

OCR 21 languages with MODI

Stronghorse_mj@hotmail.com

About how to install and call MODI to OCR 21 languages.

Introduction

Benefits of MODI (Microsoft Office Document Imaging):

- a) It's free. I'll explain this below.
- b) It's better than other free OCR engine, such as gocr, especially for CJK (Chinese, Japanese, Korean).
- c) It's well documented. You can download official document from here:
<http://www.microsoft.com/en-us/download/details.aspx?id=22333>
- d) It can OCR 21 languages as good as commercial OCR software. In fact, the low level DLLs are really come from commercial companies.

Limits of MODI:

- a) It is only included in Office 2003 and 2007. Later I'll explain how to install it with other version of Office.
- b) The source to OCR MUST be a TIFF file in disk, can't be a buffer in memory. So temp file is often be used.
- c) The interface (DLLs) is 32-bit, not 64-bit. If you really want to call it in 64-bit application, you may need to build a bridge. For example: Make a stand-alone application as OCR service, receive TIFF path and response OCR result by WM_COPYDATA message.
- d) The application that call MODI interface (create MODI object) MUST be run as Administrator in Win7/10. Otherwise you can OCR nothing.
- e) For CJK, if the result is less than 8 characters, the OCR engine will report that nothing can be recognized. In this case, duplicating source image to get a longer result is a practical solution.
- f) For Office 2007, SP1/SP2/SP3 is needed, especially for CJK. Otherwise OCR engine may be crashed.

Getting Started

For VC and C# programmers, "How to use MODI" is not a problem since you can find so many examples, such as:

<https://www.codeproject.com/Articles/17291/OCR-With-MODI-in-Visual-C>

<https://www.codeproject.com/Tips/1028098/Capture-a-Text-Image-on-the-Tracing-Layer-and-Pass>

<https://www.codeproject.com/Articles/10130/OCR-with-Microsoft-Office>

<https://www.codeproject.com/Articles/10654/The-Paperless-Desktop>

<https://www.codeproject.com/Articles/10206/Document-Processing-Part-II-Request-Driven-OCR>

<https://www.codeproject.com/Articles/29172/Converting-Images-to-Text-using-Office-2007->

OCR-Op

Since you can find tons of source code in these examples, I won't provide any more in here.

For VB6 programmers, things are more difficult:

- a) If your VB source code call MODI of Office 2007, and you run it in IDE, then you only can run it for once. In second time, it will report a NULL result or even IDE crashed. But if you call MODI of Office 2003, you can run it as many times as you want.
- b) If you compile your source code which call MODI 2007, the compiled EXE can be run unlimited times. So maybe the reason is that MODI 2007 is not compatible with VB6 IDE.
- c) If you directly create MODI object in VB source code, you must say you create it from MODI 2007 or 2003, then the user also must install the same version of MODI. But if you create object in a DLL written by VC and call it in VB, then VC source code never care about the version of MODI, so the programmer and the user can have different version of MODI.

For all programmers, the REAL problem is: How to install MODI with other versions of Office?

The official solution for this problem is in here:

<http://support.microsoft.com/kb/982760/>

Since you can freely download and install 21 languages of SharePoint Designer 2007, I said "It's free" at beginning.

The unofficial solution:

- a) Download one of the two zip packages I provided in this article. MODI_From_Office2007SP3 is recommended for VC programmer.
- b) Unzip it, read "How to install.txt", then run install.bat as Administrator. Otherwise you'll find the batch file can't copy files.

The benefits of my package:

- a) It includes minimum files for 21 languages.
- b) You can read install.bat and see the install method. Then you can rewrite your own installer if you want. Please remember that "Run as Administrator" is very important for Win 10.

The way that I found out language files to make my package:

- a) Download different language versions of SharePoint Designer 2007.
- b) Install one language MODI in one virtual machine.
- c) Remove duplicated files between virtual machines.

The way that I found out registry entries for every language:

- a) Install a clean XP SP3 virtual machine and duplicate it. Name them as VM_A and VM_B.
- b) Install an uninstall tool, for example InstallRite, in VM_B.
- c) Run the installer of SharePoint Designer from uninstall tool and monitor the modifications of registry and file.
- d) After the monitored installation of SharePoint Designer, export registry list as "aaa.reg".
- e) Import "aaa.reg" into VM_B, and copy 2 folders from VM_A to VM_B:

C:\Program Files\Common Files\Microsoft Shared\MODI

C:\Program Files\Common Files\Microsoft Shared\OFFICE12

- f) Run a home-made test program in VM_B, which will try to delete an entry read from "aaa.reg" and then test if MODI can work. If it can't work, roll back the registry and read next entry. If it works well, record the entry and read next one.

Above way also can be used to find out minimum set of DLLs, but you must minimize registry first.

Enjoy! :-)