Performing-OCR-using-Tesseract

Performing OCR on Chinese, Japanese, Korean, Greek, Arabic and Hebrew documents using SolidFramework

Author: Roger Dunham

Date: 19th March 2019

Updated: 17th December 2024

Version 1.7

Introduction

Our Solid OCR engine is being actively developed to continually improve accuracy and performance. Solid OCR is, however, capable of recognizing Latin and Cyrillic scripts only.

For other scripts, SolidFramework uses the open-source Tesseract library to perform OCR for Chinese, Japanese, Korean, Greek and Hebrew language documents. While we have committed to make Solid Framework work reliably with the Tesseract library for this set of languages, we do not do Tesseract development (i.e. no accuracy or performance improvements).

SolidFramework License Requirements

Tesseract OCR requires either a **Developer** or a **Pro with OCR** SolidFramework license.

Required files

Tesseract.dll is already included as part of the regular SolidFramework download.

“Trained Data” files are also needed. Due to their enormous size, these are not included in the standard download, but can be downloaded from the links in the table below.

Download the file (or files) that you require and place it into a folder called “tessdata” that the program that you are developing can access.

It is essential that the English traineddata file is also downloaded, since that file is used to recognise numbers and letters from the Latin alphabet that are widely used even in non-English documents.

Language	code	Link to TrainedData file
Complete set of files included in this table		https://downloads.soliddocuments.com/solidframework/tessdata/traineddata.zip
English	en	https://downloads.soliddocuments.com/solidframework/tessdata/eng.traineddata
Simplified Chinese	zh	https://downloads.soliddocuments.com/solidframework/tessdata/chi_sim.traineddata https://downloads.soliddocuments.com/solidframework/tessdata/chi_sim_vert.traineddata

Traditional Chinese	zt	https://downloads.soliddocuments.com/solidframework/tessdata/chi_tra.traineddata https://downloads.soliddocuments.com/solidframework/tessdata/chi_tra_vert.traineddata
Korean	ko	https://downloads.soliddocuments.com/solidframework/tessdata/kor.traineddata https://downloads.soliddocuments.com/solidframework/tessdata/kor_vert.traineddata
Japanese	ja	https://downloads.soliddocuments.com/solidframework/tessdata/jpn.traineddata https://downloads.soliddocuments.com/solidframework/tessdata/jpn_vert.traineddata
Greek	el	https://downloads.soliddocuments.com/solidframework/tessdata/ell.traineddata
Hewbrew	he	https://downloads.soliddocuments.com/solidframework/tessdata/heb.traineddata
Arabic	ara	https://downloads.soliddocuments.com/solidframework/tessdata/ara.traineddata
Script and Language detection traineddata is used for language detection. (should be used with any language listed above)		https://downloads.soliddocuments.com/solidframework/tessdata/osd.traineddata

Specifying the location of the TessData Folder

The traineddata files need to be available when SolidFramework initializes.

They must be located in a folder called “tessdata”.

By default, SolidFramework will look for the “tessdata” folder as a subfolder of the location where the program is executed.

Alternatively, you can specify a different location for this folder as follows:

```
SolidFramework.Imaging.Ocr.TesseractDataDirectoryLocation = [tessdataDirectoryLocation];
```

Where [tessdataDirectoryLocation] is the location of the "tessdata" subfolder that contains the downloaded files.

Checking that the Tesseract folder has been correctly found

You can use the following code:

```
var langs = SolidFramework.Imaging.Ocr.Languages;
```

and check that your Tesseract language is included in the list.

Sample code

The following code can be used to convert the PDF [sourceFilename] into the Word Document [outputFileName] using Simplified Chinese OCR.

```
//Specify the folder where the tesseract data is located.
SolidFramework.Imaging.Ocr.TesseractDataDirectoryLocation = tessdataLocation;
using (PdfToWordConverter conv = new PdfToWordConverter())
{
//Set the Text recovery language. Tesseract options are zh, zt, ja, ko and el.
conv.TextRecoveryLanguage = "zh";

//Specify the source and destination file.
conv.AddSourceFile(sourceFilename);
var result = conv.ConvertTo(outputFilename);

//Check that the conversion completed successfully, or record what the problem was.
System.Diagnostics.Debug.WriteLine("Conversion Result is: " + result);
}
```