多语言识别模型-数据使用说明文档

算法：郭辉铭

时间：2022.7

多语言模型数据集：

数据地址：

https://rrc.cvc.uab.es/?ch=15&com=downloads

注册账号和密码：

iguohm@163.com

123456

数据来源：

ICDAR 2019 Robust Reading Challenge on Multi-lingual scene text detection and recognition

数据构成

多语言文本检测数据集：

训练集

TrainSetImagesTask1_Part1 (3.5G)： ImagesPart1

TrainSetImagesTask1_Part1 (3.5G)： ImagesPart2

label: TrainSetGT (6.5M): train_gt_t13、

测试集

MLT19_TestImagesPart1.zip

MLT19_TestImagesPart2.zip

数据内容

00001 - 01000: Arabic

01001 - 02000: English

02001 - 03000: French

03001 - 04000: Chinese

04001 - 05000: German

05001 - 06000: Korean

06001 - 07000: Japanese

07001 - 08000: Italian

08001 - 09000: Bangla

09001 - 10000: Hindi

多语言文本识别数据集：

为检测图片经过裁剪之后得到。也是按序进行排队得到

法语：word_14981.png--19865

德语： 25413 -- 33908

意大利： 57062--80275

训练集

Word_Images_Part1 (The Ground truth of the word images [2 files] is here too [in the same folder with the images])

Word_Images_Part2

Word_Images_Part3

测试集

MLT19_images_task2.zip