河南农业科学,2020,49(12):159-163
Journal of Henan Agricultural Sciences
doi:10.15933/j.cnki.1004-3268.2020.12.023

# 基于 Python 爬虫和特征匹配的水稻病害图像智能采集

杨天乐[1,2],钱寅森[1,2],武　威[1,2],孙成明[1,2],刘　涛[1,2]
(1.江苏省作物遗传生理国家重点实验室/江苏省作物栽培生理重点实验室/扬州大学 农学院,江苏 扬州 225009;
2.江苏省粮食作物现代产业技术协同创新中心/扬州大学,江苏 扬州 225009)

**摘要**:为及时诊断和防治水稻病害,通过计算机技术和图像处理技术进行病害诊断。利用 Python 爬虫技术编写基于水稻病害关键词的图像爬虫程序,在此基础上使用 Matlab 图像处理模块的特征匹配对图像集进行筛选,提高图像采集的准确度。结果表明,只利用 Python 爬虫技术获取的水稻病害图像,除胡麻叶斑病外,提取的准确率均高于 50.00%,其中赤霉病提取效果最好,准确率达到 72.7%。而通过特征匹配筛选后图像错检率在 6.00%以下,不仅提高了数据采集的精度,也表明水稻病害图像智能采集方法可行。

**关键词**:水稻病害;Python;特征匹配;图像处理;农业信息
**中图分类号**:S435.111　　**文献标志码**:A　　**文章编号**:1004-3268(2020)12-0159-05

# Intelligent Acquisition of Rice Disease Images Based on Python Crawler and Feature Matching

YANG Tianle[1,2],QIAN Yinsen[1,2],WU Wei[1,2],SUN Chengming[1,2],LIU Tao[1,2]
(1. Agricultural College of Yangzhou University/Jiangsu Key Laboratory of Crop Genetics and Physiology/Jiangsu Key Laboratory of Crop Cultivation and Physiology,Yangzhou 225009,China; 2. Jiangsu Co-Innovation Center for Modern Production Technology of Grain Crop/Yangzhou University,Yangzhou 225009,China)

**Abstract**: For timely diagnose and prevent the rice diseases,computer technology and image processing technology were used for disesae diagnosis. Python crawler technology was used to compile image crawler programs based on rice disease keywords. The feature matching of Matlab image was used to filter the image set to improve the accuracy of image collection. The results showed that the extraction accuracy of rice disease images obtained only by Python crawler technology was higher than 50.00%,except bipolaris oryzae. Among them,the extraction effect of gibberellic disease was the best,with an accuracy rate of 72.7%. The false detection rate of images after the feature matching screening was below 6.00%,which not only improved the accuracy of data collection,but also showed that rice diseases image acquisition through the intelligent method was feasible.

**Key words**: Rice disease; Python; Feature matching; Image processing; Agricultural information

病害是影响水稻稳产、高产的重要因素之一[1],常见的水稻病害有稻瘟病、纹枯病、白叶枯病等[2]。水稻在不同的生长发育时期都极易受到病害的侵染,倘若没有及时发现并进行防治,很可能会

**收稿日期**:2020-02-15
**基金项目**:国家自然科学基金项目(31701355,31671615);国家博士后基金项目(2016M600448,2018T110560);国家重点研发计划项目(2016YFD0300107);2017 年江苏省优势学科项目
**作者简介**:杨天乐(1994-),男,江苏徐州人,在读博士研究生,研究方向:农业信息技术和作物栽培。
E-mail:tianley21@qq.com
**通信作者**:刘　涛(1987-),男,江苏徐州人,讲师,博士,主要从事智能农业、作物图像分析、无人机遥感研究。
E-mail:tliu@yzu.edu.com
孙成明(1973-),男,江苏宿迁人,教授,博士,主要从事作物精准栽培与农业信息技术研究。
E-mail:cmsun@yzu.edu.com

导致病害大面积发生,造成严重损失甚至绝收,同时也会使水稻品质下降[3]。早期农作物病害监测通常是通过植保人员田间取样来判断病害的危害等级,或者通过施药来提前预防,或是通过查询水稻病害图谱与病害信息进行比对,但是这些方法普遍存在着专家依赖性大、效率低、污染环境识别错误等缺陷。近年来随着信息技术的高速发展,各类新兴技术广泛应用于农作物病害的监测,并且多种作物的多项病害都已经能够成功进行监测[4-9]。

Python 是一种高级编程语言[10-11],能够提供比较完善的基础代码库,覆盖了网络、文件、图形用户界面(GUI)、数据库、文本等大量内容,被称作内置电池(Batteries included)。用 Python 开发应用程序,许多功能不必从零编写,直接使用现成的代码库即可[12]。除了内置的库外,Python 还有大量的第三方库,可直接使用。

Matlab 是 MathWorks 公司出品的一款涵盖数值分析、数学建模、图像处理等一系列功能的交互式编程软件。Matlab 应用范围极为广泛,其强大的数据可视化功能、图像处理工具箱内置丰富的专业函数,令其成为图像处理的必备工具。

Python 爬虫从技术层面来讲就是通过程序模拟浏览器请求站点的行为,把站点返回的 HTML 代码/JSON 数据/二进制数据(图片、视频)爬到本地,进而提取需要的数据[13-14]。相比其他语言和工具,Python 语法优美、代码简洁、开发效率高、支持的模块多,相关的 HTTP 请求模块和 HTML 解析模块丰富。还有强大的爬虫(Scrapy),以及成熟高效的爬虫-远程字典服务(Scrapy-redis)分布式策略爬虫框架,方便高效下载网页;多线程、进程模型成熟稳定,多线程或进程会优化程序效率,提升整个系统下载和分析能力。Python 具有非常优秀的第三方包能够模拟用户代理(User agent)的行为构造合适的请求,避免网站对于爬虫的封杀。而且,调用其他接口较方便,但缺点在于对编码的处理。

目前,关于病害诊断技术的报道较多,且 Python 爬虫技术的应用也较多,但利用 Python 爬虫技术进行病害图像分类的研究较少。鉴于此,利用该技术进行水稻不同种类病害图像的爬取,并利用特征匹配技术对错误图像进行剔除,以期完成水稻病害图像的智能采集。

## 1 材料和方法

### 1.1 材料

Python 爬虫技术、Matlab R 2018a 图像处理工具箱、百度图像网站。

软件运行环境:Python 3.7.0,Matlab R 2018a 版本。

### 1.2 方法

1.2.1 图像获取 相比于人工,Python 爬虫技术能够在短时间内获取更多的图像。在病害识别过程中,水稻病害数据库的内存量越大对于识别的精度越高,基于此种目的,本研究技术路线如下(图 1)。

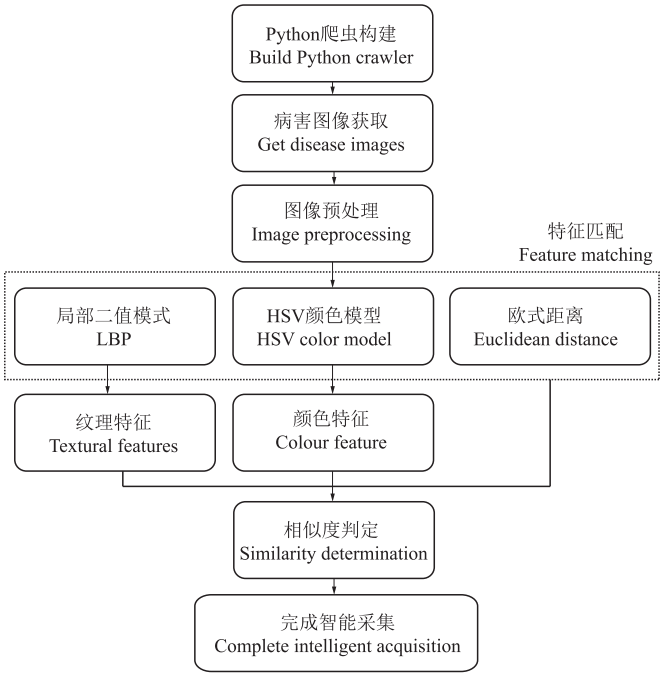


**图 1 技术路线**

**Fig. 1 Technology road**

在水稻病害 Python 爬虫构建中,首先调用以下 4 个模块,即 re、sys、urllib、requests,其中 re 模块可以为使用者直接调用进行正则匹配,减少了代码编写的繁杂程度,也使代码更加简便清晰;sys 模块同样是 Python 中自带的模块,该模块能够向使用者提供对解释器使用或维护的一些变量的访问,以及与解释器强烈交互的函数,能够更好地进行数据的收集;urllib 模块在整个程序中提供上层接口,从而能够从互联网上读取目标图像,同时此模块相较于其他有类似功能的模块优势突出、操作简便、使用门槛低;requests 模块是提供网络访问的模块,具有人性化的特点。通过以上 4 个模块的调用,初步完成爬虫基础的构建,调用代码如下:

```
# coding = utf-8
import re
import sys
import urllib
import requests
```

通过定义函数构建 Python 爬虫程序的主体,设定 1 次获取 150 张图像,通过更改函数中的参数,达到分别获取胡麻叶斑病、白叶枯病、赤枯病、稻瘟病、纹枯病 5 种常见水稻病害图像的目的。

1.2.2　特征匹配　由于 Python 爬虫在获取图像时完全依赖于关键词搜索,所以存在将不符合要求的图像误判断成目标图像的可能性,造成图像获取的不准确性。本研究通过 Matlab 进行特征匹配,并计算图像间的相似度,筛选出同一病害的图像,去除其他干扰项,以提高水稻病害图像获取的准确性。

利用 Matlab 进行图像纹理特征提取是非常成熟的手段,纹理特征提取方法众多,本研究选取局部二值模式进行特征提取。

本研究通过图像的色彩特征和纹理特征值对图像进行相似度比较,本方法采用色调(H)、饱和度(S)、明度(V)颜色模型(HSV)及局部二值模式计算特征值。对于图像相似度采用欧氏距离度量,在进行图像比较前,首先利用如下代码对通过 Python 爬虫技术获取的病害图像进行批量处理,使所有待进行特征匹配的图像统一为 215 像素×215 像素的尺寸,方便后期图像特征值的计算和验证。

```
file_path = 'G:\pictures\';
img_path_list = dir(strcat(file_path,'*.jpg'));
img_num = length(img_path_list);
if img_num>0
    for j = 1:img_num
        image_name = img_path_list(j).name;
        image = imread(strcat(file_path, image_name));
        image = imresize(image,[215 215]);
        fprintf('%d %d %s\n', i, j, strcat(file_path, image_name));
        imwrite(image, strcat('G:\size\', image_name));
    end
end
```

对统一尺寸处理后的图像进行进一步处理,将其转化为 HSV 空间(公式 1)。

$$[H1,S1,V1] = GetHSV(image1)$$
$$[H2,S2,V2] = GetHSV(image2) \quad (1)$$

HSV 空间是一种直观的颜色模型,但为了能够更好地使用颜色参数,需要利用量化函数 Quantificate 对图像 H、S、V 进行量化,将其量化为 36 维向量(公式 2)。

$$[H1,S1,V1] = Quantificate(H1,S1,V1)$$
$$[H2,S2,V2] = Quantificate(H2,S2,V2) \quad (2)$$

经以上处理后,可得到基于 HSV 空间的颜色直方图,可直观比较图像间的颜色差异。

研究发现,通过颜色特征的确能提高病害图像采集的准确度,但在叶片上病害面积较小时,仅通过颜色参数容易误把病害图像识别成非病害图像,造成结果的不准确,因此在颜色特征的基础上又利用纹理特征对图像进行进一步的处理。

利用局部二值模式对图像进行再处理,局部二值模式(公式 3)能够提取局部特征作为相似度评判依据,具有旋转不变性和灰度不变性等优点,结果以像素图的形式记录像素点和周围像素点的差异。

$$LBP(X_c, y_c) = \sum_{P=0}^{P-1} 2^P s(i_P - i_c) \quad (3)$$

其中,$(X_c, y_c)$是中心像素,亮度是 $i_c$;而 $i_n$ 则是相邻像素的亮度。

局部二值模式处理后,通过水稻纹理特征的比较,有助于进行图像相似度的判断,大致可以判断处理的 2 幅图是否为同一类型图像,有效缩小判断的范围,减小相似度判别的难度。颜色特征和纹理特征能够有效地反映图像间的相似性及差异性,但定性判断无法为剔除错误图像提供直观的依据,研究中选择利用欧式距离(公式 4)来定量评定图像间的相似度,欧氏距离越小说明相似度越高。

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} =$$

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (4)$$

## 2 结果与分析

利用 Python 爬虫技术采集水稻病害图像能够快速获取目标数量图像,耗时主要受到网速和电脑配置的限制,但相较于手动获取速度优势明显。在实际运行过程中,由于所爬取网站本身的屏蔽作用,有可能造成图像丢失,即获取图像数量少于预设值,除此之外,由于是通过关键词进行图像的收集,在大量的图像数据里难免出现非病害图像,对图像数据进行统计分析,结果如表 1。

由表 1 可知,实际数量和目标数量并不完全相同,但误差在可接受的范围内,本研究选取的 5 种水稻常见病害,除胡麻叶斑病外,提取的准确率均高于 50.00%,其中赤霉病提取效果最好,准确率达到 72.7%。

通过特征匹配后,病害图片的获取准确性大大提高,将欧氏距离值作为判断的依据,当欧氏距离值小于或等于 5 时,判断为同一种病害图像;当欧氏距离值大于 5 时,判断为不属于同一类型。

5 种常见水稻病害通过特征匹配后的获取结果如表 2。

**表 1 Python 爬虫技术获取的水稻病害图像结果**

**Tab. 1 Results of rice disease images obtained by Python crawler technology**

| 病害 Disease | 目标数量/个 Target quantity | 实际数量/个 Real quantity | 非病害数量/个 Non-disease quantity | 准确率/% Accuracy |
|---|---|---|---|---|
| 胡麻叶斑病 Bipolaris oryzae | 150 | 103 | 72 | 30.1 |
| 白叶枯病 Xanthomonas oryzae | 150 | 149 | 73 | 51.0 |
| 赤霉病 Gibberellic disease | 150 | 150 | 41 | 72.7 |
| 稻瘟病 Rice blast | 150 | 147 | 44 | 70.1 |
| 纹枯病 Banded sclerotial blight | 150 | 142 | 69 | 51.4 |

**表 2 特征匹配后获取的水稻病害图像结果**

**Tab. 2 Obtaining results of rice disease images after feature matching**

| 病害 Disease | 数量/个 Quantity | 欧氏距离≤5 数量/个 Quantity of Euclidean distance≤5 | 准确率/% Accuracy/% | 错检率/% False detection rate/% |
|---|---|---|---|---|
| 胡麻叶斑病 Bipolaris oryzae | 103 | 32 | 30.1% | 1.39% |
| 白叶枯病 Xanthomonas oryzae | 149 | 78 | 51.0% | 2.74% |
| 赤霉病 Gibberellic disease | 150 | 105 | 72.7% | -9.76% |
| 稻瘟病 Rice blast | 147 | 103 | 70.1% | 0 |
| 纹枯病 Banded sclerotial blight | 142 | 77 | 51.4% | 5.80% |

通过表 2 可知,利用特征匹配后病害图像识别准确性明显提高,目标病害的错检率均低于 6.00%,这表明本研究的方法能够高效准确地实现水稻常见病害智能获取。

## 3 结论与讨论

获取水稻病害图像对于及时准确地进行田间病害类型诊断具有积极意义,一些学者利用图像分析技术识别病害种类,这种方法要求高,操作起来相对繁琐[8];或者利用机器学习的方法进行病害类别判断[15-17],但这些方法存在原始样本量不足的缺陷。然而,大部分病害在网络上已经存在大量的图像资料,利用这些图像资料可以较为直观地进行病害种类的判别。本研究利用 Python 爬虫技术进行水稻图像的采集,极大地提高了病害分类工作的效率。另一方面,本研究针对爬虫爬取图像的不准确性,利用图像特征匹配进行改善,并达到了较为准确的水稻病害图像采集的效果。

但本研究也存在一些不足,如在获取过程中无法对图像质量进行识别,这就导致在特征匹配时效果较差,造成病害类别判断的误差;同时由于网站对于爬虫抓取的封杀,可能造成获取数量低于预期目标,接下来的研究应优化爬虫代码以提高程序运行的效率;针对图像的特征匹配准确性还有提升的空间,能够综合不同类别的特征,实现多特征融合,并优化特征提取方式,提高图像描述的准确性及全面性。与此同时,在图像采集的效率上还有待提高,如何实现采集和特征匹配同步进行并降低时间损耗是下一步研究的重点。从研究结果来看,不同类型的水稻病害采集准确性各有不同,差异较大,造成这种现象的原因除技术本身外,还与网络数据量有关,针对不同类型的水稻病害考虑采用不同的识别算法,提高诊断的准确性是一种有必要的尝试。

本研究利用 Python 爬虫技术和图像特征匹配技术实现水稻病害的快速采集,提高了水稻病害识别的工作效率,为病害种类的快速判别提供一种新的技术手段,也为其他作物病害的研究提供了参考。

**参考文献:**


[1] 王艳青. 近年来中国水稻病虫害发生及趋势分析[J]. 中国农学通报,2006,22(2):353-357.
WANG Y Q. Occurrence and trend analysis of rice pests and diseases in China in recent years[J]. Chinese Agricultural Science Bulletin,2006,22(2):353-357.

[2] 陈德西,何忠全,封传红,等. 水稻主要病害发生区划研究[J]. 西南农业学报,2014,27(3):1072-1078.
CHEN D X,HE Z Q,FENG C H,*et al*. Regionalization of major rice diseases[J]. Southwest China Journal of Agricultural Sciences,2014,27(3):1072-1078.

[3] 刘福成,刘立华,黄元财. 水稻不同时期的病虫害防治要点[J]. 吉林农业,2016(16):87.
LIU F C,LIU L H,HUANG Y C. Key points for controlling diseases and insect pests in different periods of rice [J]. Agriculture of Jilin,2016(16):87.

[4] MILLER R J,KOEPPE D E. Southern corn leaf blight:Susceptible and resistant mitochondria[J]. Science,1971,173(3991):67-69.

[5] STEDDOM K,HEIDEL G,JONES D,*et al*. Remote detection of rhizomania in sugar beets[J]. Phytopathology,2003,93(6):720-726.

[6] JING X,WANG J H,SONG X Y,*et al*. Continuum removal method for cotton verticillium wilt severity monitoring with hyperspectral data[J]. Transactions of the Chinese Society of Agricultural Engineering,2010,26(1):193-198.

[7] LI X,LEE W S,LI M,*et al*. Spectral difference analysis and airborne imaging classification for citrus greening infected trees[J]. Computers & Electronics in Agriculture,2012,79(1):32-46.

[8] 刘涛,仲晓春,孙成明,等. 基于计算机视觉的水稻叶部病害识别研究[J]. 中国农业科学,2014,47(4):664-674.
LIU T,ZHONG X C,SUN C M,*et al*. Research on rice leaf disease identification based on computer vision[J]. Scientia Agricultura Sinica,2014,47(4):664-674.

[9] 谢亚平,陈丰农,张竞成,等. 基于高光谱技术的农作物常见病害监测研究[J]. 光谱学与光谱分析,2018(7):2233-2240.
XIE Y P,CHEN F N,ZHANG J C,*et al*. Diseases common monitoring based crop hyperspectral technology[J]. Spectroscopy and Spectral Analysis,2018(7):2233-2240.

[10] OLIPHANT T E. Python for scientific computing[J]. Computing in Science & Engineering,2007,9(3):10-20.

[11] SANNER M F. Python:A programming language for software integration and development[J]. Journal of Molecular Graphics & Modelling,1999,17(1):57-61.

[12] CAI X,LANGTANGEN H P,MOE H. On the performance of the Python programming language for serial and parallel scientific computations[J]. Scientific Programming,2005,13(1):31-56.

[13] 钱程,阳小兰,朱福喜. 基于 Python 的网络爬虫技术[J]. 科学技术创新,2016(36):273.
QIAN C,YANG X L,ZHU F X. Web crawler technology based on Python [J]. Scientific and Technological Innovation,2016(36):273.

[14] 郭丽蓉. 基于 Python 的网络爬虫程序设计[J]. 电子技术与软件工程,2017(23):248-249.
GUO L R. Web crawler program design based on Python [J]. Electronic Technology & Software Engineering,2017(23):248-249.

[15] 杨昕薇,谭峰. 基于贝叶斯分类器的水稻病害识别处理的研究 [J]. 黑龙江八一农垦大学学报,2012,24(3):64-67.
YANG X W,TAN F. Research of treatment of the rice disease recognition based on bayes classifier [J]. Journal of Heilongjiang Bayi Agricultural University, 2012, 24(3):64-67.

[16] 赵建敏,芦建文. 基于字典学习的马铃薯叶片病害图像识别算法[J]. 河南农业科学,2018,47(4):154-160.
ZHAO J M,LU J W. Identification algorithm of potato diseases on leaves using dictionary learning theory[J]. Journal of Henan Agricultural Sciences, 2018, 47(4):154-160.

[17] 张红涛,朱洋,谭联,等. 基于 FA-SVM 技术的烟草早期病害识别[J]. 河南农业科学,2020,49(8):156-161.
ZHANG H T,ZHU Y,TAN L,*et al*. The recognition of early tobacco disease based on FA-SVM technology[J]. Journal of Henan Agricultural Sciences, 2020, 49(8):156-161.