ABBYY FineReader Engine性能指南

整合光学字符识别（OCR）技术将有效扩展应用程序的功能性。

OCR组件的出色性能是提高客户满意度的关键因素之一。

本文档提供关于影响常规OCR性能因素的信息，以及在软件开发套件ABBYY FineReader Engine中对其进行优化的可能性。通过利用ABBYY FineReader Engine先进的功能和选项，可以进一步提升OCR的高性能，从而获得最佳的用户体验。在评测OCR性能时，要考虑两个主要参数：

识别精准度

处理速度

哪些因素会影响OCR精准度和处理速度?

在ABBYY FineReader Engine中使用正确的参数，可以显著提高识别速度和识别精准度。

图像类型和图像质量

图像有各种不同的来源。由数字化方式创建的PDF、计算机和平板电脑设备的屏幕截图，通过扫描仪、传真服务器、数码相机或智能手机创建的图像文件—各种图像来源使不同的图像类型具有不同的质量。例如，使用错误的扫描仪设置可能会在图像上产生“噪点”、随机的黑点或斑点、模糊不均匀的字母、歪斜的线条和偏移的表格边框。对于OCR而言，这是“低质量图像”。

处理低质量图像对计算能力有较高的要求，增加了整体处理时间，还会使识别结果变差。

图像质量 = 影响OCR性能的关键因素

另一方面，在不失真的情况下处理“高质量图像”会减少处理时间。此外，读取高质量图像会使识别结果有较高的精准度。

因此，建议在OCR处理中使用高质量的图像。

通过提升图像质量来提高OCR速度和准确率。

如何获得高质量的图像?

文件扫描技巧

字体大小

如果打印文档使用的字体非常小，应该使用较高的分辨率进行扫描。

使用以下分辨率进行扫描：

- 使用300 dpi扫描典型文本（以10 pt或更大的字体打印）
- 使用400-600 dpi扫描文本（字体大小为9 pt或更小）

打印质量

质量较差的文档（例如旧报纸或书籍）应在灰度模式下进行扫描。此模式会保留扫描文本中有关字母的更多信息。

拍照技巧

合适的光线

- 确保光线均匀分布于页面，且没有暗区或阴影。
- 如有可能，请使用三脚架。将镜头位置调整到与文件页面平行，并将镜头对准文本中心区域。
- 关闭闪光灯以避免页面上出现光晕和阴影。
- 如果相机具有“白平衡”选项，请使用一张白纸对白平衡进行校准。否则，请选择最适合当前光线条件的白平衡模式。

如果光线不足……

选择更大的光圈值

- 选择更大的感光度ISO值
- 如相机不能自动对焦，则使用手动对焦

有关更多详细信息，请参阅FineReader Engine的开发人员帮助手册：

导览→最佳范例→提高处理速度

使用FineReader Engine提升图像质量

如果无法预先改变图像质量，建议在识别之前对图像质量进行提升。在FineReader Engine中，可以使用各种强大的图像预处理功能：

基本图像处理功能

图像缩放
图像裁剪
图像裁切
文本行校直
镜像和反转
去噪
局部对比增强
几何变形校正
自适应二值化

高级图像处理功能

ABBYY相机OCR技术
自动分页
去除图章和书面笔记
图像自动歪斜校正
自动侦测页面方向和旋转
图像降斑

FineReader Engine中的预定义处理配置文件

可以通过微调各个设置来优化识别精准度和处理速度。但是，预定义处理配置文件用于主要使用场景，可以通过应用其中之一来完成此操作。这些配置文件中提供的设置是最适合相应情况的。

大多数配置文件有两种形式：

- 优化设置以获得最佳识别精准度
- 优化设置以实现最高处理速度

如有必要，仍然可以通过API设置其他相关参数来更精确地调整配置文件。

场景	配置文件名称
文档转换归档 文件转换归档对文档的处理速度和视觉质量提出了很高的要求，而且所生成PDF要小。关键参数不是识别精准度最高，达到满意的水平就足够了。	DocumentArchiving_Accuracy DocumentArchiving_Speed BookArchiving_Accuracy BookArchiving_Speed
文档转换以重复使用内容 在这种情况下，最理想的标准是识别精准度和保留文档结构。由于最终文档必须正确无误，因此较低的识别率或错误的布局重建将意味着操作员需要进行额外的工作。	DocumentConversion_Accuracy DocumentConversion_Speed
图像上的文本提取，包括低质量的小文本区域。此文本信息应该可供以后搜索、提取以进行进一步处理或用于文档分类。	TextExtraction_Accuracy TextExtraction_Speed
从确切指定的页面区域捕获数据	FieldLevelRecognition
条码识别	BarcodeRecognition_Accuracy BarcodeRecognition_Speed
创建高压缩PDF文件，另存为图片	HighCompressedImageOnlyPdf
名片识别	BusinessCardsProcessing
技术图纸识别，这些图纸通常尺寸比较大，包括复杂的图表以及不同方向的文字	EngineeringDrawingsProcessing

预定义的识别模式

影响OCR性能的另一种可能性是使用为特定场景设计的识别模式。FineReader Engine提供以下预定义的识别模式：

正常

使用此模式，您将获得最高的识别精准度。

当识别内容将重新用于其他应用程序或任务中，而且高精度对这些应用程序或任务来说至关重要的情况下，强烈建议使用此模式。

平衡

此模式的识别准确性和速度适中。

通常，此识别模式比“正常”识别模式速度高，同时可达到几乎相同的准确性。

快速

使用此模式可以将处理速度提高到200-250%。

当处理速度至关重要时，例如在用于大容量文档处理归档、内容和文档管理系统中，建议使用此模式。

文件语言

ABBYY FineReader Engine能够识别单语和多语（例如用多种语言编写的）文档。指定正确的识别语言非常重要，因为错误指定的语言会大大降低文档处理速度并降低识别质量。

如果无法预先指定识别语言，则可以使用自动语言检测。

但是，预选很多识别语言将降低处理速度。因此，建议不要指定五种以上的识别语言。

为了进一步提高识别精准度，FineReader Engine提供了对多种语言的字典和词法支持。在处理包含特定主题术语或“结构”（例如产品代码、电话号码或护照号码）的文档时，可以导入自定义创建的词典以确保较高的识别质量。

文件处理阶段

就OCR而言，文档处理是一个多步骤过程。根据特定情况，每个OCR阶段应用不同的功能和参数。在FineReader Engine中，该过程包括以下阶段：

图像导入
将图像文档提供给FineReader Engine来进行处理。

图像预处理
在实际处理之前提高图像质量。

文档分析
检测单个对象，例如文本、图像或条形码。

识别
“读取”提取机器打印和手工打印的文本以及光学标记和条形码的值。

合成
重构原始文档的布局，包括页面顺序和编号、逻辑文本流、图像等。

导出
最终按要求的格式导出处理结果。

图像导入

可以直接从扫描仪发送文件到 FineReader Engine，也可以从存储系统或内存流中导入文件。从不同来源获取图像需要使用不同的方法，而且会影响识别速度。从内存中导入图像通常比从文件存储器中打开图像更快。

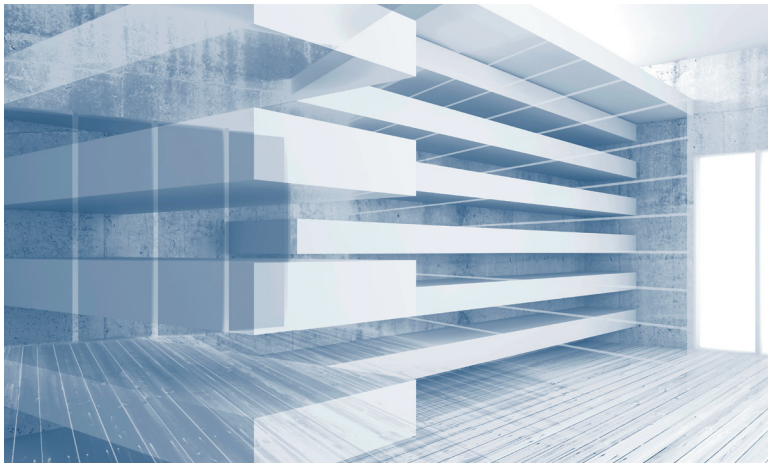
文档分析

小册子或报纸通常包含在列、表格、图表和图片中的文字。技术图纸可能是大型文档，包括具有不同文本方向的复杂工程图。对于布局如此复杂的文档，文档分析步骤将需要更多的处理时间。另一方面，布局简单的文档（如信件或合同）的分析非常快速。

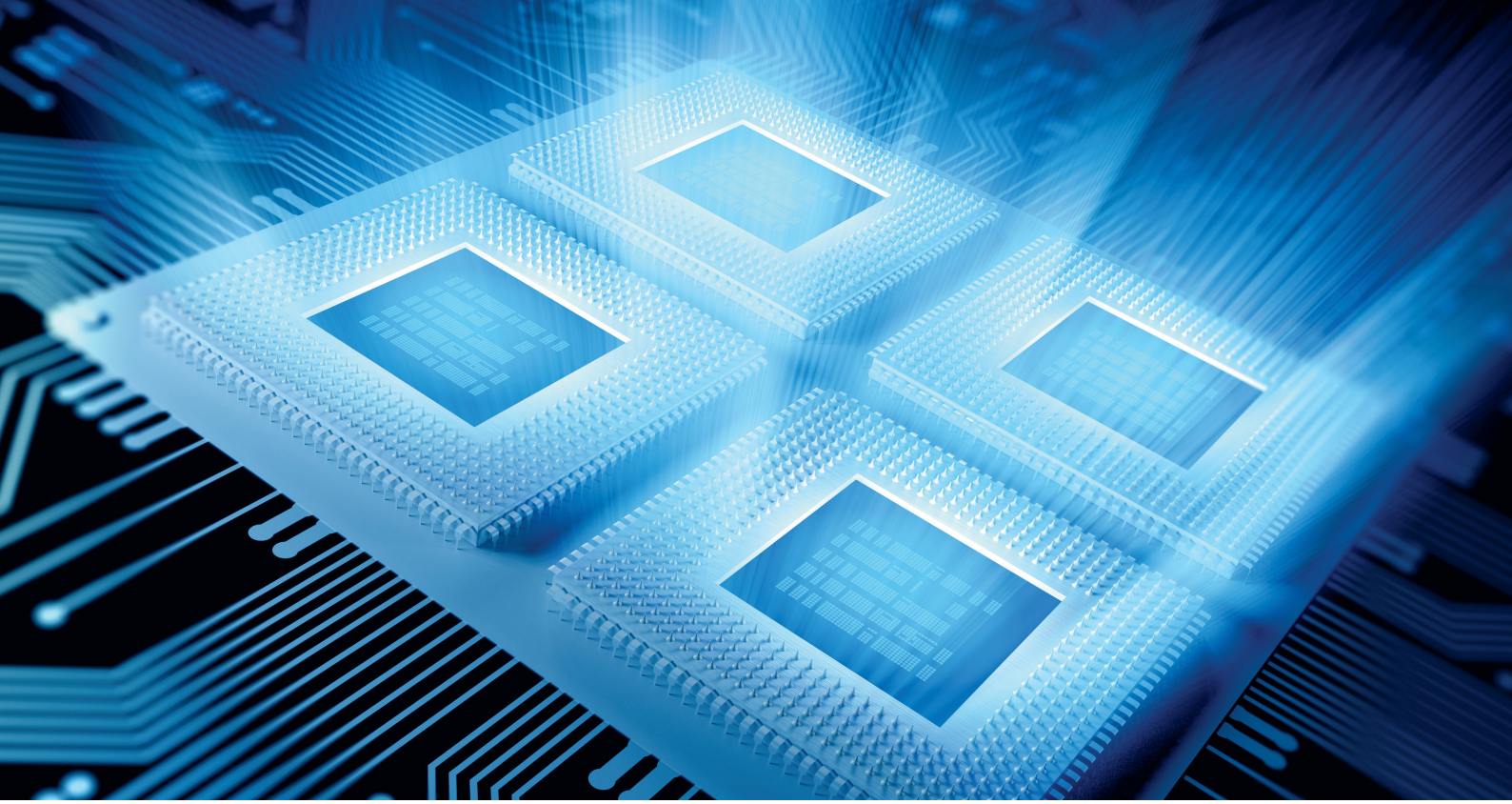
图像预处理

通常，对于高质量图像，OCR处理更快。建议相应地微调图像预处理步骤，从而节省实际处理步骤的时间。

图像可以具有不同的格式和质量。高质量图像（例如由数字化方式创建的PDF）通常不需要大量的预处理工作。对于低质量图像，例如扫描仪设置不正确的扫描文档或旧书，有必要应用高级图像预处理功能以改善识别结果。对于数码照片的预处理，则采用了专用的ABBYY相机OCR™技术。在这里，这些算法专门针对照片增强进行了优化。使用不同的预处理功能将分别影响处理速度。



不同方法和参数，用于特定的处理场景，将显著影响整体处理速度。去掉不必要的阶段可以加快整个OCR过程。例如，从预定义的文档区域提取数据时，不需要进行文档分析。将文档导出为TXT格式或“仅PDF图像”格式时，可以跳过合成阶段。



使用多核并行处理

FineReader Engine可用于构建各种规模和复杂性的应用程序—从客户端工作站到基于服务器的解决方案或数以百万计的大型项目。 涉及到OCR性能时，利用多处理器或多核系统来提高处理速度是非常有意义的。

FineReader Engine中内置支持多核，允许采用不同的方法来扩大OCR进程：

- 使用单个Engine实例
- 加载多个Engine实例

处理文档的方法有多种：

处理大型多页文档

对于具有多个页面的大型文档的并行处理，最适合使用“FRDocument”对象。在这种情况下，多页文档的页面将在可用的CPU上并行处理。最后，将结果合并到一个多页文档中。这是最易于编码的多处理方式。根据所需的因素（例如，可用的物理或逻辑CPU内核数，许可证中可用的CPU内核数以及文档的页数），将自动检测所需的进程数。如有必要，开发人员可以轻松更改多处理设置并调整要运行的进程数。

处理多个单页文档

要并行处理许多从同一来源接收的单页文档，例如扫描仪，建议使用“BatchProcessor”对象。当不需要导出文件时，例如在具有自定义输出格式的数据捕获方案中，此对象在速度方面最有效。

要并行执行许多单页文档的完整处理，建议使用Engine实例池。当输入文档应在提交后直接处理时，此方法也最适合于Web服务场景。在这种情况下，文档将从池中传递到可用的FineReader Engine实例，并立即进行处理。

ABBYY FineReader Engine提供了灵活的模块化许可证，可以选择最佳的选项组合。
例如，网络许可证允许将处理扩展到大量计算机（虚拟或物理）。

FineReader Engine—速度测试结果

下表列出了内部性能测试的结果。

请注意，测试结果始终取决于许多因素，例如图像质量、使用的识别语言和其他因素。

	单页文档；页/分钟	一个多页文档；页/分钟	数据捕获方案（不导出*）；页/分钟
顺序处理	60	51	87
FRDocument的并行处理	41	117	57
使用FRDocument并行处理，同时在处理期间将数据保留在内存中	55	141	82
使用批处理器进行并行处理	99	115	294
使用引擎池进行并行处理	165	115	102

技术测试信息

英特尔®酷睿™i5-4440（3.10 GHz，4个物理内核），8 GB RAM，4个进程同时运行。

使用“DocumentArchiving_Speed”预定义的配置文件对300个英文文档进行了性能测试。在“单页文档”和“一个多页文档”场景中，文档以PDF格式导出。

*文本是从单页文档的预定义区域中提取的。没有执行导出为任何文件格式的操作。

系统资源

在OCR过程中，将应用一系列不同的算法。这些算法取决于图像质量、文档语言、布局复杂度度和文档的页数。因此，这些算法可能需要更高的存储器资源。建议根据概述的内存需求设置系统，通过分配足够的系统内存来优化处理速度。

内存需求

处理单页文档：
至少400 MB RAM
建议1 GB RAM

处理多页文档：
至少1 GB RAM
建议1,5 GB RAM

并行处理：
350 MB RAM x CPU内核数
+额外的450 MB RAM

并行处理阿拉伯文、中文、日文或韩文的文档：
850 MB x CPU内核数
+ 750MB RAM

如何提高FineReader Engine的整体处理速度

有很多种方法可以改善系统性能：

微调图像预处理设置，为处理步骤提供最高的文档质量。

- 在处理步骤中，使用优化速度的预定义处理配置文件之一和合适的识别模式-平衡或快速。
- 指定正确的识别语言。语言不正确会大大降低文档处理速度。选择的识别语言越多，处理速度就越慢。
- 使用适当的对象（FRDocument或BatchProcessor）并启用并行处理。
- 指定适当的分析和识别参数。例如，如果图像不包含表格并且页面方向正确，则禁用表格检测和页面方向校正。
- 如果已处理的文档将导出为TXT格式或“仅PDF图像”格式，则省略合成阶段。
- 将文档导出为PDF格式时，请使用快速PDF导出配置文件。
- 使用特殊对象（ExportFileWriter），该对象适用于将非常大的多页文档导出为PDF格式。
- 有关更多信息，请参阅FineReader Engine的开发人员帮助手册：

导览→最佳范例→提高处理速度

如何提高FineReader Engine的文本识别质量

FineReader Engine提供高识别质量。识别质量始终取决于诸如图像质量、语言和其他因素。但是，有多种方法可以提高识别质量：

- 指定正确的文本类型。
- 指定适当的识别语言。
- 定义专用的语言和自定义词典，以识别特殊字符或具有特定术语的文档，例如法律或医疗保健文本。
- 将已扫描书籍的相对页面分成两个单独的图像。
- 处理数码照片时，请应用专用的相机OCR技术。
- 如果图像的分辨率与建议的分辨率有很大差异，请纠正图像的分辨率。

有关更多信息，请参阅FineReader Engine的开发人员帮助手册：

导览→最佳范例→提高识别质量

如您所知，在文档处理中，OCR过程是一项非常复杂的任务。根据各个文档处理方案的不同，OCR性能结果可能会有很大差异。ABBYY FineReader Engine中有关识别准确性和处理速度优化的提示可以帮助到您，让您的业务案例获得最佳效果。

其他信息资源

要了解有关OCR性能优化的各个方面以及SDKABBYYFineReaderEngine的更多信息，请使用以下信息源：

- ABBYY技术门户： https://abbyy.technology/
- ABBYY OCR SDK论坛： **https://forum.ocrsdk.com/**
- ABBYY FineReader Engine分布版中提供的帮助文件
- ABBYY FineReader Engine在 www.abbyy.cn/ocr-sdk上面的产品页面。

个人项目支持

如果您想讨论一个特定的项目，请与我们联系。在测试期间，您可以向我们的ABBYY技术支持提出标准技术问题，或使用ABBYY专业服务来进行高级咨询、深入的项目分析、以及个人代码评审。使用ABBYY的技术资源可以缩短开发工作并加快项目进度。 如果您想进一步探索这些选项，请联系ABBYY销售经理。

ABBYY®

如果您还有其他疑问，请联系**www.abbyy.cn/contacts** 下方列出的当地ABBYY客户代表，
或使用在线联系表**www.abbyy.cn/ocr-sdk/#request-demo。**

This software includes ABBYY® FineReader® Engine 12 recognition technologies. © 2017, ABBYY Production LLC. ABBYY, FINEREADER and ABBYY FineReader are either registered trademarks or trademarks of ABBYY Software Ltd. All product names, trademarks and registered trademarks are property of their respective owners. Windows® is a registered trademark of Microsoft Corporation in the United States and other countries. The registered trademark Linux® is used pursuant to a sublicense from LMI, the exclusive licensee of Linus Torvalds, owner of the mark on a world-wide basis. Mac® and OS X® are trademarks of Apple Inc., registered in the U.S. and other countries. Datalogics®, The DL Logo®, PDF2IMG™ and DLE™ are trademarks of Datalogics, Inc. Adobe®, The Adobe Logo®, Adobe® PDF Library™, Powered by Adobe PDF Library logo, Reader® are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States and/or other countries. Portions of this computer program are copyright © 1996-2007 LizardTech, Inc. The software is based in part on the work of the Independent JPEG Group. Portions of this software are copyright ©2011 University of New South Wales. Unicode support: © 1991-2013 Unicode, Inc. Intel® Performance Primitives: Copyright © 2002-2008 Intel Corporation. Portions of this software are copyright © 1996-2002, 2006 The FreeType Project www.freetype.org. WIBU, CodeMeter, SmartShelter, and SmartBind are registered trademarks of Wibu-Systems. All rights reserved. All other trademarks are the property of their respective owners. #11350