

DOI: 10.3724/SP.J.1006.2008.01069

基于种子图像处理的大数目玉米品种形态识别

杨锦忠^{1,2} 郝建平² 杜天庆² 崔福柱² 桑素平²

(¹ 青岛农业大学植物科技学院, 山东青岛 266109; ² 山西农业大学农学院, 山西太谷 030801)

摘 要: 玉米种子鉴别是种子质量检验和育种实践的重要内容。为了评价通过图像处理采集种子特征进行大数目品种鉴别的可行性, 扫描了 193 个品种各 50 粒种子图像, 建立和检验了由 4 大类 46 个种子形态特征及其组合组成的 6 种识别模型。大小类、形状类、纹理类、颜色类、后 3 类组合、全部 4 类组合等模型的品种检出率分别为 25%、33%、39%、95%、95%和 95%, 平均籽粒拒真率分别为 90%、90%、86%、45%、47%和 42%, 认伪率为 92%、92%、88%、46%、48%和 43%, 且后两个误判率高度正相关($r = 0.83^{**} \sim 0.91^{**}$)。机器视觉检测具有成本和速度上的优势, 能够用于大数目玉米品种的真伪鉴定, 形状+纹理+颜色组合模型最佳, 经改进技术识别率可以进一步提高。

关键词: 玉米; 图像处理; 品种识别; 种子形态; 判别分析

Discrimination of Numerous Maize Cultivars Based on Seed Image Process

YANG Jin-Zhong^{1,2}, HAO Jian-Ping², DU Tian-Qing², CUI Fu-Zhu², and SANG Su-Ping²

(¹ Plant Science & Technology College, Qingdao Agricultural University, Qingdao 266109, Shandong; ² Agronomy College, Shanxi Agricultural University, Taigu 030801, Shanxi, China)

Abstract: Seed identification plays a crucial role in seed quality testing and breeding programs in maize (*Zea mays* L.). Machine vision of seed surface features performances well based on a few experiments in maize. But the sample numbers in these studies were only 3–7 cultivars. To further examine the feasibility of image process application in discriminating numerous maize cultivars, six models were created and validated by means of principle component analysis and statistical discrimination analysis. The models comprised 4 categories or their combinations of 46 morphological traits extracted from scanned two-side images of 50 kernels each of 193 maize cultivars from Northeast and North China in recent years. Models of size, shape, texture, color, plus combination of latter 3 categories and combination of all 4 categories could correctly recognize cultivars at rates of 25%, 33%, 39%, 95%, 95%, and 95%, respectively, when cross-validated with all 9 650 kernels. Average refuse error rates were 90%, 90%, 86%, 45%, 47%, and 42%, respectively, and acceptance error ones were 92%, 92%, 88%, 46%, 48%, and 43%, respectively. These two error rates were highly and positively correlated between each other ($r = 0.83^{**} \sim 0.91^{**}$). Machine vision wins the advantages of low cost and high speed over manual or biochemical detecting methods, and is feasible to be applied to identification of numerous maize cultivars. The combination of shape, texture and color is the best model. Model performance may be promoted further with optimizing samples and structure.

Keywords: Maize (*Zea mays* L.); Image process; Variety identification; Seed morpha; Discrimination analysis

种子真实性和品种纯度鉴定在作物新品种区域比较试验管理、品种权保护、种子质量管理、种子生产与营销实践中有重要意义。以种子为材料的鉴定方法主要有形态鉴定、蛋白质电泳与DNA指纹等生化鉴定。生化鉴定方法需要价格昂贵的检测设备、精致复杂的实验技术、素质良好的专业人员, 测定成本较高, 其应用受到限制; 基于种子图像处理的

机器视觉检测方法是一种检测速度快、鉴别能力强、重复性高、可大批量检测、无疲劳的新方法, 而且还具有成本低、样品无损等生化鉴定所没有的优势。该方法用于种子形态鉴定, 在玉米^[1-2]、水稻^[3-6]、小麦^[7-10]和小扁豆^[11]上都获得了良好的效果。但上述报道中除Utku等^[9]的试验外, 均只有 3~7 个供试品种, 因而有必要针对大批量种子开展试验, 以便进

基金项目: 青岛农业大学高层次人才启动基金(630746); 山西省归国留学人员资助项目(2003049)

作者简介: 杨锦忠(1963–), 男, 山西介休人, 教授, 博士, 从事作物信息技术及其应用研究。Tel: 0532-88030340; E-mail: jzyang@qau.edu.cn

Received(收稿日期): 2007-09-17; Accepted(接受日期): 2007-12-16.

一步评价机器视觉方法的应用价值。

当前我国玉米育种单位数目庞大,育成品种数目剧增,如何在大数目背景下准确鉴别品种的真伪,已成为一个亟待解决的问题。本研究拟通过图像处理采集种子形态特征,对种子形态特征用于大数目品种识别的可行性,不同类型的形态特征在品种识别中的作用以及影响识别效果的因素进行分析和评价。

1 材料与方法

1.1 试验材料

193个普通黄玉米品种来自近几年来我国东北、华北地区和山西省玉米新品种区域试验,基本上代表了我国当代玉米品种的整体情况。

1.2 图像采集与特征提取

选取具有品种固有特征的50粒种子,采用600 dpi分辨率分别获取种子正反面的扫描图像。由于目前机器视觉的客观限制,本试验未采集种子顶部的图像。种子的全部形态特征采用自编的MATLAB程序实现自动提取。

1.3 种子形态特征指标

共测定46个形态特征指标,其中,反映种子大小的有长度、宽度、长轴长、短轴长、周长、等面圆直径、侧投影面积等7个指标;反映种子形状的有矩形度、椭圆度、凹凸比、圆形度、短/长轴长比、紧凑度、相对质心的纵坐标与横坐标等8个指标;反映种子纹理的为7个统计不变矩,种子纹理变量反映种子图像灰度值的次数分布特点;反映种子颜色的有24个特征指标,分别由RGB与HSV色彩空间的R、G、B、H、S和V分量的均值、方差、偏度和峰度组成。这些形态特征的定义详见文献[12]。因

为玉米种子的正反面的形态不尽相同,实际分析同时使用这两个侧面的上述全部46个指标,共 $46 \times 2 = 92$ 个变量。

1.4 识别模型的建立

为比较不同类型的种子形态特征在品种识别中的作用,分别建立了6种识别模型(表1)。一是完全模型,包括全部形态特征;二是简化模型,包括除大小类以外的全部形态特征;其他4个模型分别由大小、形状、纹理和颜色类特征组成。全部特征均符合正态分布假定($P=0.22\sim0.99$,详细数据未列出),识别模型使用线性判别函数^[13],把原始形态特征的前几个主成分作为输入变量,入选的主成分根据经验确定,入选要求为能够解释1%以上的总变异。若累积贡献率小于90%,则继续按贡献率从高到低增选,直到符合要求为止。利用相关矩阵计算特征值以消去变量量纲不同的影响。

1.5 识别模型的检验

识别模型的检验采用交叉验证的方法^[14],效果通过品种检出率和籽粒误判率进行评价。交叉验证法将 n 个样品中 $n-1$ 个作为训练集,建立模型,剩余1个作为测试集,检验模型。如此,分别轮换测试集和训练集,共做 n 次训练与检验。品种检出的标准是,归入目标品种的概率最大,且大于0.52%(品种内籽粒数目占测试总粒数的比例),且无并列最大者。限于篇幅,未列出各个品种的识别结果,代之以汇总统计。

品种检出率(%) = 正确识别的品种数目/品种总数目 $\times 100$; 籽粒拒真率(%) = 判别为非目标品种的籽粒数目/目标品种的籽粒总数目 $\times 100$; 籽粒认伪率(%) = 错判为目标品种的籽粒数目/判别为目标品种的籽粒总数目 $\times 100$ 。

表1 6个识别模型的概况
Table 1 Summaries of 6 discriminating models

模型 Model	大小 Size	形状 Shape	纹理 Texture	颜色 Color	简化 Reduced	完全 Whole
变量数目 ^{a)} Variable number ^{a)}	14	16	14	48	78	92
数据冗余度 Redundancy (%)	64.3	37.5	28.6	70.8	76.9	80.4
主成分数目 Principle component number	5	10	10	14	18	18
累积贡献率 Accumulative contribution (%)	98.3	98.3	98.6	95.1	90.4	90.3
最小贡献率 Minimum contribution (%)	1.34	0.99	1.04	0.99	0.89	0.93
Hotelling-Lawley 迹 Hotelling-Lawley trace	8.06	4.47	5.22	35.15	41.85	48.42

^{a)} 形态特征有正、反面图像的测量值,故变量数目是特征数目的2倍。

^{a)} Morpha traits were measured on both sides of kernels, so variable number doubled that of the traits.

2 结果与分析

2.1 利用种子大小特征识别品种

全部 14 个大小变量综合反映种子平面图像的大小的诸多方面，可以用相互独立的前 5 个主成分近似表示(表 1)，由此建立的大小模型对 193 个品种

进行识别。在交叉验证结果中，只有 48 个品种被正确检出，占 24.9%，效果很差(表 2)。籽粒拒真率大于 20%，平均高达 90.3%(表 3)，籽粒认伪率大于 50%，平均高达 92.3%(表 4)，两类籽粒误判率呈现明显的偏右分布。表明我国当代普通黄玉米品种的种子大小非常接近，品种之间严重重叠。

表 2 识别模型交叉验证的结果汇总

Table 2 Summaries of 6 discriminating models cross-validated

模型 Model	大小 Size	形状 Shape	纹理 Texture	颜色 Color	简化 Reduced	完全 Whole
检出率 Correct recognition rate (%)	24.9	33.2	39.4	95.3	95.3	95.3
正确识别品种数目 Number of cultivars correctly recognized	48	64	76	184	184	184
测试集品种总数目 Number of cultivars in test set	193	193	193	193	193	193

表 3 不同识别模型的籽粒拒真率次数分布

Table 3 Frequency distribution of refuse error rates for 6 discriminating models tested

籽粒拒真率 Refuse error rate (%)	大小 Size	形状 Shape	纹理 Tex- ture	颜色 Color	简化 Re- duced	完全 Whole
0-10	0	0	0	8	8	15
11-20	0	0	1	14	6	17
21-30	2	0	2	27	24	32
31-40	2	2	0	40	45	43
41-50	2	2	8	40	40	27
51-60	2	5	2	21	21	23
61-70	7	6	13	15	19	12
71-80	17	23	28	14	15	15
81-90	36	37	41	10	10	4
91-100	125	118	98	4	5	5
平均 Average	90.3	89.8	85.9	45.3	47.0	42.0

表 4 不同识别模型的籽粒认伪率次数分布表

Table 4 Frequency distributions of acceptance error rates for 6 discriminating models tested

籽粒认伪率 Acceptance error rate (%)	大小 Size	形状 Shape	纹理 Tex- ture	颜色 Color	简化 Re- duced	完全 Whole
0-10	0	0	1	5	7	8
11-20	0	0	0	6	6	12
21-30	0	0	0	26	17	24
31-40	0	0	0	43	33	45
41-50	0	0	1	39	54	42
51-60	1	2	0	37	33	30
61-70	4	1	7	14	17	12
71-80	11	13	24	13	14	14
81-90	36	45	72	6	8	3
91-100	141	132	88	4	4	3
平均 Average	92.3	91.8	87.6	46.0	47.8	42.9

2.2 利用种子形状特征识别品种

采用种子形状变量共 16 个，反映籽粒的平面轮廓特征。前 10 个主成分对形状的贡献率高达 98.3%(表 1)，它们建立的形状模型应用于测试集进行验证时，能够正确识别 64 个品种，占 33.2%，略高于种子大小模型(表 2)。这表明种子形状区分品种的能力略大于种子大小。与种子大小模型相比较，籽粒拒真率的表现基本一致，平均高达 89.8%(表 3)；籽粒认伪率的表现也相近，平均达 91.8%(表 4)。反映出试样品种在形状上相互之间的重叠程度也比较大。

2.3 利用种子纹理特征识别品种

共采用 14 个纹理特征变量，它们的综合变异可以用前 10 个主成分近似表示(表 1)。测试集的验证结果表明，由主成分构建的纹理模型能够正确检出 76 个品种，占 39.4%，略高于形状模型(表 2)，表明纹理特征比形状特征有更高的区分品种能力。籽粒拒真率均值为 85.9%，认伪率为 87.6%，误判率依然很高，且偏态分布明显(表 3 和表 4)，试样品种在纹理上重叠依然比较明显。

2.4 利用种子颜色特征识别品种

48 个颜色变量可以用前 14 个主成分近似表示(表 1)。对这些主成分组建的颜色模型进行交叉验证，结果有 184 个品种被正确检出，占 95.3%(表 2)，是前述 3 类种子特征的 2.4~4.8 倍，区分品种的能力非常理想。同时，籽粒拒真率和认伪率也大大下降，均值分别为 45.3%和 46.0%(表 3 和表 4)，约为前述 3 类特征的 1/2。这两类误判率的分布基本以均值为中心左右对称变小，表明试样品种在颜色上的重叠现

象有了明显改善。

2.5 利用全部种子形态特征识别品种

为了认识种子形态特征在识别品种上的极限,使用了完全模型,它是以种子全部形态特征的92个变量之前18个主成分建立的线性判别模型(表1)。对此模型进行交叉验证的结果表明,193个试样品种中有184个被正确检出,区分能力与颜色模型相当(表2)。虽然如此,完全模型与颜色模型相比,有更多品种的籽粒误判率出现在分布的低端,如前者小于21%的拒真率有32个品种,而后者为22个,表明品种之间重叠程度有所降低。

2.6 利用与种子大小无关的形态特征识别品种

大量玉米生产实践观察表明,籽粒大小的稳定性低于形状、纹理与颜色,因此实践上应该考虑采用无大小类变量的识别模型。本研究考察了剔除大小类变量的简化模型,它由形状、纹理与颜色类共92个变量的前18个主成分组成(表1)。模型的交叉验证表明,品种正确检出率达95.3%,区分能力与完全模型相当(表2),但是籽粒误判率出现在分布的低端一侧的品种数目略多一些(表3和表4)。

3 讨论

3.1 品种识别失败的特点

本研究表明,籽粒拒真率与籽粒认伪率之间高度正相关($r = 0.83^{**} \sim 0.91^{**}$);另一方面,完全模型的认伪品种数目 = $4.35 \times$ 籽粒拒真率($r = 0.93^{**}$),说明籽粒拒真率每增加1个百分点,则认伪品种数目约增加34个。这意味着如果一个品种有更多的籽粒被误归入其他品种,则有更多品种的籽粒会被误归入该品种。究其原因,如果一个品种有更多的籽粒与其他品种相似,则有更多其他品种的更多籽粒与该品种相似,相似性大到一定程度,就不能正确识别品种。当籽粒拒真率小于74%时大小模型能够正确检出品种,大于92%时完全不能检出,两者之间则有时能够检出,有时不能检出。形状模型、纹理模型、颜色模型、简化模型和完全模型的对应结果分别为[86%, 92%]、[82%, 92%]、[86%, 88%]、[82%, 90%]和[76%, 88%](详细数据未列出)。由此推断,74%的拒真率可以作为一个阈值,超过此值的品种识别结果就非常可疑。

品种内籽粒的一致性差是识别失败的重要原因。以完全模型为例,品种的籽粒拒真率与638个统计数(即分位数、平均数、变异数、偏度系数和峰

度系数共5大类)达极显著相关($r_{0.01} = 0.19, n=193$),其中变异数占1/2;相关系数大于0.5的共有22个统计数,全部为变异数,其中12个为变异系数,8个为标准差,2个为分位极差,涉及到宽度、面积、等面圆直径、椭圆度、颜色R与V分量等(详细数据未列出)。以颜色模型、简化模型和完全模型都无法正确检出的8个品种为例,它们与籽粒拒真率最低的8个品种相比,上述22个变异数的值高44%~85%,平均高63%。

3.2 影响品种识别效果的因素

品种检出率与模型中原始变量数目和主成分数目没有明显关系,却与Hotelling-Lawley迹的大小基本一致(表1和表2)。品种间的迹能够反映品种间综合差异的大小,迹越大则差异越大,可区分性越好,模型的识别效果就越好。前人报道的籽粒拒真率远低于本研究结果,可能与品种数目非常少有关(前人仅用3~7个品种,而本研究为193个)^[1-5,8,15]。品种数目越多,在籽粒形态诸特征上的交叠现象就越严重,可区分性变差,从而影响识别效果。籽粒多少反映样品对品种的代表性,为消除图像处理误差及籽粒间随机差异对分级的干扰,小麦样品最好为300~500粒^[16]。本研究样本容量为50粒,若增加至400粒,识别效果将会得到改善。

不同类型的籽粒特征在识别中作用不同,其识别效果次序为大小类<形状类<纹理类<颜色类特征(表2)。虽然种子大小是最稳定的一个玉米产量要素,但是受生态环境与栽培条件的影响仍然比较大,并且存在基因型×环境互作^[17],严重影响识别结果的稳定性。因此,除非有特殊要求,识别模型一般不应该包括籽粒大小特征。

作物的生长条件影响籽粒特征的表现型,进而可能影响籽粒形态的识别效果。虽然不同地点的籽粒外观形态和颜色有明显差异,但是并不影响品种识别,7个春小麦品种正确识别率达95%以上^[8],选择适当模型能够使5种麦类的识别精度达99.7%^[18]。尽管播期影响小麦籽粒的形状,但是品种间差异大于品种内差异,仍然能够正确识别^[15]。

值得指出,无论哪种识别模型,都有较高的平均籽粒拒真率,同时本研究没有包括母本对比,所以目前的模型适用于品种真伪识别,还不能用于品种纯度检验。应该建立和完善机器视觉技术,制订相关的技术标准体系,使之成为品种鉴定的有效辅助手段。通过选用遗传背景广泛的玉米品种,研究

种子形状、纹理和颜色特征与种子干燥及保存条件和年限、以及生态环境和栽培条件的关系,建立品种标准图谱库,筛选稳定性好的特征,结果将进一步完善玉米品种形态识别的机器视觉技术。兼顾遗传背景的复杂性和形态特征的稳定性,实践中应使用简化模型,并且以省级种子总站为单元,实现各省玉米品种动态识别。动态乃指逐年补充本省和国家区试合格、适合本省各生态区域种植的新品种,删除生产上已经淘汰的旧品种。同时,限制识别品种总数目,使训练数据集中各品种种子数目相等,采集果穗中部具有品种典型特征的完熟规则籽粒400粒以上,才能实现理想的识别效果。

4 结论

不同类型的种子特征在识别中作用不同,其识别效果为大小类<形状类<纹理类<颜色类特征。种子颜色在玉米品种识别中的作用非常明显,籽粒拒真率仅为大小、形状和纹理的1/2,其品种检出率也较后者高2~4倍。颜色模型、简化模型和完全模型的品种识别效果均优良,品种检出率达95%以上,通过图像处理方法实现玉米品种的种子形态识别是可行的。

References

- [1] Zha J-W(阚建文), Chen Y-Y(陈永艳). Recognition system for corn species by exterior parameters. *Trans Chin Soc Agric Machinery* (农业机械学报), 2004, 35(6): 115–118 (in Chinese with English abstract)
- [2] Yang S-Q(杨蜀秦), Ning J-F(宁纪锋), He D-J(何东健). A study on identification of maize cultivars by BP artificial neural network. *J Northwest Sci-Tech Univ Agric & For* (Nat Sci Edn) (西北农林科技大学学报·自然科学版), 2004, 32(S1): 162–164 (in Chinese with English abstract)
- [3] Wu J-H(吴继华), Liu Y-D(刘燕德), Ou-Yang A-G(欧阳爱国). Research on real time identification of seed variety by machine vision technology. *J Transluction Technol* (传感技术学报), 2005, 18(4): 742–744 (in Chinese with English abstract)
- [4] Huang X-Y(黄星奕), Li J(李剑), Jiang S(姜松). Study on identification of rice varieties using computer vision. *J Jiangsu Univ* (Nat Sci Edn) (江苏大学学报·自然科学版), 2004, 25(2): 102–104 (in Chinese with English abstract)
- [5] Yu Q-C(喻擎苍), Yan H-B(严红滨). Fuzzy pattern recognition method based on image contour line. *Trans CSAE* (农业工程学报), 2002, 18(1): 150–153 (in Chinese with English abstract)
- [6] Sakai N, Yonekawa S, Matsuzaki A. Two-dimensional image analysis of the shape of rice and its application to separating varieties. *J Food Eng*, 1996, 27: 397–407
- [7] Dubey B P, Bhagwat S G, Shouche S P, Sainis J K. Potential of artificial neural networks in varietal identification using morphometry of wheat grains. *Biosyst Eng*, 2006, 95: 61–67
- [8] He S-M(何胜美), Li Z-L(李仲来), He Z-H(何中虎). Classification of wheat cultivar by digital image analysis. *Sci Agric Sin* (中国农业科学), 2005, 38(9): 1869–1875 (in Chinese with English abstract)
- [9] Utku H, Koksel H. Use of statistical filters in the classification of wheat by image analysis. *J Food Eng*, 1998, 36: 385–394
- [10] Neuman M, Sapirstein H D, Shweddyk E, Bushuk W. Discrimination of wheat class and variety by digital image analysis of whole grain samples. *J Cereal Sci*, 1987, 6: 125–132
- [11] Venora G, Grillo O, Shahin M A, Symons S J. Identification of Sicilian landraces and Canadian cultivars of lentil using an image analysis system. *Food Res Int*, 2007, 40: 161–166
- [12] Hao J-P(郝建平), Yang J-Z(杨锦忠), Du T-Q(杜天庆), Cui F-Z(崔福柱), Sang S-P(桑素平). A study on basic morphologic information and classification of maize cultivars in China based on seed image process. *Sci Agric Sin* (中国农业科学), 2008, 41(4): 994–1002 (in Chinese with English abstract)
- [13] Zhang Y-T(张尧庭), Fang K-T(方开泰). Introduction of Multiple Variates Analysis (多元统计分析引论). Beijing: Science Press, 1982. pp 194–253 (in Chinese)
- [14] Lachenbruch P A, Mickey M A. Estimation of error rates in discriminant analysis. *Technometrics*, 1968, 10: 1–10
- [15] Sainis J K, Shouche S P, Bhagwat S G. Image analysis of wheat grains developed in different environments and its implications for identification. *J Agric Sci*, 2006, 144: 221–227
- [16] Sapirstein H D, Kohler J M. Effects of sampling and wheat grade on precision and accuracy of kernel features determined by digital image analysis. *Cereal Chem*, 1999, 76: 110–115
- [17] Lan J-H(兰进好), Li X-H(李新海), Gao S-R(高树仁), Zhang B-S(张宝石), Zhang S-H(张世煌). QTL analysis of yield components in maize under different environments. *Acta Agron Sin*(作物学报), 2005, 31(10): 1253–1259 (in Chinese with English abstract)
- [18] Majumdar S, Jayas D S. Classification of cereal grains using machine vision: IV. Combined morphology, color, and texture models. *Trans ASAE*, 2000, 43: 1689–1694