www.csdata.org

ISSN 2096-2223
CN 11-6035/N

文献 DOI:
10.11922/csdata.2018.0092.zh
数据 DOI：
10.11922/sciencedb.706
文献分类：信息科学

收稿日期：2018-12-14
开放同评：2019-02-14
录用日期：2019-04-28
发表日期：2019-09-19

蝴蝶物种自动识别研究的生态照片数据集

谢娟英[1]，曹嘉文[1]，马丽滨[2]，甄文全[3]，陈振宁[2,4]，

李晓东[5]，李后魂[6]，许升全[2*]

1. 陕西师范大学计算机科学学院，西安　710119
2. 陕西师范大学生命科学学院，西安　710119
3. 钦州学院海洋学院，广西钦州　535011
4. 青海师范大学生命科学学院，西宁　810008
5. 河池学院化学与生物工程学院，广西河池　546300
6. 南开大学生命科学学院，天津　300071

摘要： 蝴蝶物种的自动识别能够为生态环境监测、生物多样性保护、害虫防治等领域提供技术支持，减少对专家人工鉴定的依赖。针对目前蝴蝶自动识别研究中主要依靠标本照片，缺少蝴蝶自然生态照片数据的状况，本数据集提供了一组蝴蝶生态照片、对应的 PASCAL VOC 2007 格式的标注文件以及每张照片的掩模。本数据集包含的 721 张中国蝴蝶生态图片，涵盖 94 种蝴蝶。每张照片对应一个标注文件，标注文件中包含照片大小、照片中蝴蝶的种类和位置等信息。每张照片的掩模对应照片中的蝴蝶区域。本数据集可为自动识别、目标检测、图像分割和昆虫分类等研究提供基础数据。

关键词： 蝴蝶分类；生态照片；自动识别；目标检测

数据库（集）基本信息简介

数据库（集）名称	蝴蝶物种自动识别研究的生态照片数据集
数据作者	谢娟英、曹嘉文、马丽滨、甄文全、陈振宁、李晓东、李后魂、许升全
数据通信作者	许升全（xushengquan@snnu.edu.cn）
数据时间范围	2012–2017年
地理区域	中国
数据量	4.48 GB
数据格式	*.jpg, *.xml, *.png
数据服务系统网址	http://www.sciencedb.cn/dataSet/handle/706
基金项目	国家自然科学基金（61673251、31872273）；中央高校基本科研业务费专项资金（GK201701006）。
数据库（集）组成	本数据集由3部分数据组成：（1）JPEGImages.zip是蝴蝶生态照片，共721张；（2）Annotations.zip 是对应生态照片的721个标注文件，包含照片大小、照片中蝴蝶的种类和位置等信息；（3）masks.zip 是对应生态照片的掩模，以图片形式保存，共721张。

* 论文通信作者
许升全：xushengquan@snnu.edu.cn

引　言

蝴蝶是节肢动物门昆虫纲鳞翅目锤角亚目昆虫的统称[1]。其最引人关注的特征就是翅和身体表面由不同形态和色彩的鳞片所形成的绚丽多彩的图案和花纹。这些色彩和花纹与蝴蝶的拟态、保护色、求偶等行为有关；也是千百年来蝴蝶作为文化昆虫而被人们描绘、歌颂的原因[1]。全球目前已知的蝴蝶种类约 18 000 种，每种蝴蝶都基本固定取食少数几种植物，其发生与植物和生态系统的变化密切相关，是重要的环境指示昆虫和生物多样性监测指示生物[2-3]。对蝴蝶物种的快速准确鉴定是生物学多样性保护和蝴蝶文化发展的迫切需求，但人工对蝴蝶进行鉴别和分类是一件非常耗时耗力的工作，且对专家的分类经验依赖度很高。近年来出现了不少关于蝴蝶自动识别的研究。例如，基于内容检索的蝴蝶所属科的识别[4]、基于极限学习机的蝴蝶种类自动识别[5]和基于单隐层神经网络的蝴蝶识别[6]等。然而，这些研究所使用的数据集均为蝴蝶标本的模式照片，且涵盖的种类较少，研究结果无法应用到对蝴蝶生态照片的识别，特别是从生态照片中区别蝴蝶和环境背景。鉴于此，本数据集通过野外拍摄、蝴蝶爱好者捐赠等方式收集、筛选、整理了一组自然环境中的蝴蝶生态照片，并提供了 PASCAL VOC 2007 格式的记录照片中蝴蝶种类和位置信息的标注文件，以及每张照片的掩模。本蝴蝶生态照片数据集是已完成的第三届中国数据挖掘竞赛（国际首次蝴蝶识别大赛）的竞赛数据，包括了竞赛使用的全部蝴蝶生态照片训练数据，是目前世界上唯一的蝴蝶生态照片数据集，解决了蝴蝶自动识别领域现有数据只包含有蝴蝶标本照片且种类较少的局限，填补了蝴蝶自动识别领域尚无生态照片数据的空白。竞赛使用的训练数据还包括周尧先生的《中国蝶类志》[1]全部蝴蝶照片。希望本数据集能为昆虫分类、目标检测和自动识别领域的相关研究提供数据支持。

1　数据采集和处理方法

1.1　数据采集方法

本数据集中蝴蝶的生态照片来源于野外实地拍摄和蝴蝶爱好者的捐赠，均为高清单反相机拍摄所得，保证了照片的质量。

1.2　数据处理方法

照片中的蝴蝶依据《中国蝶类志》[1]《中国蝴蝶图鉴》[7]等蝴蝶分类文献鉴定到物种。我们使用 labelImg 工具标记照片中蝴蝶的位置和类别，并生成 PASCAL VOC 2007 格式的标注文件，蝴蝶位置由矩形框给出，一张照片对应一个标注文件。另外，为了方便图像分割领域研究者使用，我们提供了每张照片的掩模，使用 labelme 工具，用多边形描绘蝴蝶的外边缘，每张照片生成一个标注文件并转换为掩模图片，以 png 格式的图片保存。

2　数据样本描述

数据集共包含 721 张照片，涵盖 94 种蝴蝶。如图 1 所示，自然环境中蝴蝶的照片与蝴蝶标本照片的不同在于姿态各异，有些是正面照、有些是侧面照、有些正在展翅、有些双翅合拢等。特别是为了躲避天敌的捕食，大部分蝴蝶具有拟态和保护色，它们翅和身体的颜色和花纹与周围环境相似，难以辨认。



图 1　生态图片部分样本

蝴蝶生态照片的统计数据如图 2 所示，大部分种类蝴蝶的样本个数在 13 以内，每种蝴蝶至少有 1 个样本，最多包含 61 个样本，呈现典型的长尾分布。

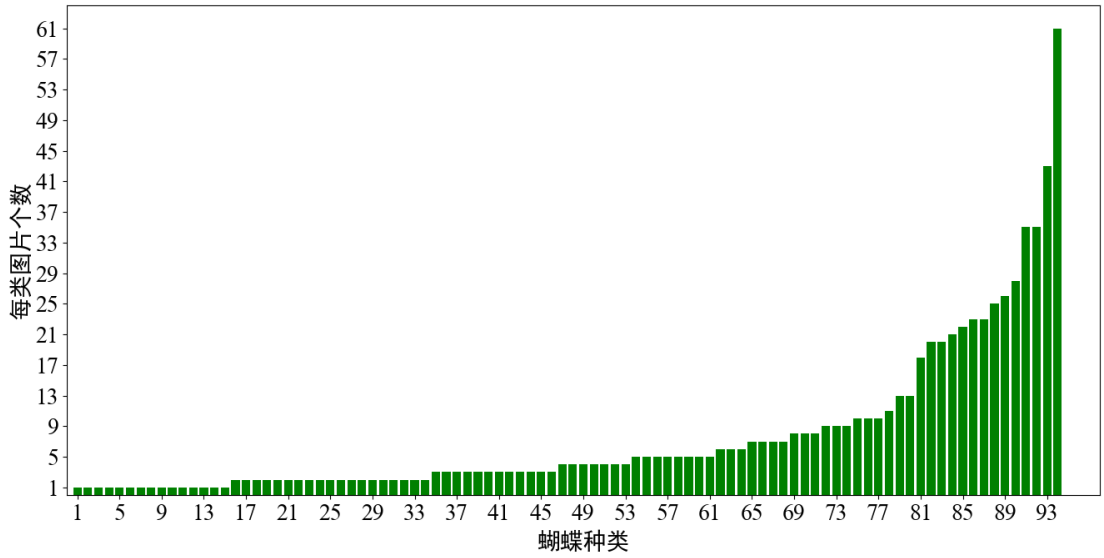


图 2　蝴蝶生态图片数据分布

每张照片都对应一个 PASCAL VOC 2007 格式的 xml 标注文件，文件中包含对应的图片名、图片大小、蝴蝶种类和蝴蝶位置信息。其中蝴蝶分类精确到种，蝴蝶的位置由矩形框标出，在标注文件中记录矩形框的对角位置信息。同时，每张照片对应一张 png 格式的掩模图片，可用于提取蝴蝶精确位置区域，为图像分割等算法评价研究提供支持。

3　数据质量控制和评估

本数据集中每张照片的蝴蝶都由昆虫分类学专家按照《中国蝶类志》等蝴蝶分类文献进行鉴定，保证了数据的准确性。

蝴蝶的位置信息均为人工标注，并经过多次核查，以确保数据质量。

本数据集已在“2018 年第三届中国数据挖掘竞赛——国际首次蝴蝶识别大赛”中发布供参赛者使用，竞赛圆满结束。此次竞赛的圆满成功也说明了本数据集提供的照片数据和标注数据质量可信，可以为相关研究提供数据支持。

4 数据价值

本数据集提供了 94 种蝴蝶在其自然生态环境中的 721 张生态照片，并给出了照片中蝴蝶的种类和位置信息。由于蝴蝶有拟态和保护色，和周围环境不易区别，所以识别生态照片中的蝴蝶种类是目标检测领域的挑战性难题。虽然本数据集只包含了 94 种蝴蝶，没有涵盖全部的中国蝴蝶种类，但蝴蝶物种的识别是同一大类内的小类间区分识别问题，属于细粒度分类，完全不同于常见的目标检测是不同大类物体的识别问题，而且 94 类的分类问题也是一个很挑战的多类分类问题；加上蝴蝶种类的识别主要依赖于蝴蝶翅膀图案，而生态环境中拍摄的蝴蝶照片，其翅膀通常不会完全展开，使得其分类特征很难获得。细粒度多类分类和高难分类特征，使得训练自动识别方法难度非常大。本数据集在为相关研究提供基础数据的同时，也能促进相关实际应用的开发，使得大众及昆虫爱好者能更加方便地了解和认识蝴蝶。

致　谢

本数据集曾在蝴蝶识别大赛中使用，在数据的整理和测试阶段得到了南京大学计算机系高阳教授、北京邮电大学杜军平教授、北京交通大学于剑教授、山东财经大学尹义龙教授、复旦大学张军平教授、南京航空航天大学谭晓阳教授的指导。特别感谢山东财经大学吕鹏副教授、北京交通大学景丽萍教授、南京大学史颖欢副教授带领各自团队在竞赛前对数据验证所做的大量工作。本数据在听取 2018 年第三届中国数据挖掘竞赛——国际首次蝴蝶识别大赛部分参赛队伍建议的基础上进行了完善，对这些队伍和队员表示感谢。

数据作者分工职责

谢娟英（1971—），女，陕西省西安市人，博士，教授，研究方向为机器学习、数据挖掘与生物医学数据分析。主要承担工作：数据集结构设计。

曹嘉文（1995—），女，陕西省西安市人，硕士研究生，研究方向为模式识别与人工智能。主要承担工作：数据的标注。

马丽滨（1980—），男，宁夏回族自治区银川人，博士，副教授，研究方向为昆虫分类学。主要承担工作：照片拍摄。

甄文全（1971—），男，辽宁省彰武县人，博士，副教授，研究方向为昆虫生态学。主要承担工作：照片拍摄。

陈振宁（1970—），男，河南省柘城县人，博士，教授，研究方向为动物系统分类与生态。主要承担工作：照片拍摄。

李晓东（1982—），男，陕西省蒲城县人，博士，副教授，研究方向为昆虫生物多样性。主要承担工作：照片拍摄。

李后魂（1956—），男，江苏省南京市人，博士，教授，研究方向昆虫分类学。主要承担工作：照片拍摄。

许升全（1967—），男，陕西省兴平市人，博士，教授，研究方向昆虫系统与进化。主要承担工作：数据集设计，照片拍摄、筛选。

参考文献

[1] 周尧. 中国蝶类志[M]. 郑州: 河南科学技术出版社, 1998.

[2] ESPELAND M, BREINHOLT J, WILLMOTT K R, et al. A Comprehensive and Dated Phylogenomic Analysis of Butterflies[J]. Current Biology, 2018, 28(5): 770-778.

[3] 马方舟, 徐海根, 丁晖, 等. 我国蝴蝶多样性的保护现状与对策[J]. 世界环境, 2016(s1): 29-31.

[4] WANG J, JI L, LIANG A, et al. The identification of butterfly families using content-based image retrieval[J]. Biosystems Engineering, 2012, 111(1): 24-32.

[5] KAYA Y, KAYCI L, TEKIN R, et al. Evaluation of texture features for automatic detecting butterfly species using extreme learning machine[J]. Journal of Experimental & Theoretical Artificial Intelligence, 2014, 26(2): 267-281.

[6] KANG S H, SONG S H, LEE S H. Identification of butterfly species with a single neural network system[J]. Journal of Asia-Pacific Entomology, 2012, 15(3): 431-435.

[7] 武春生, 序堉峰. 中国蝴蝶图鉴(全四册)[M]. 福州: 海峡书局, 2017.

论文引用格式

谢娟英, 曹嘉文, 马丽滨, 等. 蝴蝶物种自动识别研究的生态照片数据集[J/OL]. 中国科学数据, 2019, 4(3). (2019-04-08). DOI: 10.11922/csdata.2018.0092.zh.

数据引用格式

谢娟英, 曹嘉文, 马丽滨, 等. 蝴蝶物种自动识别研究的生态照片数据集[DB/OL]. Science Data Bank, 2018. (2019-04-08). DOI: 10.11922/sciencedb.706.

A dataset of butterfly ecological images for automatic species identification

Xie Juanying[1], Cao Jiawen[1], Ma Libin[2], Zhen Wenquan[3], Chen Zhenning[2,4], Li Xiaodong[5], Li Houhun[6], Xu Shengquan[2*]

1. School of Computer Science, Shaanxi Normal University, Xi'an 710119, P. R. China
2. College of Life Sciences, Shaanxi Normal University, Xi'an 710119, P.R. China
3. Ocean College, Qinzhou University, Qinzhou, 535011, P.R. China
4. College of Life Sciences, Qinghai Normal University, Xi'ning 810008, P.R. China
5. College of Chemistry and Biological Engineering, Hechi University, Hechi 546300, P.R. China
6. College of Life Sciences, Nankai University, Tianjin 300071, P.R. China

*Email: xushengquan@snnu.edu.cn

Abstract: While reducing reliance on experts, automatic species identification can provide strong technical assistance to ecological environment monitoring, biodiversity protection, pest control. However, all the datasets used for butterfly species auto-identification are currently based on specimen images, and there is a lack of ecological butterfly images. Therefore, we set up a dataset of butterfly images which includes ecological images of the butterfly species in China, corresponding annotation files in PASCAL VOC 2007 format, and a mask for each image. The dataset contains 721 butterfly images, covering 94 butterfly species. Each annotation file corresponds to an image, which contains the image size, taxonomic information and location of the butterfly. The mask file for each image corresponds to the specific location of the butterfly in the image. This dataset can support researches on automatic butterfly species recognition, object detection, image segmentation and insect taxonomy.

Keywords: butterfly taxonomy; ecological images; automatic recognition; object detection

Dataset Profile

Title	A dataset of butterfly ecological images for automatic species identification
Data corresponding author	Xu Shengquan (xushengquan@snnu.edu.cn)
Data authors	Xie Juanying, Cao Jiawen, Ma Libin, Zhen Wenquan, Chen Zhenning, Li Xiaodong, Li Houhun, Xu Shengquan
Time range	2012 – 2017
Geographical scope	China
Data volume	4.48 GB
Data format	*.jpg, *.xml, *.png
Data service system	<http://www.sciencedb.cn/dataSet/handle/706>
Sources of funding	National Natural Science Foundation of China (61673251, 31872273); Fundamental Research Funds for the Central Universities under Grant No. GK201701006
Dataset composition	The dataset consists of 3 subsets in total: (1) JPEGImages.zip is made up of 721 butterfly ecological images; (2) Annotations.zip is made up of 721 annotation files corresponding to the ecological images, and contains the image size, species and location of the butterfly; (3). masks.zip is made up of 721 mask files corresponding to the ecological images.