基于图像处理与机器学习的初烤含青烟叶辨别及含青程度识别研究*

李 峥[1]，徐志强[1**]，张晓兵[1]，林珈夷[2]，钟永健[1]，
徐均华[1]，张赵鹏[1]，焦得平[1]
（1. 浙江中烟工业有限责任公司 技术中心，浙江 杭州 310024；2. 上海创和亿电子科技发展有限公司，上海 200082）

［摘 要］为提升烟叶含青程度识别效率和准确度，通过对不同含青程度烟叶参比样的构建，采用计算机图像处理技术筛选出含青区域提取的最优颜色通道和参数阈值，并基于支持向量机（SVM）构建不同含青程度烟叶的识别模型. 结果表明：烟叶正面透光图像适合用于含青区域的提取；Lab 颜色模型的 a 通道提取图像中含青区域的效果最佳，其阈值范围为（141，142）；SVM 模型对不同程度含青烟叶样品的测试准确率达到 86.01%. 混淆矩阵显示，SVM 模型对正组和微带青烟叶的识别准确率较高. 图像处理技术和支持向量机对含青烟叶程度的划分有较好的应用效果.

［关键词］初烤烟叶；含青程度；支持向量机；图像处理；识别模型
［中图分类号］S572 ［文献标志码］A ［文章编号］1674－5639（2023）06－0019－07
DOI：10.14091/j.cnki.kmxyxb.2023.06.003

由于田间农艺管理措施不当，烟叶采收成熟度不足，或烘烤调制工艺与烟叶素质不匹配等多方面原因，将导致初烤烟叶出现含青的现象[1,2]. 烤烟国标将含青烟叶归为副组烟，外观质量表现较差，初烤烟叶按含青程度又分为微带青和青黄两种类型. 含青烟由于营养物质转化不充分，内在化学成分不协调，在烟叶评吸时表现出香气质差、香气量少、青杂气重、刺激性强等特点[3-5]. 含青烟叶由于吸食质量差的特性很难配伍到叶组配方中，工业企业为保证卷烟产品感官评吸质量的稳定，在原料分选环节会严格把控含青烟叶混入正组烟叶的比例.

随着交叉学科与检测仪器的不断发展与融合，近年来计算机图像处理技术在农业领域得到广泛应用[6-10]. 该项技术也逐步拓展到了烟叶生产中，例如：李增盛等[11]基于图像处理技术提取烘烤过程中烟叶图像的颜色特征和纹理特征，构建了烟叶烘烤阶段判别模型；潘治利等[12]利用图像处理技术提取不同产区烟叶图像特征，并进行归类分析；史龙飞等[13]基于机器视觉技术提取不同成熟度烟叶图像的颜色和纹理特征，实现烟叶成熟度的区分检测. 而目前关于初烤含青烟的研究大多只关注于含青烟叶的成因分析及对应措施，缺少应用计算机图像处理技术辨别初烤含青烟叶，并识别含青程度的相关研究. 而技术手段方面，支持向量机（SVM）因其解决非线性、高维数等问题的特有优势[12,14,15]，已被应用于图像处理识别领域. 有大量研究[10,16,17]将图像处理技术与 SVM 模型结合进行图像的识别和分类，目前已形成较为成熟的配套使用体系. 将图像处理和机器学习技术综合应用于初烤烟叶图像辅助识别，相对于传统的人工作业，将大幅度提高选叶效率和准确率.

本研究首先通过含青程度参比样的构建和图片采集，筛选出提取烟叶图像含青区域的最优颜色通道

* ［收稿日期］2023－06－30
［作者简介］李峥，男，浙江杭州人，浙江中烟工业有限责任公司助理农艺师，硕士，研究方向为烟叶外观质量数字化评价.
** ［通信作者］徐志强，男，山东莒县人，浙江中烟工业有限责任公司烟叶分级高级技师，研究方向为烟叶质量评价与工业应用，E-mail：xuzq@zjtobacco.com.
［基金项目］浙江中烟工业有限责任公司重点科技项目（ZJZY2021B001）.

和颜色参数阈值范围，之后基于 SVM 构建烤烟含青程度识别模型，并进行训练和测试，以解决不同类型含青烟的识别问题，并提高识别的准确率和效率，实现不同含青程度烟叶的快速准确检测，为智能化烟叶分级体系构建提供部分研究数据支撑.

1 材料与方法

1.1 供试材料

供试烤烟品种为云烟 87，含青烟叶样品采集于云南、湖南、湖北、贵州、河南、广西 6 个产烟省份，样品产地具体涉及 18 个地市的 41 个县（区）. 共收集 2021 年初烤烟叶样品 83 份，每份 1.5 kg. 供试材料按用途分为两类，一类为含青程度参比样，用于烟叶图像处理方式选择、含青区域颜色检测通道确定、颜色通道检测阈值筛选；另一类为模型训练及验证样品，用于搭建模型的学习数据库，并验证模型对含青烟叶类别识别的准确性.

1.2 试验方法

1.2.1 含青程度参比样制备

为保证参比样含青程度梯度界限清晰，组织 7 名省级及以上烟叶分级技术能手，成立含青程度参比样制作小组. 依据烟叶支脉含青条数及叶面含青程度对各地市的样品进行筛选. 分选出含青程度依次递增的烟叶 17 片构成参比样，经 Friedman 检验和定向成对比较检验，相邻梯度样品间差异显著，表明含青程度档次划分合理.

1.2.2 模型训练及验证样品

依据 GB 2635—92 烤烟分级标准，由上述评价小组对各地市的烟叶样品按正组烟叶、微带青、青黄 3 个档次进行分选，筛选标准如表 1 所示. 剔除不可用烟叶后，最终筛选出无病斑、虫蛀、叶面完整的正组、微带青、青黄烟叶各 405 片.

表 1 正组及不同含青程度烟叶筛选标准

组别	判定标准
正组烟叶	主、支脉及叶面无任何可见的青色，叶面颜色呈现基本色
微带青烟叶	叶脉带青，或叶面含微浮青面积在 10% 以内
青黄烟叶	主、支脉及叶面有明显青色，且青色占比在 30% 以内

1.2.3 烟叶图像采集

烟叶样品的图像采集处理作业在浙江中烟技术中心烟叶评级实验室进行，图像采集装置为定制设备，由上海创和亿公司生产. 装置主体由暗箱、拍摄装置、温湿度平衡装置组成. 将单片烟叶放置于暗箱中，打开光源，由拍摄装置进行烟叶图像采集，相机拍摄角度与工作台面垂直，以获取完整的图像信息. 拍摄环境温度 23 ℃、相对湿度 75%.

1.2.4 烟叶图像处理

为避免烟叶图像中无关区域的干扰，需对图像中除烟叶之外的区域进行剔除处理. 背景与阴影等无关区域的剔除分析方法为：

1）寻找烟叶本身与无关区域的差异，比较烟叶图像在 RGB 颜色空间、HSV 颜色空间中各颜色通道的差异性. 由于烟叶自身颜色与背景颜色存在明显差异，分析发现 HSV 颜色空间中的 S 通道能够轻易区分烟叶本身和无关区域. 在 S 通道背景区域像素值为 0，因此，提取烟叶本身的像素值范围为 S 通道的 1 ~ 255 之间. 通过设定提取烟叶本身的 S 阈值范围，获得烟叶的二值化图像.

2）计算二值化图像的所有连通区域，去除烟叶本身区域之外的所有小面积区域，获得烟叶本身的二值化掩码.

3）将提取烟叶的二值化掩码转换为 RGB 三通道图像，其中 RGB 每个通道的像素值均为烟叶掩码二值化单通道像素值.

4）将烟叶原图与烟叶三通道掩码进行矩阵乘法运算，获得剔除无关区域的烟叶本身图像.

1.2.5　主脉图像提取

由于含青程度不仅与含青面积相关，也与含青位置相关，因此需要判定含青像素点是位于叶脉还是叶面．为准确获取主脉的像素点位置坐标，进行了烟叶主脉图像提取，具体方法为：

1）对背面透光图像进行 gamma 校正，实现对图像亮暗程度的调节．gamma 校正方法如下式：

$$P_f = 255 \times (P_b/255)^{(1/G)}.$$

其中 P_f 为校正前的像素值，P_b 为校正后的像素值，G 为 gamma 校正因子，取值范围一般为（0.01，7.99）．经筛选本方法 G 取 0.2 为最优值．

2）提取校正后背面透光图像的 HSV 颜色通道的 V 通道．

3）计算 V 通道图像的最外层轮廓，并将 V 通道轮廓外的像素值赋值为 255．

4）将 V 通道像素值进行聚类成 3 个簇，依次为叶脉、叶面、背景，像素值从低到高．根据聚类中心的像素值，提取叶脉．

5）计算上一步获取叶脉区域的所有外部轮廓，保留最大的轮廓所在区域，将其作为最终的叶脉．由于叶面存在皱缩，尤其是上部烟叶，存在主脉被叶面部分遮挡的情况，此时保留最大连通区域后会丢失部分主脉区域．可通过提取烟叶骨架，判断所有连通区域是否与骨架区域存在重叠，若有重叠，则该连通区域也为主脉．

1.2.6　烟叶含青区域检测颜色通道选择

由于含青烟叶样本数量较大，为避免数据冗余，出现含青区域识别模糊、含青程度界限不明显等情况．以制备的含青程度参比样为分析对象，通过分析其在 RGB、HSV、Lab、LUV 这 4 个不同颜色空间的各个颜色参数单通道图像中青色位置的像素值情况，最终确定 Lab 颜色空间的 a 通道和 LUV 颜色空间的 U 通道，在烟叶含青区域与非含青区域存在明显差异，因此选取 a、U 两个通道进行含青区域的表征．

1.2.7　烟叶含青程度识别模型的构建

基于支持向量机（SVM）构建烟叶样品含青程度识别模型．SVM 主要用于分类和回归分析，是一种监督式学习的方法．核函数选用 rbf 径向基函数，拟合模型前启用概率估计，不施加惩罚量，不限制迭代次数，按照误差值确认停止模型拟合，所以 max_iter 设置为 −1．惩罚因子 c 表征对误差的宽容程度，其范围从 0.1 ~ 500 进行网格搜索（GridSearchCV），核函数中的参数 g 表征了模型的复杂程度，其最优值选取从 0.001 ~ 1 范围内进行网格搜索．最终选取的最优参数 c 为 100，参数 g 为 0.001．

1.3　软件工具

建模、测试及分析是在 Windows Server 2010 操作系统下，使用软件 Python 3.9 作为编程语言实现，并采用 matplotlib 3.5.2 库实现数据可视化．Numpy 1.19 和 Pandas 1.4 库用于数据整理和分析．通过 Scikit-learn 库中的 Pipeline 将数据标准化和模型训练串联，避免数据泄露．

2　结果与分析

2.1　烟叶图像背景剔除

由于烟叶样品含青部分与其他部分的颜色表征参数存在差异，采用滤波处理会过滤一些青色像素值．为避免对后期图像青色提取产生影响，未对图片进行去噪声和滤波处理．但为排除烟叶本身之外因素的干扰，对拍摄的原始烟叶图像进行了背景剔除处理．由于烟叶原始图像阴影干扰明显，而透光图像几乎不存在阴影，因此，采用透光图像对烟叶正面和背面图像进行背景剔除．处理后的烟叶样品图像如图 1 所示．

（a）剔除背景的烟叶原图　（b）经透光处理的烟叶正面图像　（c）经透光处理的烟叶背面图像

图 1　烟叶图像的处理

2.2　主脉图像提取

烟叶为典型异面叶[18]，主脉凸起部分位于叶片背面. 从拍摄的烟叶图片效果来看，在背面透光图像中，可明显地观察到主脉，且显示其与叶面存在明显差异，因此可对背面透光图进行处理以提取清晰完整的主脉图像. 提取后的主脉显示情况如图 2 所示.

图 2　主脉提取图像

2.3　确定烟叶含青区域检测的颜色通道

烟叶图像在 a 通道和 U 通道分量值的显示情况如图 3 所示，为了准确表征含青程度的颜色特征，还需计算 a、U 两个通道的百分位数特征.

（a）剔除背景的烟叶原图　（b）烟叶样品在 a 通道的分量值　（c）烟叶样品在 U 通道的分量值

图 3　烟叶在不同颜色通道分量的显示

基于青色在 a 通道与 U 通道相对较明显的结果，为从中选出最佳青色检测颜色通道的分位值，计算 17 片参比样烟叶相应颜色通道的分位数与含青程度排序结果的相关性. 将相关性最大的颜色通道分位值作为最终的烟叶含青区域检测的道，具体步骤如下：

1）计算含青程度依次递增的烟叶样品在 a、U 两个颜色通道的像素值分位数，分位数范围分别为（0.1，0.01），每一个分位数作为一个特征，总共得到 200 个分位数颜色特征.

2）计算所有分位数特征以及图片序号在内的相关系数矩阵，得到 200 ×200 的相关系数结果.

3）提取各分位数特征与含青烟叶参比样排序结果的相关系数，相关系数统计情况如图 4 所示. 由图 4 可知，a 通道与含青程度的相关系数普遍且明显高于 U 通道，因此确定 Lab 颜色空间的 a 通道为烟叶含青区域检测通道.

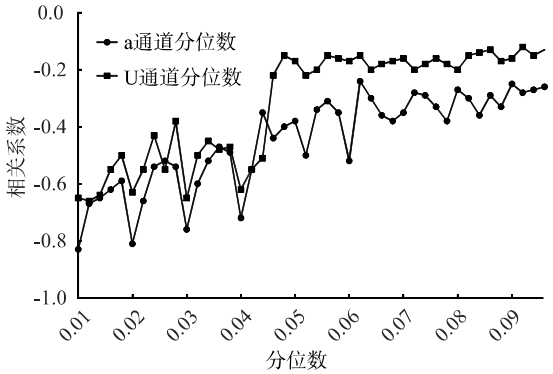

图 4　a 通道和 U 通道分位数特征与含青程度相关系数

4）将 a 通道和 U 通道分位数与含青程度的相关系数结果进行排序统计，并保留排名前 10 的分位数特征，结果如表 2 所示.

表 2　a 通道、U 通道与含青程度相关系数排名前 10 的分位数特征

分位数	a-0.01	a-0.02	a-0.03	a-0.04	a-0.05	U-0.03	U-0.01	a-0.06	U-0.02	U-0.04
相关系数	-0.83	-0.81	-0.76	-0.72	-0.68	-0.65	-0.65	-0.64	-0.63	-0.62

由表 2 可知，a 通道分位数 0.01，0.02，0.03，0.04，0.05，0.06，U 通道分位数 0.01，0.02，0.03，0.04，与含青程度的相关性较高．其中，a 通道 0.01 分位数与含青程度的相关系数最大．因此，将 a 通道 0.01 分位数作为青色检测的颜色通道．

2.4　确定烟叶含青区域检测阈值

为了在 Lab 颜色空间的 a 通道更为准确地提取青色，需明确与含青程度相关性最高的颜色阈值（L1，L2）．a 通道颜色像素值范围为（128，255），结合实际烟叶图像，设定青色像素值合理范围为（128，145）．对处于此范围的颜色阈值进行穷举，计算相应阈值下的含青比例，并将结果与专家制备的参比样顺序进行相关分析，取相关性最高的合理阈值作为青色检测的阈值．

穷举伪码为：

```
for L1 in range（128，145）：
for L2 in range（128，145）：
if L1 > = L2：
continue
```

各阈值范围下参比样的含青比例与烟叶含青程度顺序的相关系数统计情况如表 3 所示．可知，相关程度达到显著水平以上的阈值范围共 6 组，其中青色检测阈值（141，142）条件下，含青比例与含青程度的相关系数最高，因此确定烟叶含青区域检测阈值范围为（141，142）．

表 3　不同阈值范围下含青比例与含青程度相关系数

阈值范围	（141，142）	（130，141）	（130，137）	（142，144）	（130，133）	（138，140）
相关系数	0.63	0.49	0.48	0.45	0.39	0.34

2.5　含青程度识别模型测试

将筛选后的正组、微带青、青黄烟叶样本分别按照 8:2 的比例划分为训练集和测试集，采用 Lab 颜色模型的 a 通道 0.01 分位数为建模特征进行模型测试，模型对测试集的混淆矩阵如图 5 所示．测试集 243 个样本中，对于正组烟叶、微带青烟叶、青黄烟叶的综合分类准确率达到 86.01%．从混淆矩阵图可以看出，构建的含青程度识别模型对正组烟叶和微带青烟叶的区分能力较好．正组烟叶与微带青烟叶的测试集有 81 个样本，其中，正组烟叶预测准确率达到 95.61%，微带青烟叶预测准确率达到 87.65%．模型对青黄烟叶的预测准确率为 75.31%，其问题主要为青黄烟叶与微带青烟叶识别界限不清晰，识别准确率有待提升．

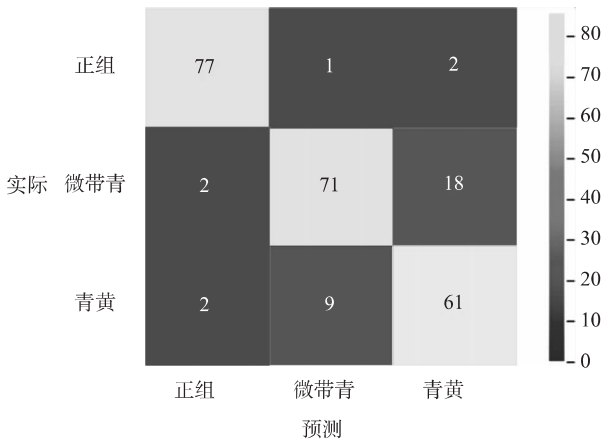

图 5　模型在测试集上的混淆矩阵

3 结论与讨论

含青烟叶的准确识别是烟叶分级领域的难点，本试验在烟叶样品整理过程中制备出一套含青程度参比样，通过已知烟叶含青程度排序进行烟叶图像处理方式的选择，以及含青区域颜色检测通道和检测阈值的筛选，保证了从烟叶图像中提取含青区域参数设置的合理性和科学性. 烟叶正面透光图像可全面反映叶脉、叶面的信息，因此针对烟叶正面透光图像进行含青区域提取. 考虑烟叶为典型异面叶，叶面正面和背面主脉图像所呈现的信息不同，本文提出了一种通过烟叶背面透光图像提取主脉图像的方式. 在实际烟叶生产中，对于烟叶等级的判定基本以叶片正面为主，并且在后续分析中发现，判断烟叶含青程度的更佳方式是通过颜色通道的分位数来表征[19]，而分位数不存在位置信息，因此本试验未对提取的主脉图像进行单独分析.

本试验采用图像处理提取识别微带青烟叶和青黄烟叶的含青区域，并基于最优颜色模型，采用SVM算法构建不同含青程度烟叶的识别模型，模型对测试样本的研究结果表明准确率达到86.01%. 混淆矩阵图显示模型对正组烟叶和微带青烟叶的识别准确率较高，而对于青黄烟叶的识别准确率相对偏低，识别错误表现为将部分青黄烟叶识别为微带青叶. 分析认为，与分级人员通过感官辨别含青程度的方式相比，由于存在像素和拍摄角度限制等问题，采用相机拍摄烟叶图像提取含青区域较易引起图像出现不同程度的失真情况[20]，后期应将烟叶图像与分级人员识别结果相结合进行再验证，最大限度地保证烟叶图像与分级人员感官的匹配，使烟叶图像所反映的信息与实际烟叶相符度更高. 此外，样本的数量与代表性对模型的训练学习效果有着重要影响. 由于不同产区烟叶外观质量存在明显差异[21,22]，该模型的普适性有待进一步认证. 本文所构建的模型为基础模型，后续可纳入更多产地烟叶的差异化特征，在模型搭建过程中进行迁移学习，使模型更具泛化能力，逐步提升模型的普适性和准确度.

[参考文献]

[1] 李丹丹，徐清泉，夏琛，等. 陈化对含青烤烟类胡萝卜素降解产物含量及吸食品质的影响 [J]. 烟草科技，2013（4）：52-55.

[2] 朱先志，刘元德，谭青涛. 烟叶含青现象产生的原因及解决措施 [J]. 现代农业科技，2011（18）：102.

[3] 李小斌，吕志峰，王科杰，等. 加酶技术提高烟叶感官质量的研究 [J]. 中国烟草科学，2007（6）：9-12.

[4] 程向红. 醇化过程中晒黄烟化学成分及感官质量的变化 [J]. 广西农业科学，2009，40（10）：1339-1341.

[5] 肖和友，邓建功，李宏图. 密集烤房烘烤环境改变对烤烟上部位叶可用性的影响 [J]. 农学学报，2016，6（12）：43-46.

[6] 刘慧力，贾洪雷，王刚，等. 基于深度学习与图像处理的玉米茎秆识别方法与试验 [J]. 农业机械学报，2020，51（4）：207-215.

[7] 周煜博，刘立群. 基于EM-PCNN的果园苹果异源图像配准方法 [J]. 农业工程学报，2022，38（5）：175-183.

[8] 马浚诚，刘红杰，郑飞翔，等. 基于可见光图像和卷积神经网络的冬小麦苗期长势参数估算 [J]. 农业工程学报，2019，35（5）：183-189.

[9] 李少敏，吴忠秀，陈升晖. 基于图像识别技术的玉米种子形态鉴别研究 [J]. 分子植物育种，2022，20（2）：667-671.

[10] 于国庆，郝若帆，马洪涛，等. 基于图像处理和支持向量机的粉碎性秸秆覆盖率的图像识别方法研究 [J]. 河南农业科学，2018，47（11）：155-160.

[11] 李增盛，孟令峰，王松峰，等. 基于图像处理的烟叶烘烤阶段判别模型优选 [J]. 中国烟草学报，2022，28（2）：65-76.

[12] 潘治利，祁萌，魏春阳，等. 基于图像处理和支持向量机的初烤烟叶颜色特征区域分类 [J]. 作物学报，2012，38（2）：374-379.

[13] 史龙飞，宋朝鹏，贺帆，等. 基于机器视觉技术的烤烟鲜烟叶成熟度检测 [J]. 湖南农业大学学报（自然科学版），2012，38（4）：446-450.

[14] 李良钰，苏铁熊，马富康，等. 基于集合经验模态分解-支持向量机的高压共轨系统故障诊断方法 [J]. 兵工学报，2022，43（5）：992-1001.

[15] 初勇志，刘成林，太万雪，等. 基于支持向量机（SVM）的不同咸化程度烃源岩总有机碳含量预测模型 [J]. 石油

实验地质，2022，44（4）：739-746.
[16] 张红涛，刘迦南，谭联，等. 基于机器视觉的烟青虫和棉铃虫雌雄蛹的分类识别［J］. 烟草科技，2020，53（2）：21-26.
[17] 程洪，DAMEROWL，BLANKEM，等. 基于图像处理与支持向量机的树上苹果早期估产研究［J］. 农业机械学报，2015，46（3）：9-14.
[18] 马留军，李峥，张瑞亚，等. 不同部位烟叶烘烤过程中颜色与化学成分之间的关系研究［J］. 天津农业科学，2018，24（9）：60-64.
[19] 乔子昂，刘涛. 颜色通道下的无参考图像质量评价［J］. 激光与光电子学进展，2020，57（12）：269-277.
[20] 孔繁锵. 结合 HVS 和相似性度量的图像质量评价测度［J］. 中国图象图形学报，2011，16（7）：1184-1191.
[21] 李峥，王建峰，程小强，等. 基于 BP 神经网络的烤烟外观质量预测模型［J］. 西南农业学报，2019，32（3）：653-658.
[22] 李峥，谭方利，贺帆，等. 基于 CIE 颜色空间构建烤烟外观质量预测模型［J］. 河南农业科学，2018，47（8）：149-154.

Research on the Containing Cyan Degree of Flue-cured Tobacco Based on Image Processing and Machine Learning

LI Zheng[1]，XU Zhiqiang[1]，ZHANG Xiaobing[1]，LIN Jiayi[2]，
ZHONG Yongjian[1]，XU Junhua[1]，ZHANG Zhaopeng[1]，JIAO Deping[1]
（1. *Technology Center of China Tobacco Zhejiang Industrial Co.，Ltd.，Hangzhou，Zhejiang，China* 310024；
2. *Shanghai Micro Vision Technology LTD，Shanghai，China* 200082）

Abstract：In order to improve the efficiency and accuracy of containing cyan degree identification of tobacco leaves. By constructing tobacco leaf reference samples with different containing cyan content. The best color channel and parameter threshold for the extraction of cyan region are screened by computer image processing technology. Construction of recognition model of tobacco leaves with different green degree based on support vector machine. The front transparent image of tobacco leaf is suitable for the extraction of containing cyan areas. The a-channel of Lab color model is the best way to extract the cyan region in the image. And its threshold range is（141，142）. The test accuracy of SVM model for different green tobacco samples reached 86.01%. The confusion matrix shows that the recognition accuracy of the model for positive group and microstrip green tobacco leaves is superior. Image processing technology and support vector machine have good application effect on the division of containing cyan tobacco leaf.

Key words：flue-cured tobacco；degree of containing cyan；support vector machine；image processing；identification model

（责任编辑：陈伟超）