基于支持向量机与多特征选择的农作物彩色病斑边缘检测①

濮永仙

(德宏师专 计科系, 德宏 678400)

摘 要: 为正确提取作物病害图像病斑, 提出了一种基于支持向量机与多特征选择的作物彩色病斑边缘检测方法. 该方法用(2d+1)×(2d+1)大小的窗口遍历图像, 计算图像亮度和色度通道的方差、均值差、最大梯度, 以及空间位置特征和均值色差作为特征向量, 实现支持向量机对病斑边缘识别. 为提高检测病斑边缘的效率, 提出了在遍历过程中, 若特征值都小于阈值时, 则跳过 d 行, d 列再遍历. 实验表明, 该方法比传统的边缘检测算子具有更好的病斑边缘识别能力.

关键词: 支持向量机; 多特征; 农作物; 彩色病斑; 边缘检测

Color Disease Spot Edge Detection of Crop Based on Multifeature Selection and Support Vector Machine

PU Yong-Xian

(Computer Science Department, Dehong Teachers' College, Dehong 678400, China)

Abstract: In order to correctly extract the lesion of crop disease images, proposes a method to detect color edge of crop disease based on support vector machines and multi feature selection. The method uses (2d + 1) * (2d + 1) as the size of the window through the image. It also is used on the image luminance and chrominance channels calculate variance, mean value difference, maximum gradient, characteristics of space position and mean color difference. In order to accurately identify the edge of disease spot, the characteristics of vector-valued input support vector machine is employed. To improve the efficiency of disease spot of edge detection, a method is proposed in the traversal process, if the eigenvalues are smaller than the threshold, skip d row, d column to traverse. The experimental shows that the method has better than the traditional edge detection operator disease spot edge recognition ability.

Key words: support vector machine; multi-feature; crops; color disease spot; edge detection

边缘检测是图像处理中的一种分割技术. 边缘是以图像局部特征不连续的形式出现, 即图像局部特征发生剧烈变化的地方, 这些特征包括灰度、颜色、纹理等. 在作物生长过程中, 有许多病害危害, 其中叶部病斑的纹理、颜色、形状等特征信息反映了作物病害类型及受害程度, 是诊断病害的依据. 因此, 准确分割病害图像病斑是提取病害特征, 准确诊断病害类别的前提. 近年来, 许多学者对作物病斑的分割进行了研究, 取得了一定成绩, 有赵进辉等[1]采用面积阈值及链码分割甘蔗赤腐病和环斑病病斑, 收到了较好的效果; 毛罕平, 张柏毅等[2,3]利用模糊 C 均值聚类对棉花、玉米病害图像的病斑进行分割, 取得了一定的成效; 任玉刚等[4]利用分水岭算法对黄瓜病害图像病斑分割, 平均正确率达 90%以上; 祁广云等[5]采用改进的遗传算法及 BP 神经网络对大豆叶片病害图像病斑分割, 能有效提取病斑区域.

本文提出基于支持向量机与多特征选择的作物彩色病斑边缘检测. 支持向量机[6](Support Vector Machine, SVM)是一款基于数据的机器学习方法, 因其出色的学习性能和泛化能力, 已成为当前机器学习

①收稿时间:2014-01-09;收到修改稿时间:2014-03-10

领域的研究热点. 有学者已将支持向量机用于图像边缘检测[7,8], 但局限于灰度图像, 且对作物病斑边缘检测的还鲜见. 彩色图像比灰度图像包含更多的信息, 已有研究表明, 彩色图像与灰度图像中大约 90%的信息是相同的, 而另外的 10%在灰度图像中是无法检测到的. 本文通过在 CIEL*a*b* 颜色空间, 用(2d+1)×(2d+1)大小的窗口遍历图像, 计算图像亮度和色度通道的方差、均值差、最大梯度, 以及空间位置特征和均值色差作为特征向量, 实现支持向量机对病斑边缘识别. 为提高速度, 在遍历过程中, 若特征值都小于阈值, 则跳过 d 行, d 列遍历, 这样提高了寻找病斑边缘的速度, 尤其在病斑较少的图像中. 这种通过边缘检测提取病斑的方式, 可以减少病害图像处理的信息量, 又能描述病斑的形态特征, 为病害特征的提取和诊断提供了前提和基础. 实验表明, 通过样本训练后的支持向量机具有很好病斑边缘识别能力.

1 支持向量机原理

支持向量机是 Vapnik 等人根据统计学理论中结构风险最小化原则提出的一种模式识别方法, 它在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势[9-11]. 基本原理是: 如有两类线性集合 $D=\{(x_1,y_1),(x_2,y_2),...,(x_i,y_i)\}$, $x\in R^d$, $y\in\{-1,1\}$ 为保证对所有样本均正确分类, 要求满足如下约束.

$$y_i[\omega\cdot x+b]-1\geq 0\,(i=1,2,\ldots,n) \quad (1)$$

式中 ω 为分类面权系数向量, 此时最大化分类间隔等价于式 1.1 约束下求式 1.2 的最小值.

$$\Phi(\omega)=\frac{1}{2}\|\omega\|^2 \quad (2)$$

解决约束最小化问题引入 Lagrange 函数, 有:

$$L(\omega,b,a)=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{n}a_i[y_i(\omega\cdot x+b)-1] \quad (3)$$

将原问题变为对偶问题, 有

$$\max\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_jy_iy_j(x_i\cdot x_j),$$

$$\text{s.t.}\sum_{i=1}^{n}a_iy_i=0,a_i\geq i,i=1,2,\ldots n \quad (4)$$

式中 a 为 Lagrange 乘子, 若 a^*为最优解, 则

$$\omega^*=\sum_{i\in s}a_iy_ix_i,\ b^*=\frac{1}{y_i}-(\omega^*\cdot x_s) \quad (5)$$

式中 s 为训练样本下标集合, 其相应的分类阈值为 b^*, x_s 为特定的支持向量. 解上述问题得最优分类函数.

$$f(x)=\text{sgn}[\sum_{i=1}^{n}a^*{}_iy_i(x_i\cdot x)+b^*] \quad (6)$$

对线性不可分问题, 只需加一个松弛变量 ξ_i, 此时约束条件为

$$y_i[\omega\cdot x_i+b]\geq 1-\xi \quad (7)$$

目标函数变为:

$$\Phi(\omega,\xi)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i \quad (8)$$

式中 C 为惩罚因子, 控制着对错分样本惩罚度的作用. 对非线性的解决是定义一个核函数 $k(x_i,x_j)$, 使其在高维空间线性可分, $k(x_i,x_j)=<\phi(x_i),\phi(x_j)>$, 相应的分类函数变为

$$f(x)=\text{sgn}[\sum_{i=1}^{n}a^*{}_iy_ik(x_i\cdot x)+b^*] \quad (9)$$

常用的核函数有线性核函数、多项式核函数、径向基核函数、Sigmoid 核函数等.

2 支持向量机检测病斑边缘的可行性

用效果较好的 canny 算子对烟草的赤星病病斑边缘检测, 其结果见图 1 所示, 在检测到病斑边缘的同时, 许多非目标边缘和杂质边缘也被检测出来. 经典算子在理想场景的效果很好, 但在复杂场景时, 很难在提取特定边缘的同时能有效抑制噪声. 因此经典算子只适用于有限的理想场合.

图 1 canny 算子检测的病斑边缘

作物病害图像由正常区域和病斑区域组成. 通常, 病斑区域的颜色呈现不同程度的黄色或褐色, 而正常区域则是不同程度的绿色. 注意到病斑边缘位于绿色与黄色或褐色之间, 边缘像素无论是与病斑内部, 还是与背景像素在某些特征上都有明显的区别. 若把边缘看作一类, 其他的看作一类, 在样本中对第 i 像素的特征取值$(x_1,x_2,\ldots,x_n)$, 对应第 i 个像素的最终输出 $y\in\{-1,1\}$, 设当像素是边缘像素时 y 值为 1, 若

像素是非边缘像素，y 值为－1. 这样可利用支持向量机实现病害图像病斑边缘检测.

3　基于支持向量机的作物病斑边缘检测

采用 SVM 对图像进行分割,首先是选择样本点生成对应的特征向量进行学习,其次计算待分割图像的每个像素点的特征向量并进行分类.

3.1 选取样本信息

1)颜色空间选择

在众多颜色模型中, CIEL*a*b*模型符合人的视觉特征, 与光线及设备无关, 处理速度与 RGB 模型同样快, 比 CMYK 模式快, 是一种均匀的彩色空间, 适合图像编辑和分析, 本文采用了 CIEL*a*b*模型. 从 RGB 空间到 L*a*b*空间的转化, 采用 D65 白点, 其中 Xn =0.950456, Yn =1, Zn =1.088754.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.607 & 0.174 & 0.200 \\ 0.299 & 0.587 & 0.114 \\ 0.00 & 0.066 & 1.116 \end{bmatrix} * \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{10}$$

$$L^* = \begin{cases} 116*(Y/Yn)^{1/3} - 16, & if(Y/Yn) > 0.008856 \\ 903.3*Y/Yn, & f(Y/Yn) \le 0.008856 \end{cases} \tag{11}$$

$$\begin{bmatrix} a^* \\ b^* \end{bmatrix} = \begin{cases} 500*((X/Xn)^{1/3} - (Y/Yn)^{1/3}) \\ 200*((Y/Yn)^{1/3} - (Z/Zn)^{1/3}) \end{cases} \tag{12}$$

2)信息提取

因传统病斑边缘检测算子会将很多非目标及杂质边缘一同检测出来, 所以仅用像素的梯度信息作为特征值是无法正确检测病斑边缘的, 这就要求在训练集合中加入其它的特征值. 根据边缘的定义, 边缘是一个“局部”概念, 边缘像素是相对于周围像素来说发生突变的地方, 其本身像素的特征并不能很好的代表边缘特征, 所以应联合邻域像素一起考虑, 加入像素邻域的特征值. 采用大小为(2d+1)×(2d+1)的窗口遍历图像, 分别计算窗口的亮度方差、色度均值差、色度方差、亮度最大梯度、色度最大梯度值、均值色差以及空间位置特征共 11 个特征向量.

①方差和均值差

方差和均值差是描述图像信息的重要特征, 若邻域内较为平滑, 则每个像素点都接近于均值, 则对应邻域的方差和均值差较小, 反之在像素值变化较为剧烈的区域, 如边缘附近, 方差和均值较大. 窗口尺寸为: (2d+1)×(2d+1), 则有以 $f(i,j)$ 为中心的邻域窗口像素的平均值 $E_{f(i,j)}$、方差 $\sigma^2{}_{f(i,j)}$ 和均值差 $CE_{f(i,j)}$ 的公式为:

$$E_{f(i,j)} = \frac{1}{(2d+1)^2} \sum_{x=i-d}^{i+d} \sum_{y=j-d}^{j+d} f(x,y) \tag{13}$$

$$\sigma^2{}_{f(i,j)} = \frac{1}{(2d+1)^2} \sum_{x=i-d}^{i+d} \sum_{y=j-d}^{j+d} (f(x,y) - E_{f(i,j)})^2 \tag{14}$$

$$CE_{f(i,j)} = \frac{1}{(2d+1)^2} \sum_{\substack{x=i-d \\ x\ne i}}^{i+d} \sum_{\substack{y=j-d \\ x\ne j}}^{j+d} f(x,y) - F_{f(i,j)}(x,y) \tag{15}$$

式(15)中, $F_{f(i,j)}(x,y)$ 为中心像素的颜色值, 该式是计算除了中心像素的均值, 再与中心像素的差, 若中心像素是边缘像素, 则均值差 $CE_{f(i,j)}$ 很大, 否则很小或为零.

②梯度特征

梯度模板选用图 2 所示的 5×5 加权梯度模板. 考虑到距离边缘像素近的像素对边缘的影响较大, 反之则小, 模板设计是一个分别对两个方向的边缘敏感, 距离中心近的像素权值大一些, 而距离中心远的权值小一些. 用窗口子图分别与 4 个模板进行卷积, 如果窗口中心像素是边缘像素, 则其中一个模板卷积绝对值会较大, 如果窗口中心像素是非边缘像素, 则所有模板卷积绝对值会很小或为零.

$$\begin{bmatrix} 1 & 2 & 3 & 2 & 1 \\ 2 & 3 & 5 & 3 & 2 \\ 0 & 0 & 0 & 0 & 0 \\ -2 & -3 & -5 & -3 & -2 \\ -1 & -2 & -3 & -2 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 & -2 & -1 \\ 2 & 3 & 0 & -3 & -2 \\ 3 & 5 & 0 & -5 & -3 \\ 2 & 3 & 0 & -3 & -2 \\ 1 & 2 & 0 & -2 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 2 & 3 & 2 & 0 \\ 2 & 3 & 5 & 0 & -2 \\ 3 & 5 & 0 & -5 & -3 \\ 2 & 0 & -5 & -3 & -2 \\ 0 & -2 & -3 & -2 & -1 \end{bmatrix} \begin{bmatrix} 0 & 2 & 3 & 2 & 1 \\ -2 & 0 & 5 & 3 & 2 \\ -3 & -5 & 0 & 5 & 3 \\ -2 & -3 & -5 & 0 & 2 \\ -1 & -2 & -3 & -2 & 0 \end{bmatrix}$$

图 2　用于检测边缘梯度的 5×5 模板

③均值色差

方差、均值差和最大梯度都是分别计算彩色图像三个通道的值作为特征值, 这样会割裂了彩色图像三个颜色通道之间的内在联系, 难免会造成各分量颜色失真, 检测出来的边缘准确率不高. 为此, 本文将色差和空间位置特征也作为特征值向量[12].

在 CIEL*a*b*空间中, 可用空间距离计算两个像素(L_1, a_1, b_1)和(L_2, a_2, b_2)之间的色差, 本文将窗口中心像素($L_{x_0y_0}$, $a_{x_0y_0}$, $b_{x_0y_0}$)到任意像素(L_{ij}, a_{ij}, b_{ij})的色差均值作为特征向量, 计算方法如式(16)所示. 若窗口内是平滑的, 即中心像素不是边缘像

素, 其均值色差就很小或为零, 反之较大.

$$LD=\frac{1}{(2d+1)^2}\sum_{i=x_0-d}^{x_0+d}\sum_{j=y_0-d}^{y_0+d}\sqrt{(L_{x_0y_0}-L_{ij})^2+(a_{x_0y_0}-a_{ij})^2+(b_{x_0y_0}-b_{ij})^2} \tag{16}$$

式中 LD 表示色差, 单位是 NBS. 色彩空间是人眼感知的均匀色彩空间, 两个颜色点之间的空间距离代表这两种颜色在人眼视觉感知上的差异. 色差值大小与人眼主观感知颜色差异程度之间的关系如表 1 所示[13]

表 1 色差与人眼颜色差异感知程度表

色差值	人眼主观感知差异
0-0.5	痕迹
0.5-1.5	轻微
1.5-3.0	可觉察
3.0-6.0	可识别
6.0-12.0	大
12.0 以上	非常大

④空间位置特征

在窗口中, 空间位置特征用以中心像素的平均行方向和列方向的像素均值差来描述.

$$x=\left|\frac{1}{2d+1}\sum_{\substack{i=i-d\\ i\neq x}}^{i+d}I_i(x)-f(x,y)\right| \tag{17}$$

$$y=\left|\frac{1}{2d+1}\sum_{\substack{j=j-d\\ j\neq y}}^{j+d}I_j(y)-f(x,y)\right| \tag{18}$$

式中 $f(x,y)$ 为中心像素的像素值, $I_i(x),I_j(y)$ 分别为行方向和列方向的像素值. 特征值取 x,y 的最大值, 若中心像素是边缘像素, 其绝对值较大, 否则绝对值很小或为零.

3)特征解释

根据以上分析, 本文选取了 11 个特征向量, 分别是三个通道的方差、均值差、最大梯度, 空间位置特征及均值色差.

特征向量解释如下:

- 亮度方差是指在亮度通道(L)中, 每个窗口子图中像素亮度的方差.
- 两个窗口色度方差是指在两个色度通道(a、b)中, 每个窗口子图中像素的色度方差.
- 亮度均值差是指在亮度通道(L)中, 计算每个窗口子图除中心像素外的亮度均值与中心像素亮度的差.
- 两个窗口色度均值差是指在两个色度通道(a、b)中, 计算每个窗口子图除中心像素色度值以外的均值, 并与中心像素色度的差.
- 亮度最大梯度值是指用图 2 所示的 4 个加权梯度模板在亮度通道(L)上分别对每个窗口子图做卷积运算, 选择梯度最大值.
- 两个色度最大梯度值是指用图 2 所示的 4 个梯度模板在两个色度通道(a、b)上对每个窗口子图做卷积运算, 选取梯度最大值.
- 空间位置特征是指分别计算窗口子图中以中心像素为据点的水平和垂直均值, 并与中心像素的差, 取其最大值.
- 均值色差是指每个窗口子图中, 中心像素与其他像素色差的均值.

3.2 支持向量机边缘识别模型的构建

利用 IBM 笔记本电脑, 配置为英特尔酷睿双核@2.4GHZ,2G 内存, 500G 硬盘, 在 Windows 2007 环境下, 用 Matlab2009a 编程实现, 其步骤如下:

1)特征值提取

a 选择几幅质量较好的图像, 用大小为(2d+1)×(2d+1)的窗口遍历图像, 经验证大小为 5×5 的窗口遍历效果较好, 计算窗口的特征向量值;

b 设阈值为 S, 若特征值大于阈值(均值色差的阈值可参看表 1, 本文选 6BNS), 则中心像素为边缘点, 支持向量机的输出值为=1,若特征值小于阈值, 则中心像素不是边缘点, 输出值为=－1;

c 为提高特征提取速度, 经手工选择参数和计算对比, 若在某窗口内计算的特征值都小于阈值 S<=0.001, 均值色差 LD<=6BNS 时, 即窗口内没有边缘像素, 则跳过 d 行、d 列再遍历图像.

2)建立 SVM 分类模型, 核函数选择和参数优化, 本文采用 5-交叉验证方法, 选择平均正确率大的核函数作为图像病斑边缘识别模型的核函数.

3)对待识别的图像进行测试.

3.3 实验与分析

本文在田间自然光照下, 用 1000 万像素佳能数码相机, 在云南德宏潞西市示范田拍摄 3 种(野火病、赤星病、蛙眼病)病害图像 300 幅, 从中选择效果较好的子图 135 幅, 修剪掉多余的部分, 使病害图像大小为 300×400, 其中每种病害 45 幅, 以 jpg 格式存储在电脑中. 核函数选择径向基核函数, 模型的训练样本每种病斑 25 幅, 测试 20 幅. 采用 5-交叉验证, 其验证结果

如表 2 所示. 表 3 为几种核函数病斑识别准确率, 其正确识别率为检测出的病斑数除以总的病斑数的商. 图 3 分别用效果较好的 canny 算子和本文方法测试的边缘检测效果图, 其中核函数选用 RBF 核函数. 限于篇幅, 只列出了每类病害识别效果图的一幅, 其他的相似.

1)从表 3 可以得出, 核函数为径向基(RBF)的病斑识别率最高平均为 97%, 核函数为 Sigmoid 的识别率较差平均为 67.3%, 这些病害中, 赤星病病斑的识别率较高, 其他的病害病斑稍差一些.

2)从图 3 可以看出, 运用本文算法可提取病害图像中的病斑边缘, 并同人眼的分割结果具有一致性.

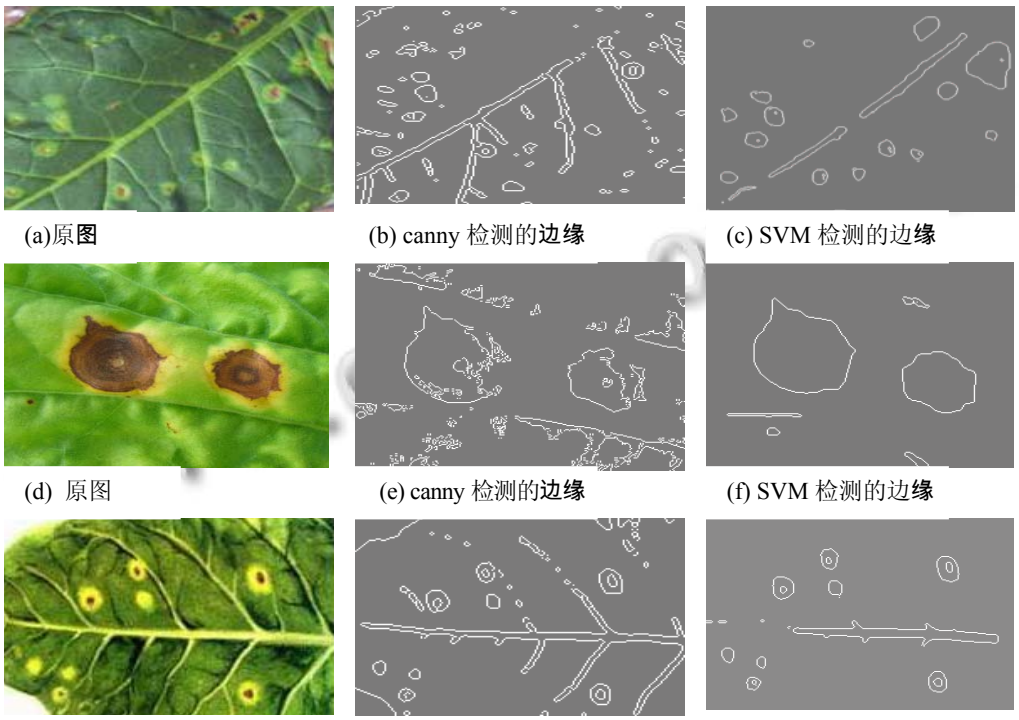
(3)本文方法与边缘检测效果较好的 Canny 算法相比, 病斑边缘都能清晰的检测出来, 较大程度地克服了噪声干扰. 图(a)和图(c)中的主叶脉边缘被检测出来, 是因叶脉区域的特征类与背景区域有明显的区别, 且与病斑区域类似. 为了区分极少数的叶脉和病斑边缘, 在统计病斑时, 只需统计近似圆形的区域, 并且半径大于一定值的二值化区域即可.

表 2　各种核函数对烟草病害图像边缘检测的最优参数

	Polynomial(缺省 coef=0)(C=0.125,g=0.125,q=1)	RBF(C=0.125,g=0.125)	Simiod(C=1,g=0.06)
交叉验证平均正确率(%)	97.412	98.282	96.753

表 3　待测病斑边缘识别准确率

方法	核函数	支持向量数	正确识别率%			
			野火病斑	赤星病斑	蛙眼病病斑	平均识别率%
支持向量机	多项式	75	86	91	92	89.7
	径向基	67	97	99	96	97.3
	Sigmoid	81	63	75	64	67.3



(a)原图　(b) canny 检测的边缘　(c) SVM 检测的边缘

(d) 原图　(e) canny 检测的边缘　(f) SVM 检测的边缘

(g) 原图　(h) canny 检测的边缘　(i) SVM 检测的边缘

图 3 算法比较

4 结语

针对传统边缘检测算子在作物病斑检测中, 噪音大、无法去除非目标边缘等, 提出了一种基于支持向量机与多特征选择的作物彩色病斑边缘检测方法. 该方法用(2d+1)×(2d+1)大小的窗口遍历图像, 计算图像亮度和色度通道的方差、均值差、最大梯度, 以及空间位置特征和均值色差作为特征向量, 实现支持向量机对病斑边缘识别. 训练后的支持向量机模型可对待测作物病斑边缘识别, 不受非目标边缘影响. 为提高速度, 提出了在遍历过程中, 计算特征值都小于阈值时, 跳过d行, d列遍历, 从而极大的提高了效率, 尤其在病斑较少的图像中. 支持向量机作为一种有监督的模式识别方法, 特征向量和参数的选定很大程度上决定着病斑边缘检测的效果, 对于它们的选定研究, 以及将本文方法移植到其它作物病斑边缘检测是下一步的主要工作.

参考文献

1 赵进辉,罗锡文,周志艳.基于颜色与形状特征的甘蔗病害图像分割方法.农业机械学报,2008,39(9).

2 毛罕平,张艳诚,胡波.基于模糊 C 均值聚类的作物病害叶片图像分割方法研究.农业工程学报,2008,24(9).

3 张柏毅,朱景福,刘勇.基于模糊 C-均值聚类的作物叶部病斑图像分割.智能计算机与应用,2011,3(10).

4 任玉刚,张建,李淼,等.基于分水岭算法的作物病害叶片图像分割方法.计算机应用,2012,32(3):752–755.

5 祁广云,马晓丹,关海鸥.采用改进的遗传算法提取大豆叶片病斑图像.农业工程学报,2009,25(5).

6 Rough Z. Sets and Intelligent Data Analysis.Information Sciences, 2002,147(1-4):1–12.

7 徐海祥,喻莉.基于支持向量机的磁共振脑组织图像分割.中国图象图形学报,2005,10(10):1275–1280.

8 李冠林,马占鸿,等.基于 K-means 硬聚类算法的葡萄病害彩色图像分割方法.农业工程学报,2010,26(2):32–36.

9 Mazzoni D, Garay MJ. An operational MISR pixel classifier using support vector machines. Remte sensing of Enviorонment, 2007,107(1):149–158.

10 Burges CJC. A totorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 1998,2 (2):121–169.

11 Gunn SR. Support vector machines for classification and regression[Technical Report]. Southampton: University of Southampton, 1998: 1–28.

l2 Canargo A, Smith JS. An image-processing based algorithm to automatically identify plant disease visual symptoms. Biosystems Engineering, 2009,102(1): 9–21.

l3 武兵.CIELAB 均匀色空间在印刷中的应用.印刷质量与标准化,2003,(5):14–17.