

Informatik OCR

Version 1.50

User Guide

May 24, 2010

Download latest version of the user guide from www.informatik.com
Make sure that you have the most current version of the program

Table of Contents

Introduction	3
Trial Version.....	4
How to Use	4
Open Image File	6
Scan.....	6
Rotation	6
Select Section.....	6
Pages.....	7
Text Output	7
Languages	8
Setup	8
DLL (Developers' Kit).....	9
Technical Support.....	9
Tesseract Copyright and License Notice	9
License, Warranty, Disclaimer	10

Introduction

Please read the entire user guide so that you have an understanding of all the options and features offered by the program.

This is a new program. New updates may be released. Please check and make sure that you have the latest version of the program.

Informatik OCR is a Tesseract-based Optical Character Recognition (OCR) program supporting most graphics images, including multi-page TIFF and PDF files.

The downloaded standard version includes the language files for English, French, German and Spanish only. For Italian, Portuguese and Dutch languages, please download the respective language files. They are free. See Languages section below.

Informatik OCR is a desktop application. The image files are displayed, letting you rotate and select sections (crop) if needed before you run the OCR function. For multi-page files you can navigate through the pages.

The text result is stored in a temporary text file (plain text or Rich Text format) and the text file is opened in the file's associated program (for example Notepad or Microsoft Word). If running OCR more than once on the same file, the output text is appended; the text is cleared when you re-open a file.

Informatik OCR is an inexpensive program intended for reasonably clean and good quality images; however it does not compete with expensive top-class OCR programs. The text output layout is in basic format.

Informatik OCR is not suitable for badly skewed text, low resolution images or text in column formats. Also, the program may not correctly read certain fonts. If the document has columnar text, select the columns and OCR the selected areas (see 'Select Sections' section below).

For higher accuracy, but slower process, check the 'AccuracyPlus' button at the bottom of the window.

The program requires Microsoft Framework 3.5 or later.

For a DLL (developer's kit), please contact Technical Support.

Trial Version

You may use the free trial version up to 60 days. After 60 days you must purchase a license or un-install the program. It is unlawful to use the trial version after 60 days.

The trial version prompts you to enter a given temporary code each time you run an OCR and/or inserts random 'DEMO VERSION' strings into the result text.

How to Use

- Open the image file by clicking on the Open Image File button or open it via the option in the File menu. Alternatively scan a document. The scan option is in the toolbar (scanner icon) and in the File menu. You can also paste an image from the Windows Clipboard (via Edit menu). If you have done a previous OCR in the same session, make sure that the output text has been saved before you open a new image file as the text will be overwritten.
- The image is displayed (first page for multi-page TIFF and PDF files).
- Select the language. English is the default. The default can be changed in Setup (setup icon in toolbar or option in File menu). The standard installation includes the Tesseract language file for English, French, German and Spanish only. Language files for many other languages can be freely downloaded. For more information about languages files see 'Languages' section below.
- If the image file is a single-page file, click on the 'OCR this Page' button. Before processing you may want to rotate or crop sections. The OCR process takes a few seconds; you see the progress in the Progress Bar at the bottom of the program window. When the page is processed, a temporary file is created with the resulting text and the text can be opened by clicking on the 'View Text' button, or it may be displayed automatically. The text is either in Rich-Text (RTF) format (default) or in plain text. The default can be changed in Setup. RTF files are opened in the associated application (typically Microsoft Word); plain text files are opened in Notepad. In Setup you can specify the text format. Generally, you need not change the default settings. Review the text, then save it; choose 'Save As' in the text viewer's File menu.
- To OCR all pages of multipage TIFF or PDF files, without preview, click on the 'OCR All Pages' button. The program runs through the pages, quickly displaying them during the OCR process. The process takes a few

seconds; you see the progress in the Progress Bar at the bottom of the program window. When all pages are processed, the resulting text file is displayed, or you may open it by clicking on the 'View Text' button. Review the text, then save it; choose 'Save As' in the text viewer's File menu. Please see the section on single-page processing above regarding the text output format.

- If you need to pause at each page in order to (for example) rotate, crop sections, or bypass pages, click on the 'OCR Page' button. With this option you have control over the pages. Press the Next Page icon (arrow icon) in the toolbar to go to the next page, or press the Select Page icon (question mark icon). After the page is displayed (and you have done the optional rotation and cropping, if needed) click on the 'OCR Page' button. The page is processed and the next logical page is displayed. When you have processed the last page, the result text is displayed. You can also view the output text by clicking on the 'View Text' button. Review the text, then save it; choose 'Save As' in the text viewer's File menu. Please see the section on single-page processing above regarding the text output format.
- If you want to OCR a section of the image, outline the area of interest (while the left mouse button is pressed), then after you release the mouse button, select the OCR option from the popup menu.

Please note...

If the output text is unreadable, check and make sure that you selected the correct Language.

If the output has many errors, check the 'AccuracyPlus' checkbox at the bottom of the window and try again. AccuracyPlus is relevant only if the source image has a very small font.

With some fonts, letters are consistently mis-read. If that happens, you may want to correct the output text file with the Replace function available in the application that opened the text file.

It is generally recommended that you use the Rich Text File (RTF) output format, rather than the Text format (TXT).

When running the OCR operation more than once on the same file, depending on the options selected, the output text will be appended to the previous text. The text is cleared when you re-open a file.

Open Image File

The following image files can be opened for OCR: TIFF, PDF, BMP, GIF, PNG, JPEG. TIFF and PDF files can be multi-page. To open a file, click on the 'Open Image File' button or choose the option from the File menu. Instead of opening an existing image file you can also scan a document. You can also paste an image from the Windows Clipboard (via Edit menu).

Scan

Informatik OCR includes a simple scan option (flatbed or ADF document feeder). The system only supports TWAIN compliant scanners. To scan, choose the Scan icon in the toolbar or select the option in the File menu.

Scanners are notorious for being ill-tempered. If the scan does not work, turn off the scanner and turn it on again.

The scanned image is saved as a temporary image file (in TIFF format). If for some reason you need to save the file permanently, go to the
C:\Documents and Settings\[user]\Application Data\OCRTesseract
folder and copy/rename the file after you have completed the OCR operation.

The scanning option is very basic and is intended to be used only within this OCR program. For a richly featured scanning software we recommend our Informatik Scan. In fact, licensees of Informatik Scan can receive a license of Informatik OCR free of charge.

Rotation

Only horizontal text can be read. If the image is vertical, you need to rotate the image before the OCR operation. To rotate, click on the rotation icons in the toolbar or right-click on the image and select the option.

Select Section

You may want to OCR a specified area of the image. Outline the area of interest (while the left mouse button is pressed), then after you release the mouse button, select the OCR option from the popup menu.

OCR by area is useful when processing documents with columnar text. For these documents you need to select and separately OCR the columns.

Pages

TIFF and PDF files can be multi-page files. To navigate thru the pages click on the Next Page and Previous Page icons (arrows) in the toolbar, or right-click on the image and select the option. To select a particular page, click on the Select Page icon in the toolbar (question mark) and specify the page number.

When navigating thru the pages, press the 'OCR this Page' button. After the last page is processed, the result text is automatically displayed.

Text Output

The text is either Rich-Text format (default, recommended) or in plain text. The default can be changed in Setup. Rich-Text files are opened in the associated application typically Microsoft Word; plain text files are opened in Notepad. To change the format for the current session, click on the RTF/TXT toggle key at the bottom of the window. Generally, you need not change the default settings.

The layout of the output text is in basic format. You may want run the text thru the spell-checker to isolate misreads. When done, save the file under your own file name (choose Save As in the text viewer's File menu).

The temporary text file is saved in the

`C:\Documents and Settings\[user]\Application Data\OCRTesseract` folder. Only one file for each extension name is kept; they are always overwritten.

Languages

The initial installation includes the necessary Tessdata language files for English, French, German and Spanish only. The default language is English but you can change the default in Setup.

The Language dropdown list includes English, French, German, Spanish, Italian, Portuguese and Dutch. Language files (if missing) are freely downloadable from

www.informatik.com/files/Tessdata_Dutch.zip
www.informatik.com/files/Tessdata_English.zip
www.informatik.com/files/Tessdata_French.zip
www.informatik.com/files/Tessdata_German.zip
www.informatik.com/files/Tessdata_Italian.zip
www.informatik.com/files/Tessdata_Portuguese.zip
www.informatik.com/files/Tessdata_Spanish.zip

Unzip the downloaded file and add the files to the Tessdata subdirectory (generally in C:\Program Files\Informatik Inc\Informatik OCR\tessdata). The files for all the required languages must be in that same folder. When you click on a language in the Language dropdown, the system will tell you the name of the subdirectory.

In Setup you can specify two more (yet unspecified supported future) languages but you will need the appropriate Tessdata files. The option is for future use if and when additional languages become available.

Setup

The following specifications can be set in Setup. Setup is accessible via the Setup icon in the toolbar or via the File menu.

Default Language:

Specify the default language

The languages require Tessdata files. The initial installation includes Tessdata files for English, French, German and Spanish only. Additional Tessdata files are freely downloadable. See Languages section above.

Text output format:

By default, the output text files are in Rich Text format. You can change the format to plain text. For the current session you can also change the format by clicking on the RTF/TXT toggle button at the bottom of the main window.

Other languages: (for future use)

Two additional languages can be added: In the 'Other Languages' fields enter the three-character code name for the language (in UPPERCASE). The appropriate Tessdata files for the languages must be copied to the Tessdata subdirectory (see Languages section above). Note, the three-character prefix of the Tessdata files are the same as the code language code name.

After creating additional languages in Setup you must restart the application.

DLL (Developers' Kit)

For a DLL version please contact Technical Support.

Technical Support

Please see the contact information shown in the www.informatik.com web site. Support can only be given for program interface issues. The program uses Tesseract OCR engine; please understand that no technical support for it can be given (like misreading of text, etc). No support can be given for the scanning function as these are normally TWAIN or scanner device problems.

Tesseract Copyright and License Notice

The program uses the TESSERACT (unaltered) free OCR engine distributed under the Apache V2.0 license.

TESSERACT Copyright and License Notice:

Copyright Protected and Licensed under the Apache License, Version 2.0 (the "License"); you may not use this file except in compliance with the License. You may obtain a copy of the License at <http://www.apache.org/licenses/LICENSE-2.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an "AS IS" BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

Informatik OCR Copyright Notice:
Graphics programming, text handling and user interface (excluding Tesseract).
Copyright 2009-210 Informatik Inc. and J Buchmann. All Rights Reserved.
Supplied AS IS without any liability

License, Warranty, Disclaimer

Please read the terms carefully before installing and using the software, as such conduct will indicate your acceptance of all of the terms of this license agreement. If you do not agree with the terms, the software cannot be licensed to you and you must un-install and return the software to Informatik Inc, or its supplier or distributor.

This License Agreement is a legal agreement between Informatik Inc. ("Licensor"), a Delaware Corporation, and you, the user ("Licensee"), and is effective the date Licensee installs the software.

This Agreement covers all materials associated with the Informatik OCR software, including, without limitation, the downloadable software product, online documentation, and any additional supporting electronic files (herein, the "Software").

The evaluation version may be used for 30 days after installation. It is unlawful to use the software after the 30 day evaluation period without licensing the software and paying the license fees. If a license is not obtained before the expiration of the 30 day evaluation period, the Software must be un-installed and destroyed.

1. GRANT OF LICENSE

Licensor hereby grants to you, and you accept, a nonexclusive license to use the Software according to the following condition:

You may use the Software on one (1) computer (PC or workstation, excluding servers) owned, leased, or otherwise controlled by you for personal or business purposes, and only as authorized in this License Agreement. The Software may not be used on other computers, nor may it be used by, or transferred to, other computers over a network. Software must not be installed on servers and Software must not be used in web applications.

2. LICENSOR'S RIGHTS

Licensee acknowledges and agrees that the Software is proprietary to Licensor and protected under international copyright law. Licensee further acknowledges and agrees that all right, title, and interests in and to the Software, including associated intellectual property rights, are and shall remain with Licensor. The License Agreement does not convey to Licensee an interest in or to the Software,

but only a limited right of use that may be revoked in accordance with the terms of this License Agreement.

3. OTHER RESTRICTIONS

This License Agreement strictly forbids distribution of the Software with Licensee's application. Distribution of the Software with Licensee's application requires separate authorization and the payment of license fees.

Licensee agrees to make no more than one (1) back-up copy of the Software. Licensee agrees not to assign, sublicense, transfer, pledge, lease, rent, or share the rights assigned under this License Agreement. Licensee agrees not to reverse assemble, reverse compile, or otherwise translate the Software.

4. TERM

This License Agreement is effective when Licensee installs the Software and shall terminate only if the terms of this License Agreement are broken. Licensee agrees to destroy the Software upon termination of this License Agreement.

5. NO WARRANTY; LIMITATION OF LIABILITY

LICENSEE ACKNOWLEDGES THAT THE PROGRAM IS PROVIDED ON AN "AS IS" BASIS WITHOUT WARRANTY OF ANY KIND. LICENSOR MAKES NO REPRESENTATIONS OR WARRANTIES REGARDING THE USE OR PERFORMANCE OF THE SOFTWARE. LICENSOR incl. DEVELOPER, COPYRIGHTHOLDER, DISTRIBUTOR) EXPRESSLY DISCLAIMS THE WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. LICENSOR SHALL HAVE NO LIABILITY TO LICENSEE OR ANY THIRD PARTY FOR ANY LOSS OR DAMAGE CAUSED, DIRECTLY OR INDIRECTLY, BY THE SOFTWARE, INCLUDING, BUT NOT LIMITED TO, ANY INTERRUPTION OF SERVICES, LOSS OF BUSINESS, LOSS OF DATA OR SPECIAL, CONSEQUENTIAL OR INCIDENTAL DAMAGES.

6. GOVERNING LAW

This License Agreement shall be construed and governed in accordance with the laws of Pennsylvania.

7. SEVERABILITY

Should any court of competent jurisdiction declare any term of this License Agreement void or unenforceable, such declaration will have no effect on the remaining terms hereof.

8. NO WAIVER

The failure of either party to enforce any rights granted hereunder or to take action against the other party in the event of any breach hereunder shall not be deemed a waiver by that party as to subsequent enforcement of rights or subsequent actions in the event of future breaches.

