次使用其中1份作为训练集,其余4份作为测试集。

[0066] 步骤15:对步骤14中的训练集,使用SVM(Support Vector Machine)支持向量机模型进行训练,获得分类模型1。根据交叉验证的结果,最终选择RBF核函数,gamma为自动选取;惩罚系数C取值为2;正负样本权重比值为2:1,其他均为默认参数。

- [0067] 步骤16:根据步骤15中获得的分类模型1,训练分类模型2。
- [0068] 步骤17:裁剪步骤10中被保留区域对应的彩色图像,提取Gabor和LBP特征并组合,按步骤13的方法进行数据预处理,获得待测的特征向量。
- [0069] 步骤18:将步骤17中的待测特征向量输入至分类器1,获得分类结果。
- [0070] 步骤19:提取步骤18中结果为白细胞的彩色图像HOG (Histogram of Oriented Gradient) 特征,按步骤13的方法进行数据预处理,并将处理后的特征向量送入分类器2中,获得最终的分类结果。
- [0071] 步骤20:将步骤19中所有检测到的白细胞在步骤1的彩色显微图像上标记出来。
- [0072] 所述步骤10中的清晰度计算采用二次模糊的方法,具体步骤为:
- [0073] 步骤10-1:计算灰度图像每个像素点的梯度值,获得原梯度图。
- [0074] 步骤10-2:对灰度图进行高斯模糊,模板大小为9*9,方差取1.5。
- [0075] 步骤10-3:对经过10-2步骤模糊后的灰度图像计算每个像素点的梯度值,获得模糊梯度图。
- [0076] 步骤10-4:计算步骤10-1得到的原梯度图和步骤10-3得到的模糊梯度图的平均绝对差值,即获得清晰度值。
- [0077] 所述步骤13中的具体步骤为:
- [0078] 步骤13-1:首先对2478维特征向量进行PCA (Principal Component Analysis) 主成分分析,仅保留99%的主要成分,获得1342维特征向量。
- [0079] 步骤13-2:对步骤13-1得到的特征进行'L2'规范化。
- [0080] 步骤13-3:对步骤13-2得到的特征进行标准化。
- [0081] 所述步骤16中的具体步骤为:
- [0082] 步骤16-1:将步骤15中模型1所有误检的负样本作为模型2的负样本集。正样本集为放缩后的人工框选白细胞。
- [0083] 步骤16-2:对样本集的灰度图提取HOG (Histogram of Oriented Gradient) 方向梯度直方图特征。Cell大小取8*8,梯度取9个方向;Block大小取2*2;步长取8。获得3600维特征向量。
- [0084] 步骤16-3:对样本集特征采用步骤13中数据预处理方式,获得2320维特征向量。
- [0085] 步骤16-4:采用与模型1相同训练方法,使用SVM模型,获得分类模型2,达到在保留白细胞的基础上进一步去除误检杂质的目的。最后选择RBF核函数,gamma为自动选取;惩罚系数C取1;正负样本权重比值为6:1,其他均为默认参数。

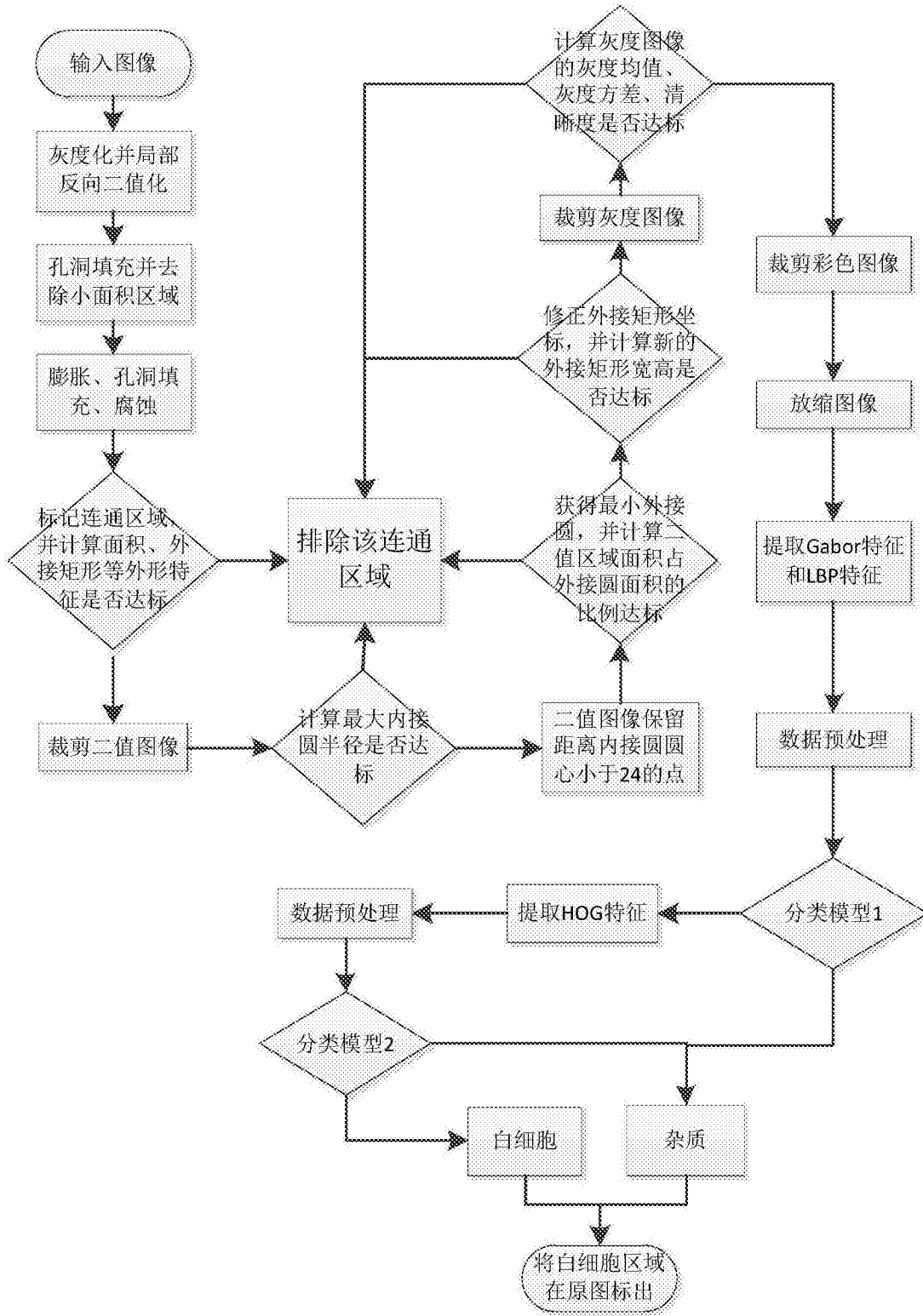


图1

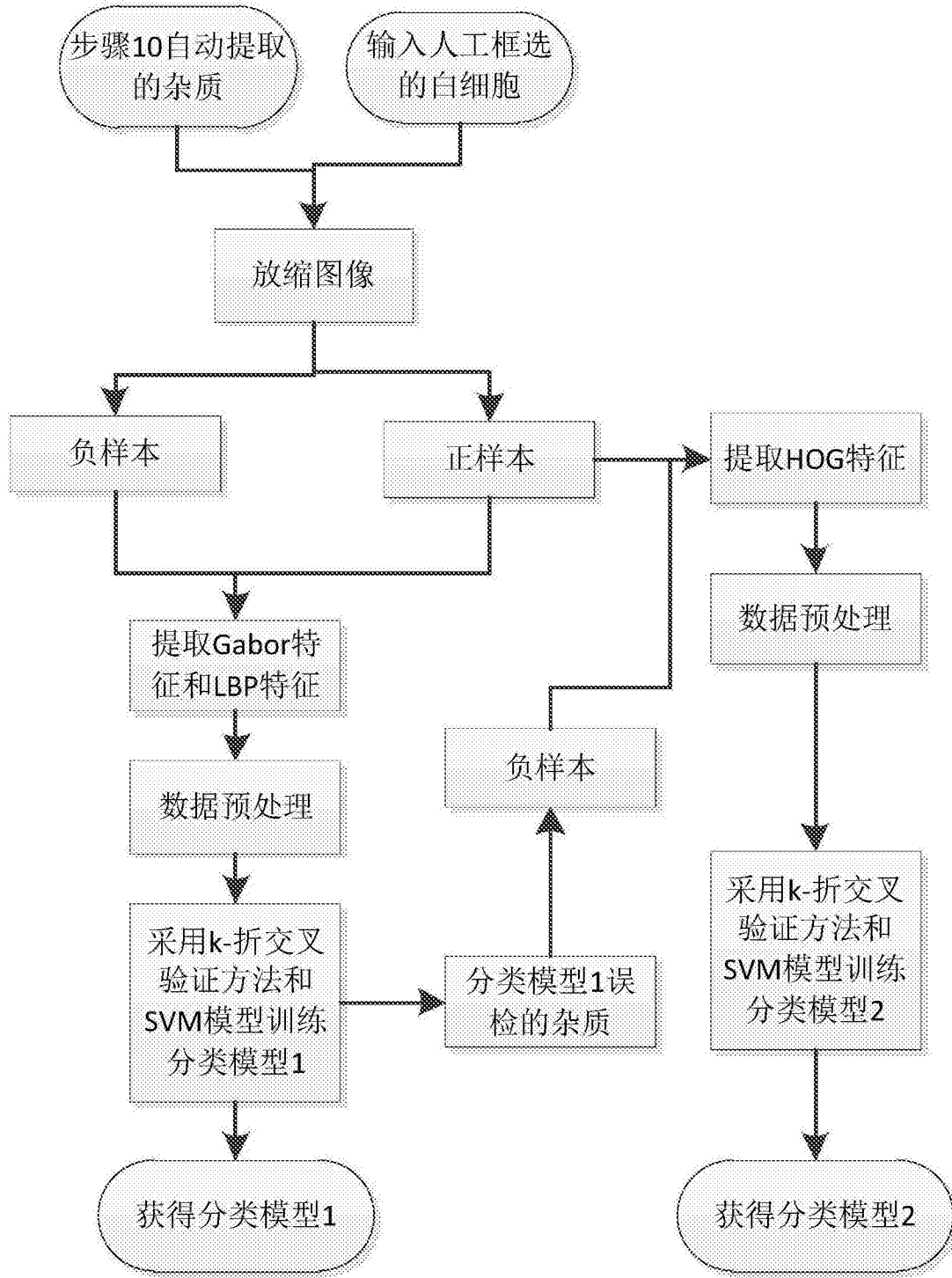


图2

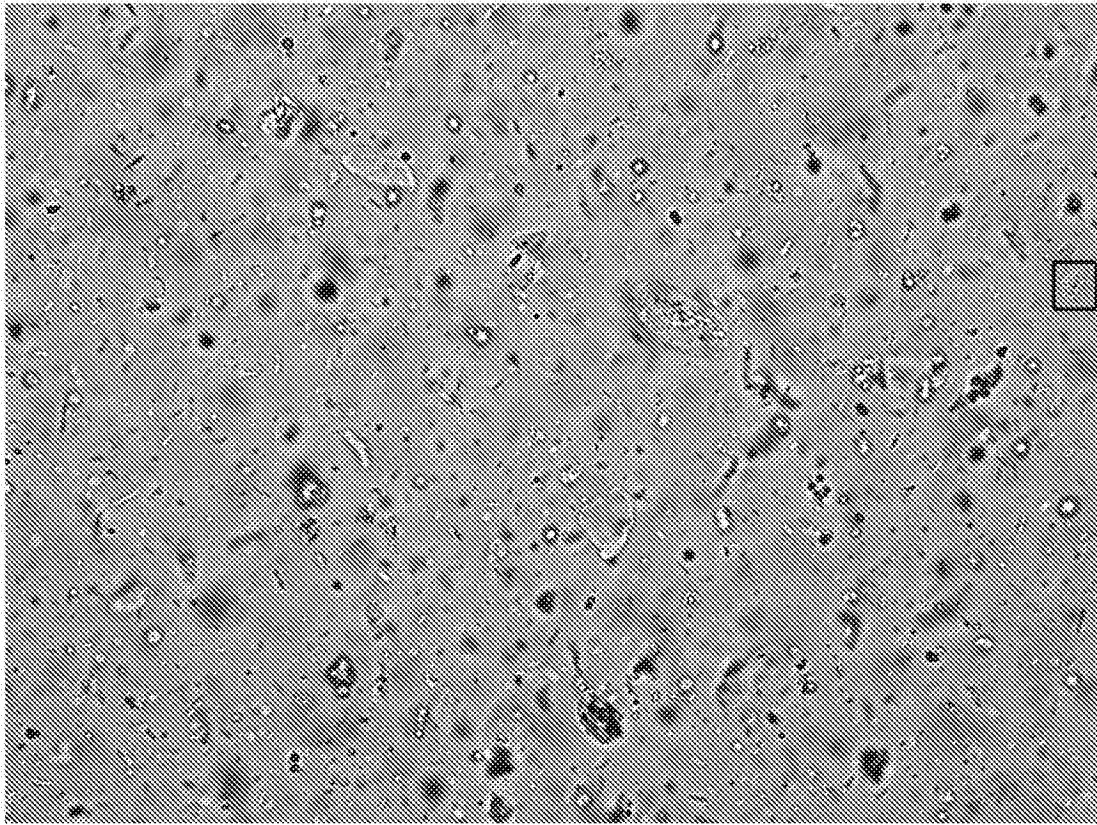


图3