
第三届中国数据挖掘大赛竞赛规则（新）

（国际首次蝴蝶识别大赛）

第三届中国数据挖掘大赛竞赛规则（新）	1
一、竞赛任务.....	2
二、训练集下载.....	2
三、训练集数据格式.....	2
3.1 蝴蝶生态介集.....	3
3.2 蝴蝶模式照片集.....	3
四、竞赛运行平台环境配置.....	4
4.1 Windows 测试平台环境配置	4
4.2 Ubuntu 测试平台环境配置.....	4
五、程序输入及输出.....	5
5.1 数据输入格式.....	5
5.2 任务结果输出格式.....	5
六、评价指标.....	6
七、竞赛提交.....	6
八、数据集版权.....	7
九、联系方式.....	7

一、竞赛任务

如何自动、快速检测蝴蝶位置和识别蝴蝶种类,进而掌握蝴蝶的生物学特征,是本竞赛需要解决的关键科学问题。本竞赛收集了一批蝴蝶的生态照片且带有准确的类别标签,要求利用《中国蝶类志》提供的标注蝴蝶模式照片以及部分人工标注蝴蝶生态照片,通过机器学习方法检测蝴蝶位置以及建立蝴蝶种类识别模型,对未标注的蝴蝶生态照片中的蝴蝶进行自动分类。具体竞赛任务如下:

(1) **蝴蝶位置检测**。对蝴蝶生态照片进行分析,通过设计相应的算法检测蝴蝶在该照片中的具体位置,并给出蝴蝶矩形框区域坐标。

(2) **蝴蝶分类**。利用提供的人工标注的蝴蝶模式照片和人工标注的蝴蝶生态照片,自动识别未标注蝴蝶生态照片中蝴蝶的种类。

二、训练集下载

下载链接已经发送至报名参赛队伍申请邮箱,请参赛队伍点击下载链接并输入密码进行下载。

注意:参赛队伍不得以任何形式对外公布本次竞赛的下载链接及密码,一旦发现参赛队伍有此类行为,竞赛委员会将取消其参赛资格,并追究相关责任。其他说明请参考《第三届中国数据挖掘竞赛参赛协议》。

三、训练集数据格式

本次竞赛训练集包括两部分:**蝴蝶生态照片集**、**蝴蝶模式照片集**。

蝴蝶生态照片集为在野外使用高清单反相机拍摄所得。其特点是每张照片包含的蝴蝶数量、种类不确定,而且蝴蝶由于自身的拟态躲避天敌的缘故,总是偏向于处在有利于隐藏自身的环境中,使得蝴蝶和周围环境较难辨别。

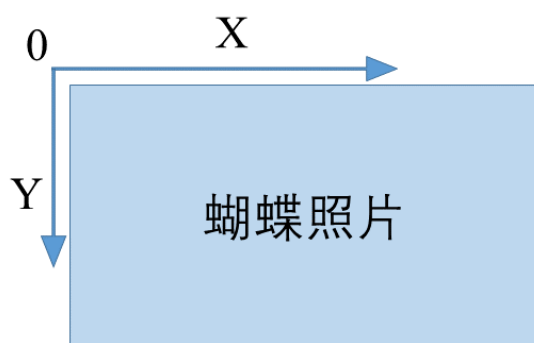
蝴蝶模式照片集(即蝴蝶标本的标准照片集)是通过扫描中国蝶类志的蝴蝶图片得来,共有 4,270 张,1176 种,包含生态照的所有种类。这类图片的特点是一张图片确定地含有一只蝴蝶。将这类图片的蝴蝶种类按照科、亚科、属、种、亚种,以及蝴蝶的雌、雄、背、腹进行分类。由于部分亚种缺失,部分蝴蝶种类雌、雄、背或腹不全,因此,为了统一,将该数据集所有模式照的蝴蝶图片均精确到种。

3.1 蝴蝶生态介集

蝴蝶生态照片集，即【第三届中国数据挖掘大赛-蝴蝶训练集】，包括两个文件夹：Annotations 和 JPEGImages。JPEGImages 文件夹中为蝴蝶生态照片。Annotations 文件夹中为蝴蝶生态照片。其中，蝴蝶标注文件的文件名与蝴蝶生态照片一一对应，且标注文件的格式为 xml 格式。蝴蝶标注文件的主要格式为

```
<size>
  <width>照片宽度</width>
  <height>照片高度</height>
  <depth>照片通道个数</depth>
</size>
<object>
  <name>蝴蝶种名</name>
  <bndbox>
    <xmin>蝴蝶位置左上角x坐标</xmin>
    <ymin>蝴蝶位置左上角y坐标</ymin>
    <xmax>蝴蝶位置右下角x坐标</xmax>
    <ymax>蝴蝶位置右下角y坐标</ymax>
  </bndbox>
</object>
```

其中，蝴蝶位置的坐标系如下图：



3.2 蝴蝶模式照片集

蝴蝶模式照片集，即【第三届中国数据挖掘大赛-蝴蝶模式照片】，包括不同科、属、种、亚种的蝴蝶模式照片，所有照片都已按照标准分类结果存放至相应种类的嵌套文件夹中。其中，照片具体的分类命名标准请参照蝴蝶模式照片文件夹种的【模式照蝴蝶命名 20180226.xlsx】文件。

四、竞赛运行平台环境配置

本次竞赛的测试运行平台包括 Win10 系统以及 Ubuntu 系统,具体配置如下:

4.1 Windows 测试平台环境配置

Windwos 测试平台配置	
OS	Win10 64-bit
Caffe	caffe-windows
Python	Anaconda 2.7
Cuda	8.0
cuDNN	5.1
数据读取目录	“D:\ccdm2018race\”

4.2 Ubuntu 测试平台环境配置

Ubuntu 测试平台配置	
OS	Ubuntu 16.04 64-bit
TensorFlow	1.3.0
Keras	2.0.5
Python	Anaconda 3.6.3
Cuda	8.0
cuDNN	6.0
数据读取目录	“/home/ccdm2018race/”

五、程序输入及输出

5.1 数据输入格式

(1) 在 Windows 平台下, 程序中数据的输入路径为: “D:\ccdm2018race\”;
在 Ubuntu 平台下, 程序中数据的输入路径为: “/home/ccdm2018race/”;

(2) ccdm2018race 文件夹包含了所有测试照片 (照片为 JPG 格式);

(3) 测试照片的命名规则为 IMG_?????.jpg, 其中符号 “?????” 代表六位阿拉伯数字, 读取照片时, 按照阿拉伯数字从小到大的顺序, 例如下图所示 (注意: 阿拉伯数字可能不是连续的)。

5.2 任务结果输出格式

任务一：蝴蝶位置检测

(1) 检测到其中的蝴蝶位置后, 输出矩形位置的左上角坐标和右下角坐标, 坐标值之间用空格隔开 (格式如下: x1 y1 x2 y2)。

(2) 测试照片中只有一只蝴蝶, 没有多个蝴蝶共存的情形。

(2) 依次将每张图片的坐标位置按行写入程序执行目录下的【队伍编号_task1.txt】文件中 (例如编号为【A007】的队伍任务一的结果文件应当命名为【A007_task1.txt】)。

任务二：蝴蝶分类

(1) 检测到生态照片中蝴蝶位置后, 应对该蝴蝶种类分类。具体的, 对蝴蝶的分类应当具体到“种”属性, 格式为: **科名(2 位大写字母)+亚科名(2 位字母)+属名(4 位数字)+种名(3 位数字)**。例如: AAaa0001002 等。

(2) **测试集中蝴蝶的种类已经全部出现在训练集中, 即 100 种左右。**

(3) 对照片中蝴蝶的种类进行分类后, 应按照数据集文件夹中的【模式照蝴蝶命名 20180226.xlsx】文件所规定的命名规则, 依次将每张图片种的蝴蝶种类字符串按行写入程序执行目录下的【队伍编号_task2.txt】文件中 (例如编号为【A007】的队伍任务二的结果文件应当命名为【A007_task2.txt】)。

六、评价指标

关于任务一、任务二的评价指标，以组委会公布的论文中的方法指标为参考，并结合任务一、任务二排名**平均加权**的方式进行最终排名，单项奖排名额外计算。

程序的运行时间不会纳入评价指标，但所有程序平均处理每张照片的时间不应超过 2s。**每张图片平均处理时间超过 2s 的程序不计入最终排名。**

七、竞赛提交

按照本次竞赛例程，参赛队伍需在 2018 年 6 月 1~15 日之间提交本次竞赛相关材料，具体包括以下四个文件夹：**【程序源代码】**、**【设计文档】**、**【可执行程序】**、**【运行文档】**。此处四个文件夹分别作如下说明：

(1) **程序源代码**。包括全部的源代码以及依赖库。

(2) **设计文档**。描述如何处理大赛任务，提供其设计思想和实现细节，并提供主要文件的简洁功能说明。

(3) **可执行程序**。包括可以直接运行的可执行文件（对于 Ubuntu 下基于深度框架的情况，需提供直接运行的脚本命令）。

(4) **运行文档**。简明并准确地描述如何运行提供的可执行程序。



说明：最后提交的结果应当严格按照上述四个文件夹命名方式命名，且上述四个文件夹应当统一压缩为一个压缩文件，且以**【参赛序号.rar】**命名（例如参赛队伍**【A007】**的提交文件应当为**【A007.rar】**）。另外，本次竞赛所提交的程序应当严格遵守以下原则：

- 请严格按照比赛环境配置提交相应程序；
- 若提交的程序无法运行或运行错误，按提交失败处理（建议参赛队伍提交前进行本地测试）。
- 组委会要求，**每支队伍只有一次提交竞赛文档的机会**，请各位参赛队伍切记谨慎、仔细整理所提交的文档。

八、数据集版权

本次竞赛所使用的数据由竞赛组委会提供，竞赛组委会拥有对竞赛数据资源的解释权和使用权。参赛者不得以任何形式公开本次竞赛所提供的数据资源，不得以任何其它非竞赛目的对本次竞赛数据资源加以利用，或者用于谋取商业利益。否则由侵权人向相关权利人承担责任。其他说明请参考《第三届中国数据挖掘竞赛参赛协议》。

竞赛组委会已经对本次竞赛中蝴蝶数据集的分析和研究已经预先发表在 <https://arxiv.org/abs/1803.06626>，供参赛队伍参考。

九、联系方式

请各位参赛队员及时关注本次竞赛的网站以及发布的通知：

<http://ccdm2018.sdufe.edu.cn/sjwjjs.htm>

如有任何本届竞赛相关疑问，请随时联系竞赛委员会。

联系人：吕鹏

联系电话：188-1033-0787

通讯地址：济南市二环东路 7366 号山东财经大学计算机科学与技术学院

邮 箱：ccdm2018race@126.com。

邮 编：250014