文献 DOI：
10.11922/csdata.180.2015.0008

文献分类：生物学

收稿日期：2015-07-16
发表日期：2016-11-11

中国凤蝶标本图像特征数据集

王江宁[1]，韩艳[1]，纪力强[1*]

摘　要：中国凤蝶标本图像特征数据集是在对蝴蝶标本图像自动识别的研究中产生的。本数据集收集了使用规范方法从 390 幅经过处理后的中国凤蝶标本图像中提取的三个最常用特征的数值数据，即颜色、形状和纹理特征。本数据集的每条记录都包含了蝴蝶的分类信息、图像编号以及特征信息，为模式识别、昆虫分类等研究提供了基础数据。

关键词：凤蝶科；标本图像；图像特征；模式识别

数据库（集）基本信息简介

数据库（集）中文名称	中国凤蝶标本图像特征数据集		
数据库（集）英文名称	Feature dataset of Chinese papilionidae specimen image		
通讯作者	纪力强（ji@ioz.ac.cn）		
数据作者	王江宁，韩艳，纪力强		
数据时间范围	1994 年		
地理区域	中国		
数据格式	.xlsx	数据量	1.4 MB
数据服务系统网址	http://www.sciencedb.cn/dataSet/handle/8		
基金项目	国家自然科学青年基金（31501841）		
数据库（集）组成	收集了 390 幅中国凤蝶标本图像的分类信息，从中提取的颜色、形状、纹理特征数据，以及原始标本缩略图的索引		

引　言

基于图像的模式识别被应用于各个领域[1]，但多数是在与人体或者机器人相关的领域，对于昆虫图像的研究相对较少。这由两方面原因造成：一方面，昆虫学方面的数据库不少[2]，但是适合模式识别研究的图像数据库却不多[3]；另一方面，昆虫图像不少，但是带有可靠分类信息的昆虫图像的数据不如人脸、指纹等图像数据容易获取。因此，适合模式识别研究的昆虫学数据集相对较少，限制了模式识别在昆虫学中的应用。本数据集收集整理了《中国蝶类志》（1994 版）[4] 中的标本图片，经过一系列处理后形成单一背景图像的特征数据。这些特征均使用作者提出的特征提取方法[5] 从图像中提取而来，并配有准确的分类学信息，对于模式识别方法、昆虫分类学的研究人员有重要的意义。

1　数据采集和处理方法

原始数据是扫描处理存档的蝴蝶标本的图像数据。这些图像经过背景切割后形成纯色背景图像，保留了主体的色彩。我们使用已经公开发表的固定方法[5] 提取处理后的图像特征，本方法的主要处理模块如图 1 所示。

1. 中国科学院动物研究所，北京 100101；
* 通讯作者（Email: ji@ioz.ac.cn）

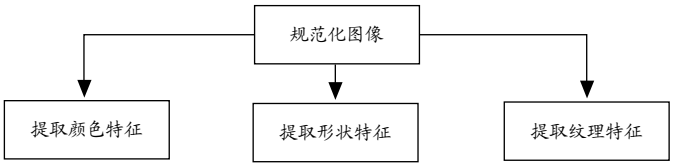

图 1 图像特征提取的主要模块

1.1 规范化图像

本过程由 2 个步骤构成：

步骤 1：去除背景颜色，保留主体（蝴蝶标本），获取主体的最小包络矩形。

步骤 2：从图像中剪裁出最小包络矩形，将剪裁的图像缩放到 512×512 px，获得规范图。

1.2 提取颜色特征

本过程由 4 个步骤组成：

步骤 1：计算主体的重心以及长径（主体上到重心最远点的距离）。

步骤 2：以重心为圆心，以长径的 1/2 作圆，将图像分为内、外两区。

步骤 3：分区提取颜色直方图，按照 R、G、B 三通道和 16 位量化数进行统计。

步骤 4：规一化直方图到 [0, 1]，得到两组 48 维的直方图向量（内区一组，外区一组）。

1.3 提取形状特征

本过程由 5 个步骤组成：

步骤 1：同 1.2 节步骤 1。

步骤 2：以重心为原点，以向右的水平线段（与长径等长）为扫描线，顺时针扫描图像，按照间隔 1° 的角度和 1/100 的长径进行采样，将规范图转换成极坐标下的扫描图，扫描图应为 360×100 px。

步骤 3：统计每条扫描极线中主体部分的数量，获得形状直方图，得到一组 360 维的值在 [0, 100] 的向量。

步骤 4：将上述向量值重新量化成 20 位，并将 360 个重新量化到 10 位的向量值重新统计，获得极线的直方图（20 维 [0, 360) 的向量），即 20 个级别的极线各有多少根极线。

步骤 5：将上述直方图进行规一化处理，得到一组 20 维 [0, 1] 的向量。

1.4 提取纹理特征

本过程由 5 个步骤组成：

步骤 1：同 1.2 节步骤 1。

步骤 2：将彩色图转换成灰度图，使用公式 v=(R+G+B)/3 转换。

步骤 3：以重心为原点，以向右的水平线段为扫描线，按照间隔 1° 顺时针扫描图像，统计每条扫描极线中主体部分的相邻灰度值变化大于 16 的数量，并将统计值量化成 10 级，得到一组 360 维（360 条极线）的值在 [0, 10] 的向量。

步骤 4：将 360 个重新量化到 10 位的向量值重新统计，获得极线的直方图（10 维 [0,360) 的向量），即 10 个级别的极线变化各有多少根极线。

步骤 5：将上述直方图进行规一化处理，得到一组 10 维 [0, 1] 的向量。

1.5 特征数据集

1.1 节中步骤是提取各种特征的必需步骤，通过 1.2、1.3、1.4 节描述的步骤提取了颜色、形状、纹理特征。而每幅图像的分类信息是人工标记的。最终，特征数据、图像代号以及分类信息构成本数据集。

2 数据样本描述

数据集以表格形式进行存储，表格的字段说明见表 1。表 2 给出了一组数据的示例。其中，图像代码的命名规则为：“-”前的数字是标本图像在《中国蝶类志》中的页码；“-”后的数字是自定义的图像编号，按顺序编号。数据集中还存放有原始标本缩略图的索引，便于用户追溯原始图像标本。本数据集中的特征代码意义如表 3 所示。其中特征代码中的下划线前半部是特征向量的代码，后半部是特征向量的维数编号。如 cbirT6_0 表示形状特征向量的第 1 个值，cbirT9_48 表示外区色彩特征向量的第 1 个值。分类学名字段以“属名 种名”格式提供。本数据集结构较简单，此设计便于模式识别研究人员直接使用或者稍作调整后使用。

表 1 数据集的元数据描述

字段名	字段意义	字段类型
tk_name	特征代码	字符
tk_value	特征值	数值（双精度实数）
img_path	图像代码	字符
taxa_latin	分类学名	字符（拉丁学名）

表 2 范例数据

tk_name	tk_value	img_path	taxa_latin
cbirT6_0	1.00000000000000000000	100-1.png	*Troides magellanus*
cbirT6_1	0.10961250000000000000	100-1.png	*Troides magellanus*
cbirT6_10	0.62510380000000000000	100-1.png	*Troides magellanus*
cbirT6_11	0.84378030000000000000	100-1.png	*Troides magellanus*
cbirT6_12	0.62506570000000000000	100-1.png	*Troides magellanus*

表 3 数据集中的特征代码

特征代码	意义
cbirT9_[0-47]	48 维内区色彩特征
cbirT9_[48-95]	48 维外区色彩特征
cbirT6_[0-19]	20 维的形状特征
cbirT7_[0-9]	10 维纹理特征

3 数据质量控制和评估

本数据集的特征提取方法已经公开发表，并得到了领域同行的肯定。数据提取过程也已程序化，设计并完成了相应的组件。该组件经过多次人工验证，运行稳定可靠。本数据集通过程序调用该组件生成，由固定特征提取模块进行特征提取的操作，保证了特征数据的可靠性，不会产生误差数据。

原始图像信息经过人工采集、校对和三次核对，保证了分类信息数据的可靠性。鉴于分类学中分类系统的会随学科的发展而变化，本数据集中图像的分类信息以 1994 版《中国蝶类志》分类系统为标准。

4 数据使用方法和建议

本数据集数据形式简单，在使用时注意：

1. 特征向量值需要将图像编号和特征代码组合后再使用。

2. 颜色特征向量的 0–47 维和 48–95 维可以分成两组向量使用。

3. 特征向量值本身是32位的double型数据，在Excel中显示受到该软件的约束，因此在使用时请根据实际需要酌情选择精度。

致 谢

感谢研究组张荣在原始图像处理中所做的工作。

数据作者分工职责

王江宁（1982—），男，博士，助研，研究方向：昆虫图像识别。主要承担本数据集的规划、建设和维护。

韩艳（1972—），女，学士，工程师，研究方向：生物多样性信息学。主要承担本数据集原始数据的采集和整理。

纪力强（1961—），男，博士，研究员，研究方向：生物多样性信息学。主要承担本数据集的设计。

参考文献

[1] Sá J P M D. Pattern recognition: concepts, methods and applications[M]. Springer, 2001.

[2] 中国科学院动物研究所，昆明动物研究所，上海植物生理生态研究所，等 . 中国动物主题数据库 [DB]. 北京：中国科学院动物研究所，2009.

[3] 王江宁，纪力强 . 分布式分类学图像智能检索框架设计 [C]// 第十届科学数据库与信息技术学术研讨会 . 北京：兵器工业出版社，2010: 481 ～ 485.

[4] 周尧 . 中国蝶类志 [M]. 郑州：河南科学技术出版社，1994.

[5] 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.

引用数据

(1) 王江宁，韩艳，纪力强 . 中国凤蝶标本图像特征数据集 [DB/OL]. Science Data Bank. DOI: 10.11922/sciencedb.180.8.

Feature dataset of Chinese papilionidae specimen image

Wang Jiangning, Han Yan, Ji Liqiang

ABSTRACT Feature dataset of Chinese papilionidae specimen image is created in the researches of butterfly image recognition. This dataset collects the color, texture and shape features from 390 pre-processed specimen images of Chinese papilionidae by standardized feature extraction methods. Each record contains the classification information, image identifier feature ID and feature value of an image. The dataset could support research on pattern recognition and entomology.

KEYWORDS Papilionidae; specimen image; image feature; pattern recognition

引文格式： 王江宁，韩艳，纪力强 . 中国凤蝶标本图像特征数据集 [J/OL]. 中国科学数据，2016, 1(3). DOI: 10.11922/csdata.180.2015.0008.