

知乎

首发于
生物科研

科研干货 | InftyReader: 最强公式识别软件, 写论文必备神器



myhalic

生物科研 课题设计 实验检测

+ 关注他

13 人赞同了该文章

不少科研党就是写论文时经常要输入一堆定理公式, 有的公式是书上或者文献里面的原样公式, 一个一个往Word或者LaTeX里面敲非常累。此时如果能有识别公式的软件, 那肯定会提升不少效率。今天就来推荐一款公式识别软件。【文末附软件安装包】

InftyReader是一款日本人研发的OCR应用软件, 能识别公式和外语文字, 我们主要是用它来识别pdf文章或者图片中的公式。与ABBY不同的是, 它可以将复杂的数学公式识别和翻译成LaTeX, MathML, XHTML, HRTeX, IML和Microsoft Word文档!

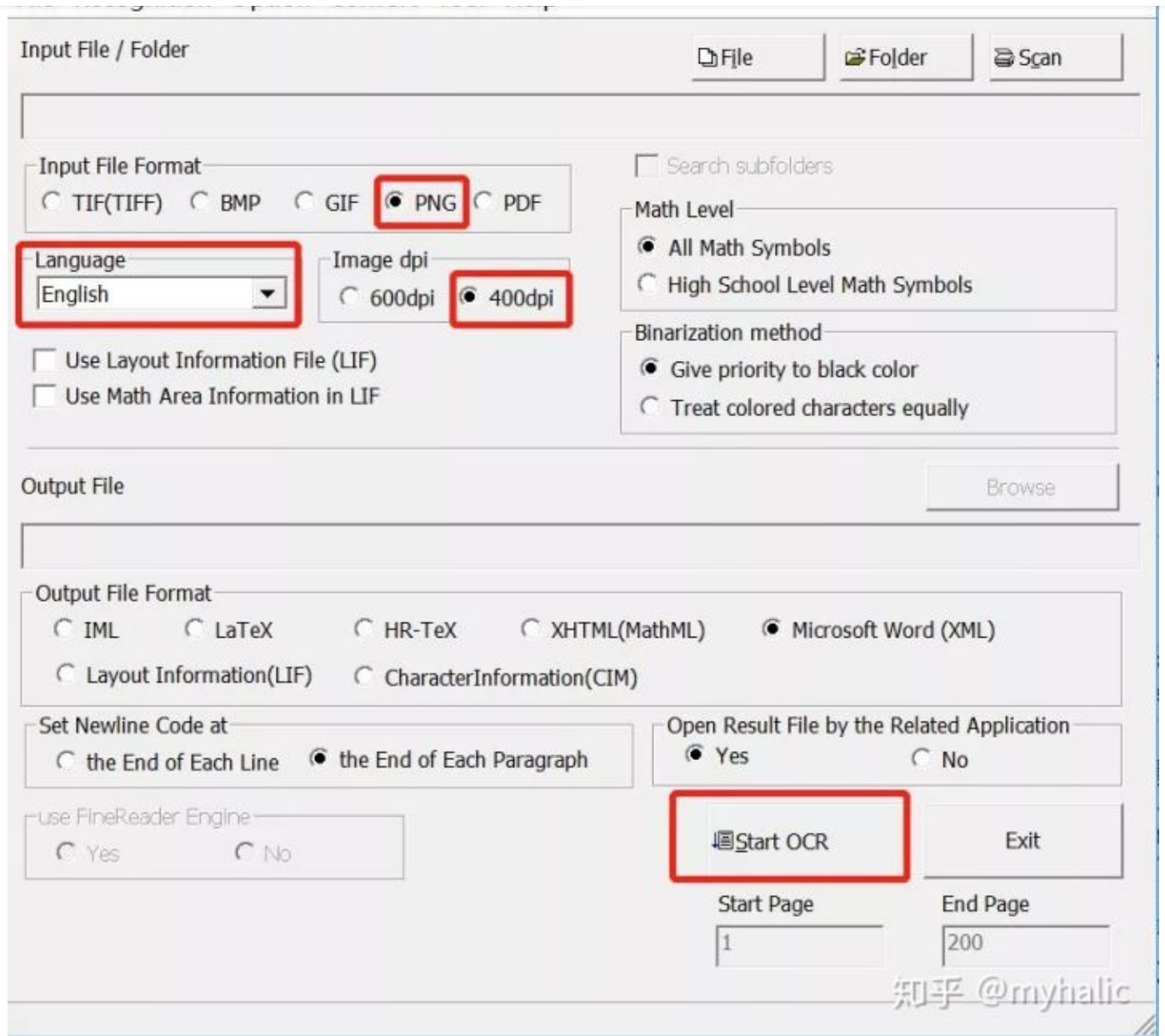
▲ 赞同 13



● 添加评论

➤ 分享

★ 收藏



Input File / Folder

File Folder Scan

Input File Format

☐ TIF(TIFF) ☐ BMP ☐ GIF ☒ PNG ☐ PDF

Language: English

Image dpi: ☐ 600dpi ☒ 400dpi

☐ Search subfolders

Math Level

☒ All Math Symbols ☐ High School Level Math Symbols

Binarization method

☒ Give priority to black color ☐ Treat colored characters equally

☐ Use Layout Information File (LIF)

☐ Use Math Area Information in LIF

Output File

Browse

Output File Format

☐ IML ☐ LaTeX ☐ HR-TeX ☐ XHTML(MathML) ☒ Microsoft Word (XML)

☐ Layout Information(LIF) ☐ CharacterInformation(CIM)

Set Newline Code at

☐ the End of Each Line ☒ the End of Each Paragraph

Open Result File by the Related Application

☒ Yes ☐ No

Use FineReader Engine

☐ Yes ☐ No

Start OCR

Exit

Start Page: 1

End Page: 200

知乎 @myhalic

此软件为全英文操作环境。打开软件后首先找到file按钮后的文档类型选择区域，我们选择pdf，然后再回到file按钮，打开需要识别的文档。在output file format 选择latex，之后找到start ocr回车即可。

下面是一些说明：

- 1.这个软件能识别英文、日文文献，不能识别中文文献。
- 2.软件识别还是会有一定错误，需要一些经验进行判断。
- 3.本软件对于文字版pdf识别效果最好，或者需要清晰的扫描版。
- 4.jpg图像等亦可识别，但同样要求尽量高的清晰度。

▲ 赞同 13 ▼

● 添加评论

➤ 分享

★ 收藏

知乎

首发于
生物科研

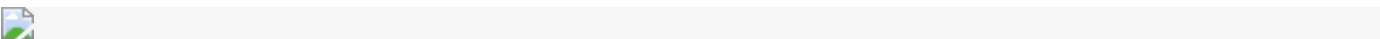
6.软件原理是PDF文件转换成图片，然后针对图片进行识别，所以速度会很慢，最后生成XML用word打开。

第一步：将pdf的论文转换成png格式。

打开PDFtoPNG，如下图所示，注意dpi修改为400。



第二步：打开InftyReader，选择试用，注意试用版一天只能解析5张图片的公式。



第三步：打开软件后，如下界面，按照红框操作即可。要点是选择PNG、英语、400dpi、输出word格式，最后点击Start OCR即可，注意试用版一次只能解析一张图片。也可以用PDF格式直接导入，就是比较慢。



▲ 赞同 13

▼

● 添加评论

➦ 分享

★ 收藏

原图部分如下:

After these settings, two classes of conventional SMC laws to achieve the objective of the paper have the form [10]

$$\mathbf{u} = Y\bar{\mathbf{z}} - [k_1 \operatorname{sgn}(s_1), \dots, k_n \operatorname{sgn}(s_n)]^T \quad (10)$$

$$\bar{z}_i = -\bar{a}_i \cdot \operatorname{sgn}\left(\sum_{j=1}^n s_j Y_{ji}\right) \quad i = 1, \dots, m \quad (11)$$

$$k_i = \sum_{j=1}^n F_{ij}(q, \dot{q}) |s_j| + \rho_i + \delta_i \quad i = 1, \dots, n \quad (12)$$

or

$$\mathbf{u} = Y\bar{\mathbf{z}} - [k_1 \operatorname{sat}(s_1 / \phi_1), \dots, k_n \operatorname{sat}(s_n / \phi_n)]^T \quad (13)$$

$$\bar{z}_i = -\bar{a}_i \cdot \operatorname{sat}\left(\sum_{j=1}^n s_j Y_{ji}\right) \quad i = 1, \dots, m \quad (14)$$

$$k_i = \sum_{j=1}^n F_{ij}(q, \dot{q}) |s_{\Delta j}| + \rho_i + \delta_i \quad i = 1, \dots, n \quad (15)$$

$$s_{\Delta j} = s_j - \phi_j \cdot \operatorname{sat}(s_j / \phi_j) \quad j = 1, \dots, n \quad (16)$$

Here, $\operatorname{sgn}(\cdot)$ and $\operatorname{sat}(\cdot)$ denote the sign and the saturation functions, respectively. For $i = 1, \dots, m$ and $j = 1, \dots, n$, \bar{a}_i, ϕ_j , and δ_j denote the estimation of upper bound of $|a_i|$, the boundary layer widths of the saturation-type control and the selected positive constants, respectively, while F_{ij} , $1 \leq i, j \leq n$, are chosen to satisfy

$$|C_{ij}(\mathbf{q}, \dot{\mathbf{q}})| < F_{ij}(\mathbf{q}, \dot{\mathbf{q}}). \quad (17)$$

知乎 @nyhalic

识别如下, 可见识别公式正确率很高, 且公式可编辑。

and $|d_i(t\lambda \leq \rho_i(q, \dot{q}, f))|$ (9)

for all $i = 1, \dots, n$, where $(\cdot)_i$ denote the i th entry of a vector. After these settings, two classes of conventional SMC laws to achieve the objective of the paper have the form [10]

$$u = Y\bar{z} - [k_1 \text{sgn}(s_1), \dots, k_n \text{sgn}(s_n)]^T \quad (10)$$

$$\bar{z} = -\bar{a} \cdot \text{sgn}(\sum_{j=1}^n s_j Y_{ji}) \quad i = 1, \dots, m \quad (11)$$

$$k_j = \sum_{i=1}^n F_{ij} (|C_{ij}(q, \dot{q})| + \rho_j + \delta_j) \quad i = 1, \dots, n \quad (12)$$

or

$$u = Y\bar{z} - [k_1 \text{sat}(s_1/h), \dots, k_n \text{sat}(s_n/l\phi_n)]^T \quad (13)$$

$$\bar{z} = -\bar{a} \cdot \text{sat}(\sum_{j=1}^n s_j Y_{ji}) \quad i = 1, \dots, m \quad (14)$$

$$k_j = \sum_{i=1}^n F_{ij} (|C_{ij}(q, \dot{q})| + \rho_j + \delta_j) \quad i = 1, \dots, n \quad (15)$$

$$s_{\Delta j} = s_j - \phi_j \cdot \text{sat}(s_j/l\phi_j) \quad j = 1, \dots, n \quad (16)$$

Here, sgn and sat denote the sign and the saturation functions, respectively. For $i = 1, \dots, m$ and $j = 1, \dots, n$, \bar{a}_i , ϕ_j , and δ_j denote the estimation of upper bound of $|a_i|$, the boundary layer widths of the saturation-type control and the selected positive constants, respectively, while F_{ij} , $1 \leq i, j \leq n$, are chosen to satisfy

$$|C_{ij}(q, \dot{q})| < F_{ij}(q, \dot{q}) \quad (17)$$

Although possibly achieving the desired performance, the two classes of conventional SMC laws have their drawbacks. First, the discontinuity of sign-type controls as given by (10) – (12) leads to chattering, which is generally undesirable in practical applications since it involves extremely high control activity and might further excite high-frequency dynamics neglected in the course of modeling. Second, although a continuous one, the saturation-type controllers as given by (13) – (16) can only ensure that the system state enters the boundary layer but does not provide asymptotic stability in the presence of disturbances [16]. Instead of sign- and saturation-type controllers as given by (10) – (12) and (13) – (16), in the following we present a

官方下载地址: inftyreader.org/InftyRe...

注:

转PNG也可以用其他软件。

如果只想识别一页PDF上的公式, 可以先用PDF阅读器先提取出那一页PDF, 然后用InftyReader将那一页PDF转换成PNG。

单个公式的截图转换会很快, 正确率基本上在90%左右。

不要为了简单直接截图, 因为dpi不合适, 会导致识别效果很差! 如果有耐心的话, 还是建议直接PDF识别。

还可以尝试用国产的MathOCR, 一款开源的.jar软件。

来源: 虫部落/万能搜吧

▲ 赞同 13

▼

● 添加评论

➦ 分享

★ 收藏

知乎



首发于
生物科研

发布于 2019-02-22

论文

生物

文章被以下专栏收录



生物科研
生物科研实验及文章撰写经验分享

关注专栏

推荐阅读

你的屏幕脏了，我帮你擦干净

苏豪同学



零基础做科研，4招搞定，让导师好好过节

myhal...

发表于生物科研

还没有评论

写下你的评论...



▲ 赞同 13



💬 添加评论

➦ 分享

★ 收藏