镍基单晶高温合金文献PDF转TXT方法对比

大多数的领域文献都是以PDF格式存储，这种格式对于文献的呈现与传播比较有利，但是PDF作为一种不可编辑的文本，对于研究者直接使用并不友好，所以涌现了许多PDF转化为可编辑文本的方法，包括OCR，pdfminer，pdfplumber等。

镍基单晶高温合金领域文献包含较多的图片、特殊字符、公式等等信息，单纯使用上述说的pdfminer与pdfplumber等方法虽然能够较快识别文字，但是在识别效果上不尽如人意，所以本文主要使用光学字符识别（OCR）方法识别镍基单晶高温合金领域文献，将不可编辑的PDF文本转化为可编辑的TXT文本。

光学字符识别是一个起步比较早的研究领域，专用的OCR库有很多。总体来说，现有的OCR库对于单个字的识别准确率都很高，但是文献PDF转TXT不能只看单个字的准确率，句子、段落的准确率也非常重要。同时OCR本身也是一个机器学习任务，它的速度和效率也十分重要。

本文主要测试了调研找到的四个OCR库（本质是三个）：pytesseract，paddleocr（CPU版本），easyocr和tesserocr。使用这四个python库，对20篇镍基单晶高温合金文献进行识别、清洗，比较他们单篇平均处理时间和处理效果。

文献和测试标准

实验使用的文献是随机选取的镍基单晶高温合金文献，尽可能保证了测试文献涵盖绝大多数的出版商和文献格式模板。

测试标准上，对于识别速度，比较单篇平均识别时间；对于识别效果，由于不太好计算篇幅的准确率，所以采用人工比较检查的方法。

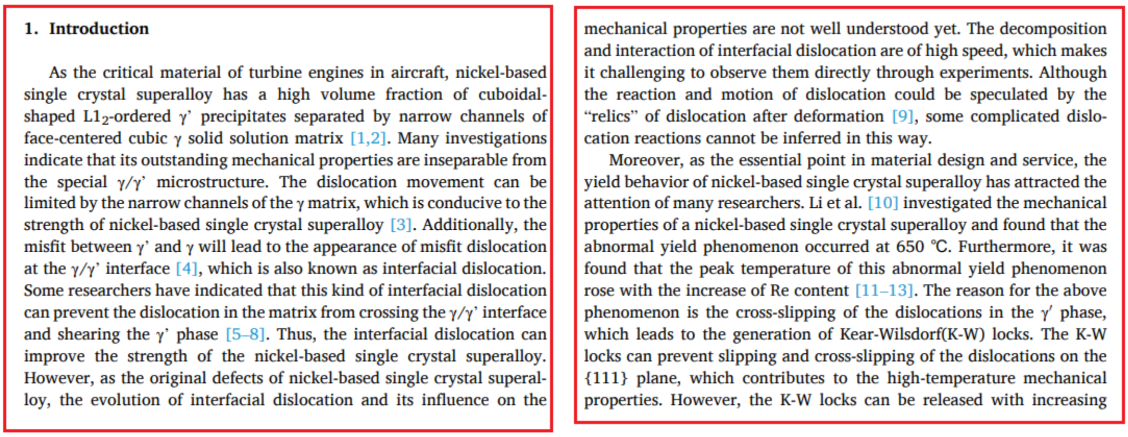
实验

pytesseract与tesserocr

Tesseract是一款由HP实验室开发由Google维护的开源OCR（Optical Character Recognition , 光学字符识别）引擎。这两个库本质都是使用了tesseract，只是调度方法和封装不同导致运行速度不同，他们的最终效果几乎完全相同。在实验过程中，一开始使用了pytesseract库，进行了很多的配置，但是速度较慢，后来是在Stack Overflow上看到对于tesseract的加速，才找到了tesserocr这个库，也使整体的识别速度提升了5倍。

识别速度上，pytesseract平均花费418秒，是四者最长的；tesserocr平局花费75秒，是四者最短的。

识别效果上，由于两者的内核是一样的，所以他们的识别效果几乎完全一样。而且tesseract能够对于较好解决同行两列的现象，准确地从左往右识别。识别效果在良好的清洗后是最佳的。



1. **Introduction**

As the critical material of turbine engines in aircraft, nickel-based single crystal superalloy has a high volume fraction of cuboidal-shaped $L1_2$-ordered γ' precipitates separated by narrow channels of face-centered cubic γ solid solution matrix [1,2]. Many investigations indicate that its outstanding mechanical properties are inseparable from the special γ/γ' microstructure. The dislocation movement can be limited by the narrow channels of the γ matrix, which is conducive to the strength of nickel-based single crystal superalloy [3]. Additionally, the misfit between γ' and γ will lead to the appearance of misfit dislocation at the γ/γ' interface [4], which is also known as interfacial dislocation. Some researchers have indicated that this kind of interfacial dislocation can prevent the dislocation in the matrix from crossing the γ/γ' interface and shearing the γ' phase [5–8]. Thus, the interfacial dislocation can improve the strength of the nickel-based single crystal superalloy. However, as the original defects of nickel-based single crystal superalloy, the evolution of interfacial dislocation and its influence on the mechanical properties are not well understood yet. The decomposition and interaction of interfacial dislocation are of high speed, which makes it challenging to observe them directly through experiments. Although the reaction and motion of dislocation could be speculated by the "relics" of dislocation after deformation [9], some complicated dislocation reactions cannot be inferred in this way.

Moreover, as the essential point in material design and service, the yield behavior of nickel-based single crystal superalloy has attracted the attention of many researchers. Li et al. [10] investigated the mechanical properties of a nickel-based single crystal superalloy and found that the abnormal yield phenomenon occurred at 650 ℃. Furthermore, it was found that the peak temperature of this abnormal yield phenomenon rose with the increase of Re content [11–13]. The reason for the above phenomenon is the cross-slipping of the dislocations in the γ' phase, which leads to the generation of Kear-Wilsdorf(K-W) locks. The K-W locks can prevent slipping and cross-slipping of the dislocations on the {111} plane, which contributes to the high-temperature mechanical properties. However, the K-W locks can be released with increasing

paddleocr

paddleocr是由百度开发的ocr方法，他的识别准确率也很不错，受到很多使用者的青睐。本次实验使用的是飞浆平台的paddleocr（CPU）版本，没有使用GPU版本的原因在于一开始只是希望查看一下效果，CPU和GPU版本只是速度上存在差异，所以只测试了CPU版本。

识别速度上，paddleocr每篇平均花费374秒，排名倒数第二。

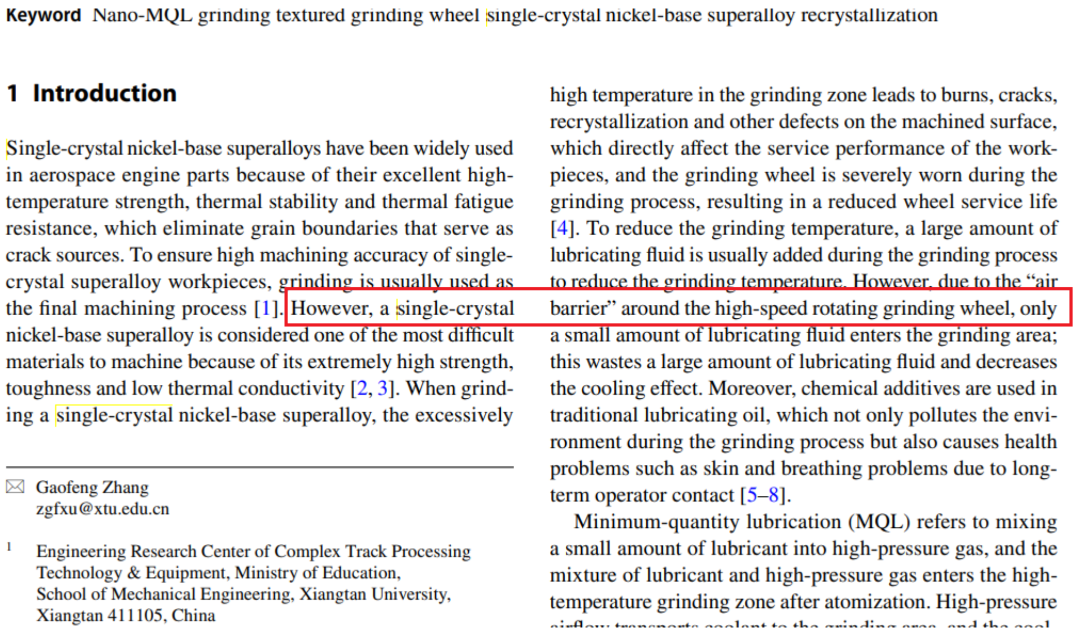
识别效果上，略差于tesseract，主要体现在单字符和跨段落识别上，就镍基单晶高温合金的数据来说，识别错误概率更高，同时会偶发性出现同行两列内容识别错误现象。

easyocr

easyocr是基于pytorch的一个比较高效的ocr库，他的模型较小，识别速度较快，部署较容易。

识别速度上，easyocr略慢与tesserocr，平均花费90秒，速度较快。

识别效果上，单个字体识别并没有因为模型小而质量差，但是却出现了较为严重的同行两列一起识别现象，所以无法采用这一工具。



Keyword Nano-MQL grinding textured grinding wheel single-crystal nickel-base superalloy recrystallization

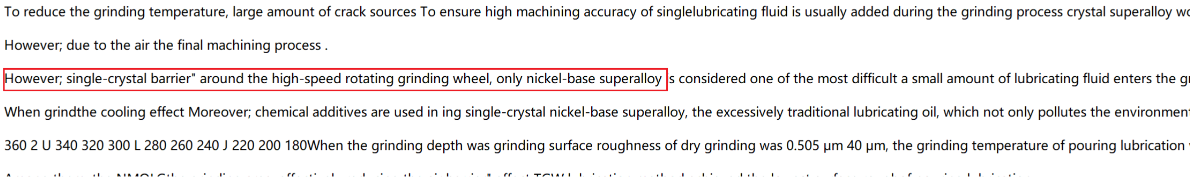
1 Introduction

Single-crystal nickel-base superalloys have been widely used in aerospace engine parts because of their excellent high-temperature strength, thermal stability and thermal fatigue resistance, which eliminate grain boundaries that serve as crack sources. To ensure high machining accuracy of single-crystal superalloy workpieces, grinding is usually used as the final machining process [1]. However, a single-crystal nickel-base superalloy is considered one of the most difficult materials to machine because of its extremely high strength, toughness and low thermal conductivity [2, 3]. When grinding a single-crystal nickel-base superalloy, the excessively high temperature in the grinding zone leads to burns, cracks, recrystallization and other defects on the machined surface, which directly affect the service performance of the workpieces, and the grinding wheel is severely worn during the grinding process, resulting in a reduced wheel service life [4]. To reduce the grinding temperature, a large amount of lubricating fluid is usually added during the grinding process to reduce the grinding temperature. However, due to the "air barrier" around the high-speed rotating grinding wheel, only a small amount of lubricating fluid enters the grinding area; this wastes a large amount of lubricating fluid and decreases the cooling effect. Moreover, chemical additives are used in traditional lubricating oil, which not only pollutes the environment during the grinding process but also causes health problems such as skin and breathing problems due to long-term operator contact [5–8].

Minimum-quantity lubrication (MQL) refers to mixing a small amount of lubricant into high-pressure gas, and the mixture of lubricant and high-pressure gas enters the high-temperature grinding zone after atomization. High-pressure airflow transports coolant to the grinding area, and the cool

✉ Gaofeng Zhang
zgfxu@xtu.edu.cn

[1] Engineering Research Center of Complex Track Processing Technology & Equipment, Ministry of Education, School of Mechanical Engineering, Xiangtan University, Xiangtan 411105, China



To reduce the grinding temperature, large amount of crack sources To ensure high machining accuracy of singlelubricating fluid is usually added during the grinding process crystal superalloy wo

However; due to the air the final machining process .

However; single-crystal barrier" around the high-speed rotating grinding wheel, only nickel-base superalloy is considered one of the most difficult a small amount of lubricating fluid enters the gr

When grindthe cooling effect Moreover; chemical additives are used in ing single-crystal nickel-base superalloy, the excessively traditional lubricating oil, which not only pollutes the environment

360 2 U 340 320 300 L 280 260 240 J 220 200 180When the grinding depth was grinding surface roughness of dry grinding was 0.505 μm 40 μm, the grinding temperature of pouring lubrication

	pytesseract	paddleocr	easyocr	tesserocr
识别速度（秒）	418	374	90	75
识别效果	最佳	一般	同行两列问题	最佳

实验说明

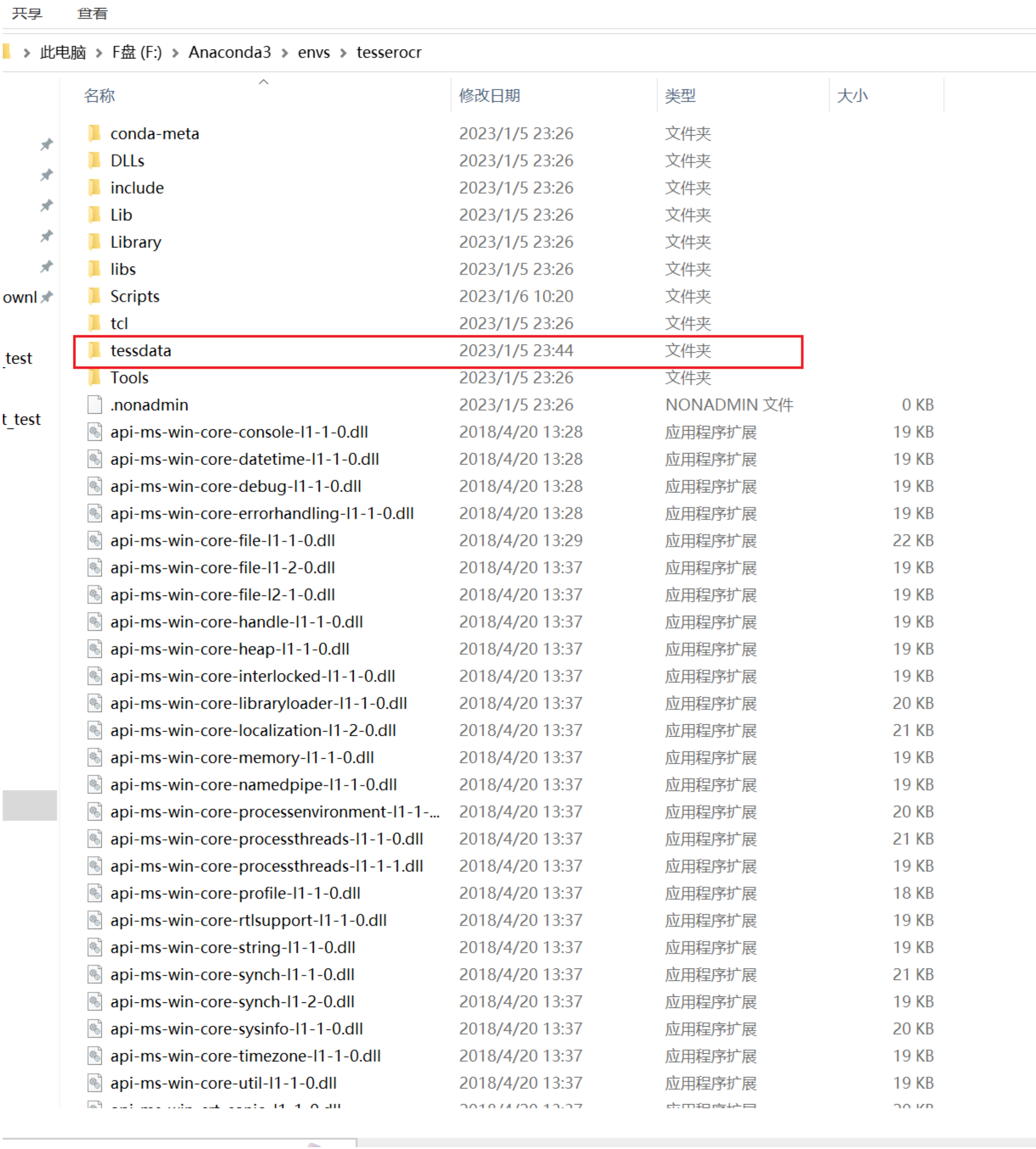
这里所用的ocr方法全部以默认和简单作为策略进行设置，不排除在使用GPU、编写更好的清洗代码等等措施下能够使得现有的结果得到提升。但是目前情况上看tesserocr无论是在速度、效果还是项目需求上都十分契合，因此决定采用tesserocr作为项目PDF2TXT的核心方法。

tesserocr使用说明

环境配置

1. 使用whl方法安装tesserocr库。直接pip可能会出现错误，所以直接登录https://github.com/simonflueckiger/tesserocr-windows_build/releases根据自己python的版本下载对应的whl文件，然后在命令行中进行安装。
2. 使用tesserocr.image_to_text（“path”）报错：运行错误：初始化API失败，可能是无效的tessdata路径。

   如果遇到上述这个问题，是你的tessdata位置存在问题，需要将tessdata文件夹放到你下载的python下，如图



此电脑 > F盘 (F:) > Anaconda3 > envs > tesserocr

名称	修改日期	类型	大小
conda-meta	2023/1/5 23:26	文件夹	
DLLs	2023/1/5 23:26	文件夹	
include	2023/1/5 23:26	文件夹	
Lib	2023/1/5 23:26	文件夹	
Library	2023/1/5 23:26	文件夹	
libs	2023/1/5 23:26	文件夹	
Scripts	2023/1/6 10:20	文件夹	
tcl	2023/1/5 23:26	文件夹	
tessdata	2023/1/5 23:44	文件夹	
Tools	2023/1/5 23:26	文件夹	
.nonadmin	2023/1/5 23:26	NONADMIN 文件	0 KB
api-ms-win-core-console-l1-1-0.dll	2018/4/20 13:28	应用程序扩展	19 KB
api-ms-win-core-datetime-l1-1-0.dll	2018/4/20 13:28	应用程序扩展	19 KB
api-ms-win-core-debug-l1-1-0.dll	2018/4/20 13:28	应用程序扩展	19 KB
api-ms-win-core-errorhandling-l1-1-0.dll	2018/4/20 13:28	应用程序扩展	19 KB
api-ms-win-core-file-l1-1-0.dll	2018/4/20 13:29	应用程序扩展	22 KB
api-ms-win-core-file-l1-2-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-file-l2-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-handle-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-heap-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-interlocked-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-libraryloader-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	20 KB
api-ms-win-core-localization-l1-2-0.dll	2018/4/20 13:37	应用程序扩展	21 KB
api-ms-win-core-memory-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-namedpipe-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-processenvironment-l1-1-...	2018/4/20 13:37	应用程序扩展	20 KB
api-ms-win-core-processthreads-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	21 KB
api-ms-win-core-processthreads-l1-1-1.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-profile-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	18 KB
api-ms-win-core-rtlsupport-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-string-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-synch-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	21 KB
api-ms-win-core-synch-l1-2-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-sysinfo-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	20 KB
api-ms-win-core-timezone-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-core-util-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	19 KB
api-ms-win-crt-conio-l1-1-0.dll	2018/4/20 13:37	应用程序扩展	20 KB

环境配置如有其它问题，可以参考<Tesserocr库安装与使用 - 知乎 (zhihu.com)>

具体使用

OCR方法解决PDF2TXT，一共分为三步：**PDF转化为图片、图片OCR**和**文本清洗。**

tesserocr库在图片OCR这一步上最核心的方法是

```
result = tesserocr.image_to_text(image)
```

详细的内容见具体的程序代码，同时压缩包也会附带tessdata和3.7版本的whl文件（windows）便于快速安装使用tesserocr库。