# 基于局部学习的玉米种子近红外高光谱图像鉴选

唐金亚 黄 敏 朱启兵

江南大学轻工业过程先进控制教育部重点实验室, 江苏 无锡 214122

**摘要** 将局部学习算法引入到种子的近红外高光谱图像最优波段选择中,并建立偏最小二乘判别分析分类预测模型,实现少波段条件下的玉米种子的快速鉴选。实验共采集了6类样本共720粒的玉米种子在874~1734 nm波段范围内的256幅近红外高光谱图像,利用局部学习算法获得波段的特征权重,并依据特征权重选择了最优波段。实验结果表明局部学习算法可有效获取最优鉴选波段,在13个最优波段条件下,对6组玉米种子可以获得平均纯度为95.97%的鉴选结果,为实现玉米种子的快速鉴选提供了一个合适的技术途径。

**关键词** 光谱学;局部学习;近红外高光谱;玉米种子;偏最小二乘判别分析;波段选择

**中图分类号** O439 **文献标识码** A

**doi:** 10.3788/LOP52.041102

# Discrimination of Maize Seeds by Near Infrared Ray Hyperspectral Imaging with Local Learning

Tang Jinya Huang Min Zhu Qibing

*Key Laboratory of Advanced Process Control for Light Industry, Minister of Education, Jiangnan University, Wuxi, Jiangsu* 214122, *China*

**Abstract** The local learning algorithm is introduced into the optimal wavelength selection of near infrared ray hyperspectral imaging of maize seeds. These obtained wavelengths are used to develop a discrimination model coupled with partial least squares discriminant analysis to implement the rapid discrimination of maize seeds using less wavelengths. 256 near infrared ray hyperspectral images between 874~1734 nm wavelengths are acquired using a hyperspectral imaging system for 720 maize seed samples including six varieties. Local learning algorithm is proposed to calculate the weight values of wavelengths, and the optimal wavelengths are selected according to the weight values. The experimental results show that local learning algorithm can effectively select the optimal wavelengths. Using 13 optimal wavelengths, six groups of maize seeds achieve an average purity of 95.97%, which can provide a suitable technical way for the rapid discrimination of maize seeds.

**Key words** spectrocsopy; local learning; near infrared hyperspectral; maize seeds; partial least squares discriminant analysis; wavelength selection

**OCIS codes** 110.2970; 200.4560; 300.2140; 300.6340

## 1 引 言

种子是农业生产最基本、最主要的生产资料,是各项农业技术和农业生产资料发挥作用的载体,对农业的增产、增收起到了关键性作用,因此种子的质量安全问题关乎整个农业生产的发展[1]。种子纯度体现了种子品种在特征特性方面典型一致的程度,是反映种子质量的重要指标。受制于多种因素的影响,种子间的混杂是一个难以避免的现象。加强种子纯度检验以减少种子混杂,是种子生产企业亟待解决的一个重要课题。

随着计算机图像技术和光谱技术的不断发展,基于机器视觉和光谱分析技术的检测技术已经得到了广

**收稿日期**: 2014-10-09; **收到修改稿日期**: 2014-11-14; **网络出版日期**: 2015-03-25

**基金项目**: 国家自然科学基金(61271384, 61275155)、江苏省自然科学基金(BK2011148)

**作者简介**: 唐金亚(1990—),女,硕士研究生,主要从事农产品无损检测方面的研究。E-mail: tangjinya2013@163.com

**导师简介**: 黄 敏(1974—),女,博士,教授,主要从事农产品无损检测方面的研究。E-mail: huangmzqb@gmail.com (通信联系人)

泛的应用[2-6]。但是在种子纯度检测中,传统的机器视觉提取的仅仅是种子外在的形态学特征信息,近红外光谱分析技术也只能获得种子的光谱信息,无法得到种子的空间信息,这两种方法获得的种子特征信息都较少,随着种子品种数目愈来愈多,种子间特征的可区分性变差,从而降低了最终的识别效果[7]。集图像和光谱技术等优点于一身的高光谱图像技术作为一种快速、无损的检测方式正被广泛应用于无损检测领域[8-11]。张初等[12]利用近红外高光谱图像技术研究了西瓜种子的分类识别问题,获得了100%的预测和建模集精度,但是该研究是针对于批量种子样本的识别。朱启兵等[13]利用400~1000 nm的高光谱图像技术研究了10类玉米种子的单粒识别问题,利用233个波段的图像熵信息并结合支持向量数据描述方法获得了94.14%的平均识别精度。由于使用了233个波段信息,导致检测速度无法满足生产企业的在线快速鉴选需要。为了减少波段数目,朱启兵等[14]利用偏最小二乘投影算法对玉米高光谱图像进行了最优波段选择,选取了65个最优波段,获得了98.90%的识别精度,但是这一波段数目仍然偏多,无法满足快速鉴选的实际需要。

高光谱数据的波段数目多,在利用全波段进行建模分析时,不仅增加数据的存储空间,而且影响检测的实时性能。解决这一问题的主要途径就是对全波段进行筛选,选出对分类最为重要的波段,从而设计多光谱系统。目前,已有多种特征选择算法被应用到高光谱波段选择中[15-16],但是这些波段选择方法多对算法参数设置敏感,且选择的波段依赖于最终分类器(回归机)模型的选择等问题。因此如何结合玉米种子的快速鉴选要求,研究一种具有稳健性的波段选择算法来选择最优波段,并开发多光谱的种子纯度鉴选方法仍是一个需要解决的课题。

本文将局部学习算法[17]引入到种子高光谱图像最优波段选择中,以达到选择最少最优波段的目的,且选择结果具有稳定性,从而实现玉米种子的快速鉴选,为开发种子纯度的多光谱图像快速鉴选方法提供途径。

## 2 材料与方法

### 2.1 实验材料

玉米种子样本由江苏省农业科学院提供的6个品种:江玉608(JY608)、苏玉30(SY30)、苏玉35(SY35)、苏玉24(SY24)、苏玉33(SY33)、豫玉18(YY18)构成,其中每类120粒,总共720粒玉米种子样本。上述种子样本部分具有亲缘关系,可较好代表制种过程中的去杂或母本去蘼不彻底这一常见的种子混杂现象。

### 2.2 高光谱成像系统和数据采集

实验采用的高光谱成像系统主要由N17E-QE线扫描成像光谱仪(Spectral Imaging Ltd. Oulu, Finland)、两个150 W的光纤卤素灯(2Specim, Spectral Imaging Ltd., Oulu, Finland)、带有聚焦棱镜的电荷耦合器件(CCD)相机、IRCP0076型电控移位平台(Isuzu Optics Corp, Taiwan, China)、计算机等部分组成。图1为高光谱成像系统图。在874~1734 nm波长范围内,波长间隔为3.36 nm时得到256个波段下的近红外高光谱图像。

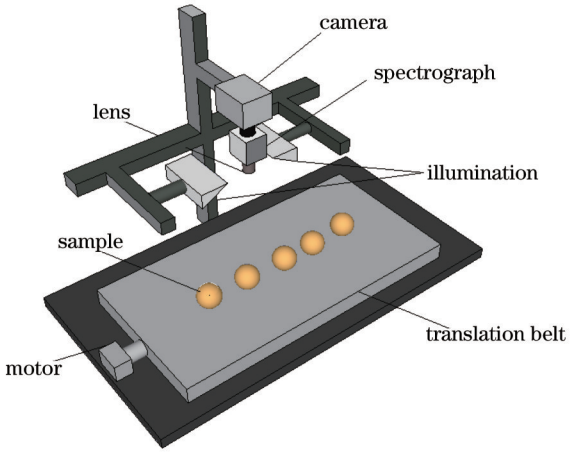


图1 近红外高光谱成像系统图

Fig.1 Hyperspectral imaging system

图像采集软件采用台湾五铃光学公司提供的高光谱成像系统采集软件,其主要控制图像采集的参数包括平台的移动速度、相机的曝光时间等。主要参数设置:平台的移动速度为13.8 mm/s, 相机的曝光时间为

3.5 ms。图像参数调整的目的是为了得到清晰、不失真、大小合适的高光谱图像，从而为更好地进行下一步实验做准备。在采集高光谱图像时，需要关闭暗箱以避免外部光线的干扰。高光谱数据处理和光谱数据分析软件分别为 ENVI 4.3 (Research System, Inc, America)和 Matlab 2009b ( MathWorks, America)。

## 3 数据分析

### 3.1 高光谱图像的校正

由于受光源在不同波段上的强度分布不均匀以及摄像头中存有的暗电流的影响，会导致光强比较弱的波段所含有的噪声过大，因此需要对采集到的高光谱图像进行校正。实验中，为了降低光源变化及系统噪声的影响，采集了10次标准白板图像和10次标准黑板图像，分别用10次平均后的白板图像和黑板图像对每次采集到的高光谱图像进行校正，校正公式为

$$R_c = \frac{R_s - R_b}{R_w - R_b}, \tag{1}$$

式中 $R_c$ 为校正图像，$R_s$ 为原始图像，$R_w$ 和 $R_b$ 为标准白板图像和黑板图像。接下来的图像处理和分析都是在校正后的 $R_c$ 上进行的。

### 3.2 高光谱图像的处理

高光谱图像处理主要包括两部分：图像预处理和特征提取。实验中首先对高光谱图像进行滤波、增强等预处理操作，然后对处理后的高光谱图像利用阈值分割的方法提取玉米种子的轮廓，获得感兴趣区域，最后在874~1734 nm波段范围内共256个波段下，提取每一粒玉米种子的感兴趣区域在这些不同波段下的光谱均值作为表征该玉米种子的特征参数。图2为SY24在1274.27 nm波段下的分割结果。

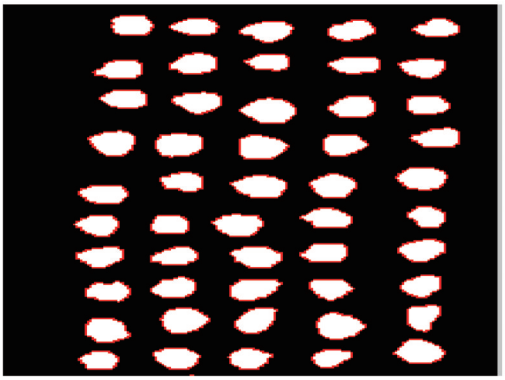

图2 SY24在1274.27 nm波段下的分割结果

Fig.2 Segmentation results of SY24 at 1274.27 nm wavelength

### 3.3 局部学习算法

局部学习算法[17]是近年来提出的一种特征自适应选择方法，该算法的核心内容主要包括两方面：1)该算法通过计算每个样本到其同类最近邻和异类最近邻之间的距离来定义最近邻间隔，从而可以实现将一个复杂的非线性问题转换成线性问题，大大简化了计算的复杂度；2)该算法引入了特征权重 $\boldsymbol{w}$，通过梯度下降的方法对权重值进行迭代调整，使得间隔内样本在权重值变换的特征空间下留一交叉验证误差最小[即解决(2)式的优化问题]，实现从全局的角度计算波段间的相关性，从而避免因对局部结构的学习导致实验结果的过拟合。

$$\min \sum_{n=1}^{N} \log[1 + \exp(-\boldsymbol{w}^{\mathrm{T}} \bar{z}_n)] + \lambda \|\boldsymbol{w}\|_1, \tag{2}$$

式中 $\boldsymbol{w}^{\mathrm{T}} \bar{z}_n$ 为在特征权重 $\boldsymbol{w}$ 下的最大样本间隔，$\lambda$ 为调整参数。

权重 $\boldsymbol{w}$ 反映了波段的重要性，根据权重的大小选择合适数量的波段，用于后续分类器的构建和多光谱系统的设计。

### 3.4 分类器设计和评价指标

采用偏最小二乘判别分析(PLSDA)算法建立玉米种子的分类预测模型，是基于偏最小二乘(PLS)方法

建立的样本分类变量与近红外高光谱特征间的回归模型。具体算法可以参考文献[11]。为了消除不同波段特征之间参量存在的差异性过大而对建模结果产生的影响，在建立PLSDA模型之前对各类玉米种子的各个波段特征参量进行归一化处理。实验最终的结果以训练/预测精度以及训练/预测纯度作为评价标准，其计算公式为

$$\begin{cases} A = \dfrac{T_P + F_P}{T_P + T_N + F_P + F_N} \\ P = \dfrac{T_P}{T_P + F_N} \end{cases}, \tag{3}$$

式中 $T_P$ 为主要类别正确识别个数，$T_N$ 为主要类别错误识别个数，$F_P$ 为杂质类别正确识别个数，$F_N$ 为杂质类别错误识别个数。

## 4 实验结果与分析

### 4.1 样本集划分

将每类120粒共6类玉米种子设计成6组实验数据，每组数据采用如下方法构成：1)利用Kennard-Stone算法[18]将JY608、SY30、SY35、SY24、SY33、YY18中的任意一类120粒种子划分成90粒和30粒，且对应类别标签为1；2)将其余5类样本的每一类利用Kennard-Stone算法划分为8粒和112粒，且所有这5类样本的类别标签记为2；3)将类别标签为1的90粒种子和类别标签为2的 $5\times8=40$ 粒种子构成训练集样本，并从剩余的 $5\times112=560$ 粒种子中随机抽取10粒和类别标签为1的30粒种子构成预测样本，最终每组实验数据得到130粒训练样本，40粒预测样本。这样设计的目的在于波段选择算法是基于监督学习的方法，传统的监督学习需要大量的标记样本才能达到理想的分类效果，然而在实际应用中要获得大量的标记样本需要花费的代价相当昂贵[19]，而且更多情况下生产厂家只需要对主要类别的种子进行纯度鉴定，对于其里面含有的杂质的类别并不关心。另外，从快速性角度上来说，对多类别融合的种子纯度鉴定，在选择波段数目和运行效率上无疑会比单一类别种子纯度鉴定来的复杂，因此6组实验数据中每一组包含一类主要种子类别，其余五类看作是掺杂的杂质，因不需要知道它们的具体类别，故总体标签记为2。

### 4.2 采用局部学习选取有效波段

根据前面样本集的划分，依次将6组数据中每一组的130×256(130个样本，256个波段)的训练样本矩阵作为局部学习算法的输入量，通过局部学习算法对初始权重 $w$ 不断更新迭代，直至 $\boldsymbol{w}$ 收敛为止，从而使得在最终得到的权重 $\boldsymbol{w}$ 变换下的特征空间内留一交叉验证的分类误差最小。因为在迭代学习过程中不断寻求最小的权重 $\boldsymbol{w}$ 直至收敛，从而导致一些不相关波段特征的权重值很小，趋近于0。最后将波段权重值由大到小排列，根据实际需要选择有效波段作为输入量，建立PLSDA分类预测模型。图3为第一组数据通过局部学习得到的权重值分布图。从图中可以看出，权重值大的波段大都分布在900~950 nm和1700~1734 nm之间。文献[20]指出在这两个波段范围内大多是C-H键一级倍频区，例如-CH3和-CH2等基团，这些和玉米种子中淀粉以及糖类等碳水化合物相吻合。图4为6类样本的平均反射光强类内均值曲线图，相比较于权重值

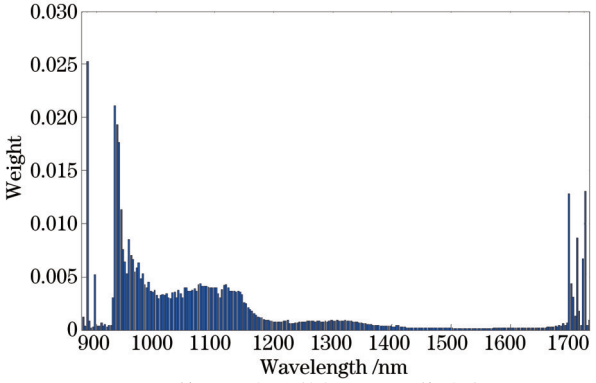


图3 第一组实验数据的权重值分布

Fig.3 Distribution of weight values for the first group experimental data

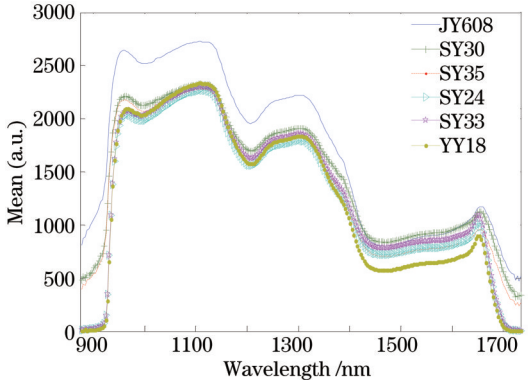


图4 6类别样本平均反射光强类内均值曲线

Fig.4 Mean curve of average reflected light intensity for six kinds of maize seeds

分布图可以看出在权重值大的部分对应的曲线差异性较大,从而可以从侧面证明波段选择的可靠性。

在局部学习过程中涉及到三个参数的调整,依次是核宽度 $\sigma$,调整参数 $\lambda$ 以及学习步长 $\eta$。前两个参数的调整可以参考文献[17],在一定的范围内,由于逻辑回归中隐含的留一交叉验证使得它们的选择对实验结果的影响很小。实验最终选定 $\sigma=2$,$\lambda=0.5$,学习步长 $\eta$ 由线性搜索得到,最终选择 $\eta=0.1$。

### 4.3 建模分析

在实验中,根据所得的波段权重值由大到小排列选择前5~15个波段依次作为输入量,建立PLSDA分类模型。为了模拟实际生产中种子混杂现象,每组样本都进行了5次随机实验(即测试样本中杂质类别样本进行了5次随机采样),最后以5次随机后的平均精度和纯度最为每组最终的实验结果。图5为每类玉米种子5次随机后的平均预测精度,图6为每类玉米种子5次随机后的平均预测纯度。由图5、6可以看出,在5~15个波段范围内不同类别玉米种子可以达到的预测精度在88%~98%之间,预测纯度在94.7%~98.7%之间。图7为6组数据在前5~15个波段下的平均训练精度和平均预测精度,图8为6组数据在前5~15个波段下的平均训练纯度和平均预测纯度。从图7、8可以看出,在第13个波段下平均训练精度和平均预测精度分别达93.26%和92%,平均训练纯度和平均预测纯度分别达到95.96%和95.97%,基本上可以满足现在的生产需求。实验结果表明:经局部学习算法选择的波段可以实现在选取少波段情况下达到较好的纯度检验结果。

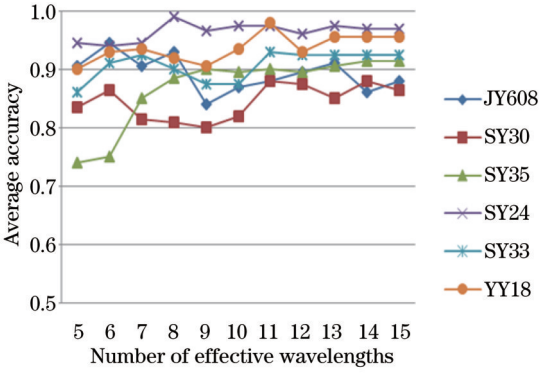


图5 每类玉米种子5次随机平均预测精度

Fig.5 Average test accuracy of five random for every kind of maize seeds

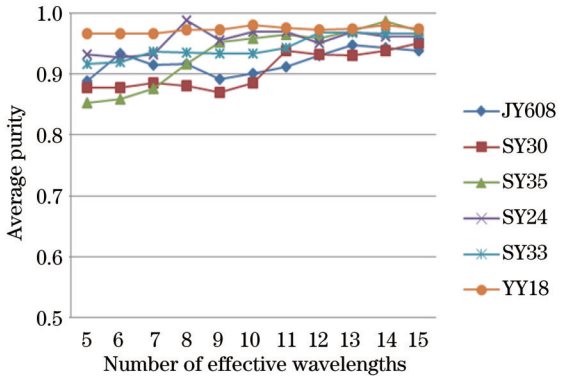


图6 每类玉米种子5次随机平均预测纯度

Fig.6 Average test purity of five random for every kind of maize seeds

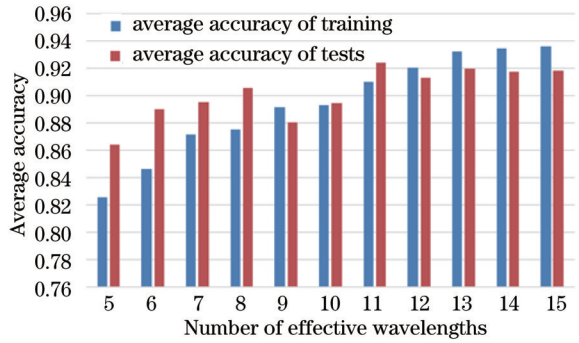


图7 6类平均训练精度和预测精度

Fig.7 Average accuracy of training and test for six kinds of maize seeds

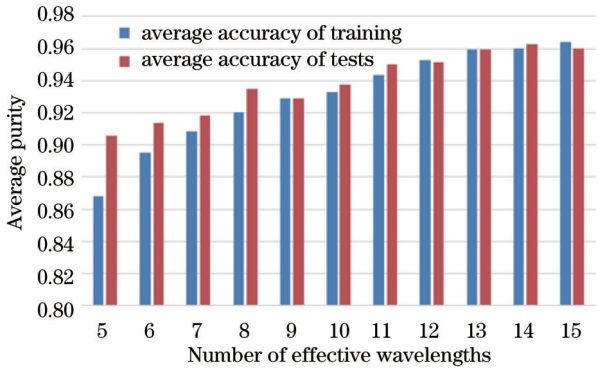


图8 6类平均训练纯度和预测纯度

Fig.8 Average purity of training and test for six kinds of maize seeds

## 5 结 论

利用波段范围在874~1734 nm下的近红外高光谱图像研究了基于局部学习的波段选择算法实现玉米种子的快速鉴选。实验根据样本的类别设计了6组实验数据,将这些数据通过局部计算算法学习权重,再根据权重大小选取有效波段,建立PLSDA分类预测模型,最后通过平均训练精度,平均预测精度,平均训练纯度以及平均预测纯度作为种子纯度检验的指标。实验结果表明,利用局部学习算法进行波段选择所选取出

来的波段特征通过PLSDA建模可以实现玉米种子的快速鉴选,而且鉴选结果不受玉米种子类别影响,具有很好的稳定性。因此局部学习算法是一种有效的波段选择算法,可以实现玉米种子的快速鉴选。

## 参 考 文 献


1 Cheng Xuefeng, Zhang Fengyun. Status and prospect of seed testing technology [J]. Seed, 2009, 28(8): 58-62.
成雪峰, 张凤云. 种子检验技术的现状与展望[J]. 种子, 2009, 28(8): 58-62.

2 Meng Qingkuan, He Jie, Qiu Ruicheng, *et al*.. Crop recognition and navigation line detection in natural environment based on machine vision [J]. Acta Optica Sinica, 2014, 34(7): 0715002.
孟庆宽, 何 洁, 仇瑞承, 等. 基于机器视觉的自然环境下作物行识别与导航线提取 [J]. 光学学报, 2014, 34(7): 0715002.

3 Liu Yande, Ying Yibin, Cheng Fang, *et al*.. Research of machine vision in purity inspection of seed [J]. Transactions of the Chinese Society for Agricultural Machinery, 2003, 34(5): 161-163.
刘燕德, 应义斌, 成 芳, 等. 机器视觉技术在种子纯度检验中的应用[J]. 农业机械学报, 2003, 34(5): 161-163.

4 Zhao Jiewen, Bi Xiakun, Lin Hao, *et al*.. Visible-near-infrafed transmission spectra for rapid analysis of the freshness of eggs [J]. Laser & Optoelectronics Progress, 2013, 50(5): 053003.
赵杰文, 毕夏坤, 林 颢, 等. 鸡蛋新鲜度的可见—近红外透射光谱快速识别[J]. 激光与光电子学进展, 2013, 50(5):053003.

5 Guo Peiyuan, Lin Yan, Fu Yan, *et al*.. Research on freshness level of meat based on near- infrared spectroscopic technique [J]. Laser & Optoelectronics Progress, 2013, 50(3): 033002.
郭培源, 林 岩, 付 妍, 等. 基于近红外光谱技术的猪肉新鲜度等级研究[J]. 激光与光电子学进展, 2013, 50(3): 033002.

6 Zhang Haidong, Li Guirong, Li Ruocheng, *et al*.. Determination of tea polyhenols content in puerh tea using near-infrared spectroscopy combined with extreme learning machine and GA-PLS algorithm [J]. Laser & OptoelectronicsPprogress, 2013, 50(4): 043001.
张海东, 李贵荣, 李若诚, 等. 近红外光谱结合极限学习机和GA-PLS算法检测普洱茶茶多酚含量[J]. 激光与光电子学进展, 2013, 50(4): 043001.

7 Yang Jinzhong, Hao Jianping, Du Tianqing, *et al*.. Discrimination of numerous maize cultivars based on seed image process [J]. Acta Agronomica Sinica, 2008, 34(6): 1069-1073.
杨锦忠, 郝建平, 杜天庆, 等. 基于种子图像处理的大数目玉米品种形态识别[J]. 作物学报, 2008, 34(6): 1069-1073.

8 Hong Tiansheng, Qiao Jun, Li Zhen, *et al*.. Non-destructive inspection of Chinese pear quality based on hyperspectral imaging technique [J]. Transactions of the CSAE, 2007, 23(2): 151-155.
洪添胜, 乔 军, 李 震, 等. 基于高光谱图像技术的雪花梨品质无损检测[J]. 农业工程学报, 2007, 23(2): 151-155.

9 Xue Long, Li Jing, Liu Muhua. Detecting pesticide residue on navel orange surface by using hyperspectral imaging [J]. Acta Optica Sinica, 2008, 28(12): 2277-2280.
薛 龙, 黎 静, 刘木华. 基于高光谱图像技术的水果表面农药残留检测试验研究[J]. 光学学报, 2008, 28(12): 2277-2280.

10 Zhao Jiewen, Hui Zhe, Huang Lin, *et al*.. Quantitative detection of TVB-N content in chicken meat with hyperspectral imaging technology [J]. Laser & Optoelectronics Progress, 2013, 50(7): 073007.
赵杰文, 惠 喆, 黄 林, 等. 高光谱成像技术检测鸡肉中挥发性盐基氮含量[J]. 激光与光电子学进展, 2013, 50(7): 073007.

11 Feng Zhaoli, Zhu Qibing, Zhu Xiao, *et al*.. Maize variety recognition using hyperspectral Image [J]. Journal of Jiangnan University (Natural Science), 2012, 11(2): 149-153.
冯朝丽, 朱启兵, 朱 晓, 等. 基于光谱特征的玉米品种高光谱图像识别[J]. 江南大学学报(自然科学版), 2012, 11(2): 149-153.

12 Zhang Chu, Liu Fei, He Yong, *et al*.. Fast identification of watermelon seed variety using near infrared hyperspectral imaging technology [J]. Transactions of the Chinese Society of Agricultural Engineering, 2013, 29(20): 270-277.
张 初, 刘 飞, 何 勇, 等. 利用近红外高光谱图像技术快速鉴别西瓜种子品种[J]. 农业工程学报, 2013, 29(20): 270-277.

13 Zhu Qibing, Feng Zhaoli, Huang Min, *et al*.. Maize seed identification using hyperspectral imaging and SVDD algorithm [J]. Spectroscopy and Spectral Analysis, 2013, 33(2): 517-521.
朱启兵, 冯朝丽, 黄 敏, 等. 基于高光谱图像技术和SVDD的玉米种子识别[J]. 光谱学与光谱分析, 2013, 33(2): 517-521.

14 Zhu Qibing, Feng Zhaoli, Huang Min, *et al*.. Maize seed classification based on image entropy using hyperspectral imaging technology [J]. Transactions of the Chinese Society of Agricultural Engineering, 2012, 28(23): 271-276.
朱启兵, 冯朝丽, 黄 敏, 等. 基于图像熵信息的玉米种子纯度高光谱图像识别[J]. 农业工程学报, 2012, 28(23): 271-276.

15 Wang Jiahua, Han Donghai. Analysis of near- infrared spectra of apple SSC by genetic algorithm optimization [J].

Spectroscopy and Spectral Analysis, 2008, 28(10): 2308-2311.
王加华, 韩东海. 基于遗传算法的苹果糖度近红外光谱分析[J]. 光谱学与光谱分析, 2008, 28(10): 2308-2311.
16 Wang Shuang, Huang Min, Zhu Qibing. Optimal wavelength selection of hyperspectral scattering images based on UVE-PLS projection analysis [J]. Acta Photonica Sinica, 2011, 40(3): 428-432.
王 爽, 黄 敏, 朱启兵. 基于无信息变量和偏最小二乘投影分析的高光谱散射图像最优波段选择[J]. 光子学报, 2011, 40(3): 428-432.
17 Sun Y, Todorovic S, Goodison S. Local- learning- based feature selection for high- dimensional data analysis [J]. Pattern Analysis and Machine Intelligence, 2010, 32(9): 1610-1626.
18 Macho S, Callao M P, Larrechi M S, *et al*.. Monitoring ethylene content in heterophasic copolymers by near-infrared spectroscopy: Standardisation of the calibration model [J]. Analytica Chimica Acta, 2001, 445(2): 213-220.
19 Cao Hui, Liu Yufeng. Not marked sample research in the application of a semi- supervised learning method [J]. Guangxi Journal of Light Industry, 2008, (12): 80-82.
曹 慧, 刘玉峰. 未标记样本在半监督学习中的应用方法研究[J]. 广西轻工业, 2008, (12): 80-82.
20 Williams P, Norris K. Near- Infrared Technology in the Agricultural and Food Industries [M]. Saint Paul: American Association of Cereal Chemists, 1987.

栏目编辑: 苏 岑