中国农业大学学报 1999,4(3):11～14
Journal of China Agricultural University

玉米种子蛋白电泳图谱图象中泳道边界提取方法研究①

廖树华② 余建华 郑丽敏 宋同明
(中国农业大学植物科技学院)

摘 要 对玉米种子蛋白电泳图谱的特点进行了分析,在此基础上进一步提出电泳图谱图象中泳道边界的分析方法——边界提取时的阈值迭代技术及其相关的图象锐化回归技术。

关键词 玉米种子;电泳图谱;图象处理;边缘提取
分类号 S338;TP391.41

Study on Lane Boundary Detection of Corn Seed Protein Electrophoregrams in Image Processing

Liao Shuhua Yu Jianhua Zheng Liming Song Tongming
(College of Plant Science & Technology,CAU)

Abstract The image features of corn seed protein electrophoregram was studied and a method, iterative computing thrade-off value in boundary detecting and regression analysis in image sharpening, of its lane boundary detection was proposed, which resulted in resolution of the problem of the lane boundary detection. In this way, the band features of electrophoregram will be extracted more effectively by computer.

Key words corn seed; electrophoregram; image processing; boundary detection

玉米是我国第二大粮食作物,但每年由于大量的伪劣玉米种子充斥市场,给国家与农户造成了不可弥补的损失。因此,品种的鉴定,包括真实性鉴别和纯度测定,对于种子贸易、玉米产量提高及其品种的保护有着十分重要的意义。实践中,目前已有经济有效、准确可靠的检测方法,这就是乳酸聚丙烯先胺凝胶电泳实验系统。该实验系统操作简单方便、重复性好、图谱清晰易辨。玉米种子在该检测系统里产生众多的谱带条纹。谱带条纹数及谱带条纹之间位置结构保持一定的不变性,俗称这些谱带条纹为玉米品种的"指纹"。实践中,正是利用品种的这一"指纹"特征(条纹数、结构上保持一定的不变性)来对玉米品种进行鉴定。但在实际工作中也常常遇到这样一些困难:①谱带条纹标准的把握(不同的人可能数出不同数目的条纹)与谱带条纹结构特征的处理;②图谱特征的量化和标准化,即遗传信息的量化和标准化;③谱带条纹"身份"的鉴别。因此,需要考虑用计算机进行分析处理。

电泳谱带条纹的计算机处理实质上是一图像处理和模式识别问题。在图像处理方面,国外80年代就有人开始研究用计算机去处理蛋白电泳图谱图象的分析问题[1]。目前,国外已有专门的基于计算机技术的电泳谱带条纹分析系统。这些系统在用于提取电泳图谱的基本特征:谱带条纹的中心位置、边界、灰度级、体积等方面有一定适用性,但没有鉴别功能,且各家的分析效果也并不一致。国内这方面的研究还很薄弱,特别是要研究鉴定问题,这只能结合具体的生物种类进行研究。本研究是在玉米品种上进行初步的探索。

收稿日期:1997-11-14
①国家自然科学基金资助项目 39470470
②廖树华,北京圆明园西路2号中国农业大学(西校区),100094

1 电泳图谱图象处理面临的问题

电泳实验产生的谱带条纹数量及分布基本决定于品种的基因型，但实验时间、环境温度、电压及凝胶酸碱度等诸多因素对图谱位置、明显程度等有影响。为尽可能消除环境因素的影响以便于图谱结构分析，本电泳实验系统引入了标准品种作参照系，为图谱的计算机处理提供了方便。

从图 1 可以看出，一张电泳图谱图象中通常有多条泳道、许多谱带条纹，各谱带条纹的大小、清晰程度不一。电泳图谱图象表面粗糙杂点也多，颜色深浅变化呈锯齿式的波动。因环境因素影响也使图象质量多变。另外，由于凝胶本身的物理特点（水分多、透明、有厚度），经拍照冲洗或扫描仪输入等二次处理后的图象质量也会发生一些变化，某些较弱但又清晰的图象会变得不清。另外，实验用的凝胶这一介质比较柔软，很易扭曲变形、破损，电泳图谱图象在实验过程中也会发生一定的偏移和弯曲，使不同区域，图象明暗度不一样。这些都给图象的处理带来很大困难，特别是图象处理时某些参数（所选邻域大小、灰度阈值、灰度变化阈值等）的选取上很难指定，本研究的阈值的迭代技术就是因此而产生。

2 泳道边界提取方法描述

在图谱图象处理中，有 2 种方式完成谱带条纹、泳道提取及关联：一是分别提取谱带条纹和泳道，然后实现关联；第二种是先提取泳道，然后提取谱带条纹。与前者相比，后者更为简便。因为，泳道提取显然要比谱带条纹提取容易多，泳道数量少、内外区域分明，而谱带条纹却没有此特点；其次，一旦图象中的泳道分割出后，很容易将谱带条纹分析的二维的图象处理问题转化为一维的谱线处理问题，可大大地简化图谱特征提取问题的分析难度。

根据上述分析，在图象的处理中，将其简化为以下几个过程：整幅图象的预处理（旋转、平滑等）；图象中泳道边界提取；泳道上的谱带条纹特征提取；利用所提取的特征进行进一步的品种鉴定与纯度测定。本研究论述的只是其中的一个过程——泳道边界提取。

图象边缘提取和分割是图象处理中的难点之一。它的解决对我们进行更高层次的处理如特征描述、识别有着重大的意义，因而也就产生许多的方法如：梯度算子、Kirsh 算子、Laplacian 算子等微分算子类方法以及 Hueckel 方法、Haralick 斜面模型、Prewitt 多项式等基于曲面拟合的边缘检测方法。本研究在参考这些方法的基础上根据玉米蛋白质电泳图谱图象的特点提出自已的处理方法。

本算法分为几个过程：①图象数据量的压缩；②图象的锐化预处理；③边界的提取。

压缩算法的过程较简单，主要是为了减少数据的处理量。一幅正常的扫描图象通常有上百万的数据量，需要进行适当压缩。

2.1 图象锐化预处理

边缘的锐化一般通过求其曲面梯度来实现，经锐化后该点的新值往往取的是其梯度的模，常用差分来近似其梯度。由于图象较粗糙，本研究提出另一种锐化方法——回归处理技术：

A）在某一象素点 $P(i,j)$ 的邻域内沿 2 相互垂直方向（通常为纵、横取向）取 2 组个数相同的相邻点，其灰度值记为 $\{x_1,x_2,\cdots,x_N\}$ $\{y_1,y_2,\cdots,y_N\}$。

B）分别在这 2 类点上作直线回归（自变量值为 $1,2,3,\cdots,N$；因变量为所取的灰度值），求出它们的回归系数，分别记为 α_x,α_y。

C)将这 2 个回归系数当作二维向量，向量的模作为象素点 $P(i,j)$ 锐化后的新值。其模通常取为：

i) $\max(|\alpha_x|,|\alpha_y|)$；　ii) $|\alpha_x|+|\alpha_y|$；　iii) $(|\alpha_x|^m,|\alpha_y|^m)^{1/m}$, $m=1,2,\cdots$

上述过程只是原理性的描述，在程序的编制中还需进一步对算法进行简化。

我们知道，一维线性回归的基本公式为：$b_1=L_{1y}/L_{11}$，其中 $L_{1y}=\sum_{i=1}^{N}x_iy_i-(\sum_{i=1}^{N}x_i\cdot\sum_{i=1}^{N}y_i)/N$，$L_{11}=\sum_{i=1}^{N}x_i^2-(\sum_{i=1}^{N}x_i)^2/N$。由于每一象素上邻域点取法相同，即 L_{11} 的值都一样，可不必计算。根据第 A)步所述，上述式子中 x_i 取值为 i，经简化后，所需的只是计算公式

$$L_{1y}=\sum_{i=1}^{N}iy_i-(N+1)\cdot\sum_{i=1}^{N}y_i/2=\sum_{i=1}^{N}(i-(N+1)/2)y_i \tag{1}$$

的值。(1)是 α_x,α_y 的计算公式，若 $N=5$，则(1)式简化为：$L_{1y}=2(y_5-y_1)+(y_4-y_2)$。公式(1)的特点之一是当 N 为奇数时结果仍为整数。在图象处理中象素的灰度级常用整数表示。为防止数值出现上溢，N 不宜过大，且应对结果进行判断。我们在 LX-P5/PCI133 机上取 $N=5$，对 800×600 象素图象进行锐化处理，1 s 之内即完成从处理到效果显示的操作。

2.2　边界的提取

边界提取实际上就是对象素进行分类，某类点称为边界点，其他为非边界点。锐化的目的就是将边缘点突出，而其他点消隐。图象经锐化后的象素值用 $P(x,y)$ 表示，称 $P(x,y)$ 为象素点 (x,y) 的锐化灰度。$P(x,y)$ 值越大，点 (x,y) 越有可能为边界点。利用这一特性，我们就可将突出的边缘区域分割出，其分离的思想就是选取某个阈值作为判断条件的依据，进行边界点的分离。实际上，阈值的选择是各种图象处理都要解决的问题。由于图象本身的特点，要求在不同的泳道片段采用不同的阈值，而在泳道边界提取之前这些片段是是未知的。本研究采用迭代思想来实现阈值的确定问题，其方法描述如下：

①STEP0：确定泳道方向　按锐化灰度级作直方图。泳道的起点是边界变化最剧烈的点，各条泳道的起始边界变化特征相近。根据起始点边缘占据的图片面积较小(通常不到总面积的 1/10)这一先验知识确定锐化灰度阈值。利用此阈值容易找出起始边缘，并完成图象整体校正(因为实验时各泳道起始点基本保同一水平直线上)，不妨认为校正后泳道朝向为 Y 轴方向。

②STEP1：图象片断划分　根据所定确的方向，将图像划分不同片断，为简化，图象中每条 Y 轴方向的垂直光栅线就是一片断，若图象为大小为 $M\times N$ 象素点，则有 N 个片断。

③STEP2：确定泳道数的目标值。

④STEP2A：初始阈值的确定　根据图象的锐化灰度直方图和图片中泳道边界面积的先验估计(通常不到总面积的 1/3)，按灰度级从大到小对直方统计值进行累加，当累加结果达到整幅图象总点数的 1/3 或更小的值时，该锐化灰度级 α 便为初始阈值。

⑤STEP2B：估计泳道目标值　利用 α 值找出各片断可能的边缘区域及其数目，并对区域数作直方图。分析直方图，若中部有一条或几条相邻的明显峰线，取这几条峰线中的中间一条所对应的边缘区域数为目标值，记作 G，并进入 STEP3。否则，增加阈值水平 α，重复本步过程。

⑥STEP3：校正各片断的阈值　取各片断的初始阈值为 STEP2B 所产的阈值，即 $\alpha_i^0=\alpha$，$i=1,2,\cdots,N$。阈值的校正原则为：片断的区域数高于目标值 G，阈值加大；低于 G，阈值减小。与目标值偏差越大，校正幅度应适当增加。

⑦STEP4：将边缘区域数与目标值 G 相近的片断进行统计，记作 N_G，若 $N_G>2N/3$ 或 N_G 的值不增加，则进入 STEP5。否则，返回 STEP3。

⑧STEP5：边界点聚类　标出各片断边缘区域的中心点，先对边缘区域数为 G 的片断上的标记点根据顺序关系及空间位置进行聚类，以此为基础，根据边缘区域数与 G 的偏差程度逐步将其他片断的标记点归并到相应的边界类中，其基本依据是空间关系。

⑨STEP6：利用边界分类结果及起始点确定各泳道的区域。

3 结果与讨论

我们利用本文所述方法对许多张的电泳图谱进行了计算机分析，效果良好。图 1 是玉米品种 M017×B73 的原图，图 2 是锐化处理后的图象，图 3 是计算机标记出的边界点。一幅大小为 800×600 象素点图像的处理时间在 P5/133 机上只需 3～4 s。

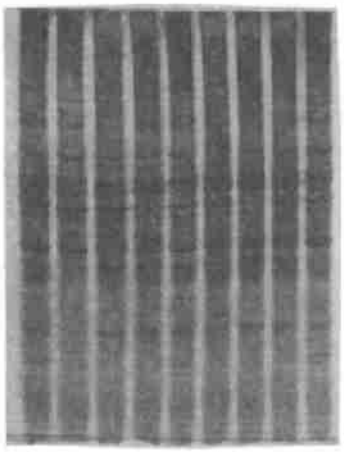

图 1　1Mo17XB73 玉米品种电泳图

图 2　经锐化处理后的图

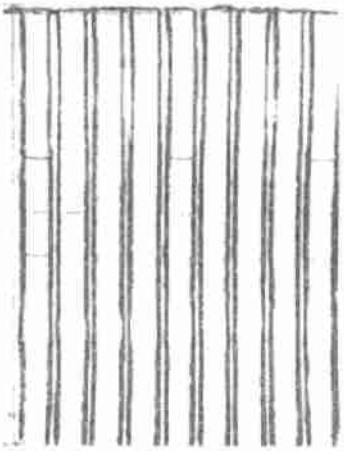

图 3　泳道边界的散点图

图象边缘提取和分割是图象处理中的重要内容。它是我们进行更高层次图象处理如特征描述、识别等内容的基础。本研究只是根据所要处理的图象类型特点，研究相应的边界提取方法。实际上，每类图象处理问题都可根据自身特点，找出更有效的图象处理方法。

以上内容只是本项目研究工作中的一部分，根据所述方法，我们完成了图谱图象泳道边界的提取。下步所要进行的是泳道内谱带条纹特征的提取与分析，并进一步利用这些特征进行计算机的玉米品种“指纹”的鉴定。限于篇幅，将不再详述。

参 考 文 献

1 Jerry E S, Michael G H. A robust high-sensitivity algorithm for automated detection of proteins in two dimensional electrophoress gels. Computer Applications in the Biosciences, 1993, 9(2): 68～74

2 钟玉琢，乔秉新，李树青. 机器人视觉技术. 北京：国防工业出版社，1994

3 田捷，沙飞，张新生. 实用图象分析与处理技术. 北京：电子工业出版社，1995

4 程兴新，曹敏. 统计计算方法. 北京：北京大学出版社，1989

5 余建华，廖树华，宋同明等. 玉米杂交种、自交系鉴定技术研究进展(综述). 中国农业大学学报，1998，3(1)：68～74