



大田作物病害识别研究图像数据集

陈雷^{1*}, 袁媛^{1*}

1. 中国科学院合肥智能机械研究所, 合肥 230031

摘要: 根据联合国粮农组织报告, 每年农业病虫害造成的自然损失率超过 37%, 农业病虫害识别与防治对于提高农业产量具有重要意义。传统人工识别方法依赖经验, 主观因素较大, 不够准确。近年来计算机视觉方法逐渐发展, 该方法更加客观, 并支持实时在线诊断, 但需要大规模训练样本的支持。因此, 构建可供机器学习建模使用的图像数据集对于实现高效的农业病虫害识别至关重要。为此我们构建了农业病虫害研究图库 (IDADP), 涵盖农业病虫害图像采集、分类、标记、存储与建模等多方面的内容, 面向科研学者与农技人员两大类用户群体提供农业病害在线诊断及相关的技术咨询等服务。本数据集目前包括以水稻、小麦、玉米为主的大田作物的高质量农业病害图像数据约 200 GB。与现有大多仅含有 3-5 幅典型症状图像的农业病害图谱类资源存在本质区别, 本图像数据集由高分辨率和高相似度的同类农作物病害原始图像数据构成, 每种病害的图像数量有几百乃至上千幅, 可作为病害识别建模的训练样本使用。本数据集将为农业病害识别研究领域提供宝贵的基础数据资源, 同时可作为大数据环境下机器学习建模的标准图库, 对促进农业病害图像识别研究的发展具有重要的实际应用价值。

关键词: 农业病害; 大田作物; 病害识别; 标准图库; 训练样本

文献 DOI:

10.11922/csdata.2019.0008.zh

数据 DOI:

10.11922/sciencedb.745

文献分类: 信息科学

收稿日期: 2019-03-20

开放同评: 2019-04-17

录用日期: 2019-06-26

发表日期: 2019-12-31

数据库 (集) 基本信息简介

数据库 (集) 名称	大田作物病害识别研究图像数据集
数据作者	陈雷、袁媛
数据通信作者	陈雷 (chenlei@iim.ac.cn); 袁媛 (yuanyuan@iim.ac.cn)
数据时间范围	2013-2018年
地理区域	中国境内
数据量	200 GB
数据格式	SQL Server
基金项目	中国科学院信息化专项 (XXH13505-03-104); 国家自然科学基金面上项目 (31871521)。
数据服务系统网址	http://www.scidb.cn/journalDetail?dataSetId=633694461276192770&code=5c36e22c13f6b34064283d5e&tID=journalOne&dataSetType=journal&language=zh_CN&lan=2
数据库 (集) 组成	本数据集共有水稻、小麦和玉米3种大田作物的15种病害图像, 每种病害对应一个文件夹。其中水稻病害6个文件夹, 包括水稻白叶枯病、水稻稻曲病、水稻稻瘟病、水稻胡麻斑病、水稻纹枯病、水稻细菌性条斑

* 论文通信作者

陈雷: chenlei@iim.ac.cn

袁媛: yuanyuan@iim.ac.cn

数据库（集）组成

病；小麦病害5个文件夹，包括小麦白粉病、小麦赤霉病、小麦梭条斑花叶病、小麦雪霉叶枯病、小麦叶锈病；玉米病害4个文件夹，包括玉米大斑病、玉米南方锈病、玉米小斑病、玉米锈病。每个文件夹中包含该病害图像以流水号命名的原始JPG文件，以及介绍该病害基本信息与防治方法的intro.txt文件。本数据集共有高质量的jpg图像17 624张。

引言

近些年来，利用计算机视觉和人工智能等技术进行农作物病虫害防治，为农作物病虫害的无损检测和智能化诊断提供了新的方式和思路。尤其是2006年Hinton提出深度学习以来，在诸多领域尤其是图像分类方面取得了显著的效果，为进一步提高基于计算机视觉技术的农业病虫害图像识别效果提供了新的思路。而深度学习方法的效果依赖于大规模的训练数据，因此出现了图像识别最大数据库ImageNet、最有影响的人脸图像数据库LFW等著名的标准图像数据集。同样，在农业病虫害图像识别领域则需要大规模的农业病虫害基础图像资源。农业病虫害图像数据库的规模和质量在很大程度上决定了病虫害图像识别系统的效果。建设规模化、标准化、可共享的农业病虫害图像资源是该领域应当先行的基础研究。

目前现有的农业病虫害图像资源大多是图谱的形式，如纸质出版物的病虫害图谱包括《中国蔬菜病虫原色图谱》^[1]《中国果树病虫原色图谱》^[2]《中国经济作物、粮食作物、药用植物病虫害原色图鉴》^[3]等。几经改版，现已成为主流的病虫害图谱书籍。本世纪以来，网络版的农业病虫害数据库在国内外开始出现并应用，国际上最权威的是国际农业和生物科学中心编辑出版的作物保护大全检索系统（CABI-CPC, Center of Agriculture and Biological International-Crop Protection Compendium）^[4]，收录了世界上150多个国家和地区、3000多种有害昆虫、病害和天敌的信息，部分配有图片。国内电子版的病虫害图谱有中国农业科学院作物科学研究所建立的作物病虫害数据库、河北科技师范学院研制的智能蔬菜病虫害诊断与防治专家系统^[5]、安徽省农业科学院农业经济与信息研究所开发的农业病虫害图文基础数据库、中国医学科学院药用植物研究所植物保护中心开发的药用植物病虫害数据库等。以上这类图谱，对每种病虫害仅给出几张典型症状图片，主要是以科普介绍和形象对照应用为主，不能作为机器学习方法的训练数据集，无法应用于后续的病虫害图像计算机识别方法的研究。由于农作物品种繁多、地域差别以及学科交叉等原因，目前我国还没有建立起可供机器学习使用的农作物病虫害识别研究标准图像数据集。

文章作者在有关项目的支持下，通过采集、整合数据，建设了农业病虫害研究图库（IDADP）。此次公开的数据集包含大量的水稻、小麦、玉米等作物病害图像资源，每种病害有几百乃至上千幅图片，其原始图片分辨率达到2000万像素，建立了一个可为机器学习建模提供训练和测试样本的农作物病虫害识别研究图像数据集。

1 数据采集和处理方法

本研究中图像采集设备为Canon EOS 6D型数码单反相机，配备佳能EF 17-40mm f/4L USM镜头与佳能EF 100mm f/2.8L IS USM微距镜头，以及索尼DSC-RX100M3数码相机。拍摄时采用相机的最优画质与最大分辨率5472×3648，其中佳能数码单反相机采用原始的RAW格式，其后在计算

机上使用佳能 Digital Photo Professional 软件将 RAW 文件转化为 JPG 图像文件，索尼数码相机直接采用 JPG 格式进行拍摄。拍摄时采用光圈优先模式，调整适当的光圈使得图像景深足够大，以保障被拍摄的作物器官在画面中有一定的清晰度。

在采集作物病害图像时，主要是在露天或大棚的自然光照条件下，拍摄角度使光路尽量垂直于作物器官所在平面，并利用侧面光保证作物器官受光均匀，所拍摄的作物器官占据画面的中央主要位置。

2 数据样本描述

本数据集以图像数据库的形式进行存储与管理。其中第一层图像数据词典以 Microsoft SQL Server 的数据表形式存储，数据表的各字段说明以示例如表 1 所示，其中图像路径是指向存放在计算机硬盘上的原始图像文件的索引，作为数据库的主键。此外还有一些图像数据的关键信息，如拍摄时间、拍摄地点、图像大小等，已经蕴含在图像文件的 exif 信息中，因此不在数据表中重复表示。本数据集结构设计较简单，便于数据集管理人员的日常维护与农业病害识别研究人员与农技专家等用户的直接使用。

表 1 作物病害图像数据词典的数据表字段与示例

列 1	列 2	列 3	列 4	列 5
作物名称	器官名称	病虫害名称	图像路径	备注
小麦	叶部	赤霉病	小麦赤霉病\IMG2015.jpg	-
小麦	叶部	白粉病	小麦白粉病\IMG1882.jpg	-
小麦	叶部	白粉病	小麦白粉病\IMG1883.jpg	-
水稻	叶部	稻瘟病	水稻稻瘟病\DSC18_2083.jpg	-
水稻	穗部	稻曲病	水稻稻曲病\IMG17_5657.jpg	-

第二层是图像基础数据，保存在计算机硬盘上。本数据集按照作物病害名称建立文件夹，共有 15 个文件夹。其中水稻病害 6 个文件夹，包括水稻白叶枯病、水稻稻曲病、水稻稻瘟病、水稻胡麻斑病、水稻纹枯病、水稻细菌性条斑病；小麦病害 5 个文件夹，包括小麦白粉病、小麦赤霉病、小麦梭条斑花叶病、小麦雪霉叶枯病、小麦叶锈病；玉米病害 4 个文件夹，包括玉米大斑病、玉米南方锈病、玉米小斑病、玉米锈病。每个文件夹中包含该病害图像以流水号命名的原始 JPG 文件，以及介绍该病害基本信息与防治方法的 intro.txt 文件，每张图像代表一个数据样本。例如本数据集中水稻白叶枯病有 974 张图像，则文件夹“水稻白叶枯病”中含有 974 张 JPG 图像样本以及 1 个介绍水稻白叶枯病基本信息与防治方法的 intro.txt 文件。本数据集中的部分图像样本示例如图 1 所示。



图1 大田作物病害识别研究图像数据集中的样本示例

3 数据质量控制和评估

本研究中采集的作物病害图像来源主要有两种方式：一种是人工接种，该方式下所拍摄的作物病害图像完全能够保障分类准确；另一种方式是拍摄各类生产基地的作物病害图像，并没有经过人工接种，病害是自然发生的，因此所拍摄的病害图像后期经过植保专家的实验室分析与鉴定，以保障病害图像的分类准确。同时，图像采集工作由从事农业病害图像识别研究的专业技术人员按照标准操作流程和规范进行，并在后期对所采集的图像进行人工筛选，将不符合要求的文件剔除，保证本数据集中作物病害图像数据来源的质量和可靠性。

4 数据价值

本数据集与现有作物病虫害图谱的最主要区别是本数据集中的每种病害图像具有几百乃至上千张，可以通过机器学习方法建立病害识别相关的模型，为作物病害图像研究领域提供基础数据资源。随着本数据集的发展，今后将建设成为国内标准的农业病虫害图像识别研究数据资源，为业内相关研究人员提供统一的训练集与测试集数据，从而使得不同的方法可以在同一数据集下进行比较，推动农业病虫害图像识别研究的发展。

5 数据使用方法和建议

由于作物病害发生是一个复杂的过程，同种病害在不同品种、不同部位上的表现症状也不完全

相同，因此使用本数据集的图像数据时可按照具体需求进行二次筛选。同时由于作物病害图像原始数据文件较大，建立在用于机器学习模型训练时根据所选取的框架进行相应的压缩或裁剪。

本数据集中的大田作物病害图像没有对病害位置进行人工标记，在构建模型时需要一定数量的无病害的作物图像作为正样本，所需的正样本图像数据可前往 IDADP 网站下载 (http://www.icgroupcas.cn/website_bchtk/index.html)，因此正样本数据不单独在本数据集中提供。

此外，由于数据版权原因，本文中所上传的图像数据均带有水印，如需获取无水印的原始图像数据需要前往图库网站联系数据库负责人，下载并签署相应的使用协议。

致 谢

感谢安徽省农业科学院的戚仁德、王士梅、王大刚、张爱芳、陆丽娟、黄亮、汪涛、董伟和安徽农业大学金秀提供作物病害图像拍摄基地，协助对作物病害图像进行鉴定与分类。

数据作者分工职责

陈雷 (1981—)，男，安徽省巢湖市人，博士，副研究员，研究方向为机器学习理论方法及在大数据环境下的应用。主要承担工作：本数据集平台的规划、建设和维护。

袁媛 (1981—)，女，安徽省肥东县人，博士，副研究员，研究方向为计算机视觉方法及在农业病虫害图像识别中的应用研究。主要承担工作：本数据集原始数据的采集、整理与相关软件工具的研发。

参考文献

- [1] 吕佩珂. 中国蔬菜病虫原色图谱[M]. 北京: 中国农业出版社, 1992.
- [2] 吕佩珂. 中国果树病虫原色图谱[M]. 北京: 华夏出版社, 1993.
- [3] 吕佩珂. 中国粮食作物、经济作物、药用植物病虫原色图鉴[M]. 呼和浩特: 远方出版社, 1999.
- [4] International C, Wallingford (RU). Crop protection compendium[M]. Berlin: Blackwell Verlag GmbH, 2006.
- [5] 王久兴, 刘介丹, 陈秀敏. 基于 .NET 的智能蔬菜病虫害诊断与防治专家系统构建[J]. 河北科技师范学院学报, 2013, 27(2): 18-22.

论文引用格式

陈雷, 袁媛. 大田作物病害识别研究图像数据集[J/OL]. 中国科学数据, 2019, 4(4). (2019-06-11). DOI: 10.11922/csdata.2019.0008.zh.

数据引用格式

陈雷, 袁媛. 大田作物病害识别研究图像数据集[DB/OL]. Science Data Bank, 2019. (2019-03-20). DOI: 10.11922/sciencedb.745.

An image dataset for field crop disease identification

Chen Lei^{1*}, Yuan Yuan^{1*}

1. Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei 230031, P.R. China

*Email: chenlei@iim.ac.cn, yuanyuan@iim.ac.cn

Abstract: According to the report of Food and Agriculture Organization of the United Nations, the annual natural loss rate caused by agricultural pests and diseases reached more than 37%. Identification and control of agricultural pests and diseases is significant for improving agricultural yield. Traditional manual recognition methods are not accurate enough since they rely on subjective experience. In recent years, computer vision-based methods have developed gradually. These methods are more objective and support real-time online diagnosis. As these methods depend on large-scale training samples, building an image dataset for machine learning modeling is very important for efficiently identifying agricultural diseases and pests. Therefore, we have constructed an image dataset for agricultural diseases and pests research (IDADP) which covers such aspects of agricultural diseases and pests as image acquisition, classification, labeling, storage and modeling. Meanwhile, this image dataset provides online diagnosis of agricultural diseases and related technical consultation services for scholars and agricultural technicians. The image dataset currently has about 200 GB of high-quality agricultural disease images, including field crops such as rice, wheat and corn. Essentially different from existing agricultural disease map resources which mostly contain only 3 to 5 typical symptom images, our dataset consists of the original image data of the same kind of crop diseases with high resolution and high similarity. Each disease has hundreds or even thousands of images, which can be used as training samples for machine learning modeling of disease identification. As a standard dataset for machine learning modeling in large data environment, this image dataset will provide valuable basic data resources. And it has important applicability in promoting the development of agricultural disease identification.

Keywords: agricultural disease; field crops; disease identification; standard image dataset; training sample

Dataset Profile

Title	An image dataset for field crop disease identification
Data corresponding author	Lei Chen (chenlei@iim.ac.cn), Yuan Yuan (yuanyuan@iim.ac.cn)
Data authors	Lei Chen, Yuan Yuan
Time range	2013 – 2018
Geographical scope	China
Data volume	200 GB
Data format	SQL Server
Data service system	< http://www.scidb.cn/journalDetail?dataSetId=633694461276192770&code=5c36e22c13f6b34064283d5e&tID=journalOne&dataSetType=journal&language=zh_CN&lan=2 >
Sources of funding	The 13th Five-year Informatization Plan of Chinese Academy of Sciences (XXH13505-03-104); National Natural Science Foundation of China (31871521).

Dataset composition	The dataset contains 15 disease images of rice, wheat and maize, each of which corresponds to a folder. Concretely, there are 6 folders of rice diseases, including bacterial blight of rice, rice false smut, rice blast, rice brown spot, rice sheath blight and rice bacterial leaf streak; 5 folders of wheat diseases, including wheat powdery mildew, wheat head blight, wheat spindle streak mosaic virus, <i>gerlachia nivalis</i> and wheat leaf rust; and 4 folders of maize diseases, including corn northern leaf blight, southern corn rust, corn southern leaf blight and corn rust. Each folder contains the original JPG files named by pipeline number of the disease image and the intro.txt file which introduces the basic information of the disease and its control methods. This dataset contains 17 624 high quality JPG image data.
----------------------------	---