Amyuni OCR Module

For PDF Creator Version 4.5

ActiveX and .NET

Developer's Guide

Updated October 2010

AMYUNI Consultants – AMYUNI Technologies
www.amyuni.com

Contents

Legal Information

Information in this document is subject to change without notice and does not represent a commitment on the part of AMYUNI. The software described in this document is provided under a license agreement or nondisclosure agreement. The software may be used or copied only in accordance with the terms of the agreement. It is against the law to copy the software on any medium except as specifically allowed in the license or nondisclosure agreement.

The licensee may make one copy of the software for backup purposes. No part of this manual may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or information storage and retrieval systems, for any purpose other than the licensee's personal use, without express written permission of AMYUNI.

Amyuni and the Amyuni logo are trademarks of Amyuni Technologies Inc.

Microsoft, the Microsoft logo, Microsoft Windows, Microsoft Windows NT and their logos are trademarks of Microsoft Corporation.

All other trademarks are the property of their respective owners.

Acknowledgments

This software uses the deflate algorithm developed by Jean-loup Gailly (jloup@gzip.org) and Mark Adler (madler@alumni.caltech.edu). This software is also based in part on the work of the Independent JPEG Group and on parts of the FreeType library.

The Software includes the Open Source Tesseract library in binary format. The Tesseract library source-code can be obtained from http://code.google.com/p/tesseract-ocr/ and is licensed under the Apache 2.0 license. The Tesseract source-code is not included with the Amyuni distribution.

Copyright Notice and License Agreement

AMYUNI CONSULTANTS – AMYUNI TECHNOLOGIES
DEVELOPER/APPLICATION LICENSE AGREEMENT
FOR OCR MODULE PRODUCTS

NOTICE TO USER:
THIS IS A CONTRACT. BY INSTALLING THIS SOFTWARE YOU ACCEPT ALL THE TERMS AND CONDITIONS OF THIS AGREEMENT.

This AMYUNI ("Amyuni") Developer License Agreement accompanies all Amyuni OCR Module product and related explanatory materials ("Software"). The term "Software" also shall include any upgrades, modified versions or updates of the Software licensed to you by Amyuni.

Please read this Agreement carefully. You will be asked to accept this agreement and continue to install or, if you do not wish to accept this Agreement, to decline this agreement, in which case you will not be able to use the Software.

Upon your acceptance of this Agreement, Amyuni grants to you a perpetual but nonexclusive license to use the Software, provided that you agree to the following:

1. Use of the Software. You may include the Software as part of your own applications to install it on your client's machine at the same time as your own applications. The applications can be either in-house or for external distribution. The Software can be used on either workstations or servers without limitations on the number of users. One Application **license is required for every application that includes the Software.The activation code that is provided to you by Amyuni should be kept confidential and not be revealed to end-users, even in the case where the developer's products are sub-licensed to other developers.**

2. Copyright and Trademark Rights. The Software is owned by Amyuni, and its structure, organization and code are the valuable trade secrets of Amyuni. The Software also is protected by Canadian Copyright Law and International Treaty provisions. This Agreement does not grant you any intellectual property rights in the Software.
The Software includes the Open Source Tesseract library in binary format. The Tesseract library source-code can be obtained from http://code.google.com/p/tesseract-ocr/ and is licensed under the Apache 2.0 license. The Tesseract source-code is not included with the Amyuni distribution.

3. Restrictions. You agree not to modify, adapt, translate, reverse engineer, decompile, disassemble or otherwise attempt to discover the source code of the Software. You may not sell the product as a standalone application. You will make all necessary steps to prevent users of your applications to use the Software other than from inside your applications. These steps are described in this manual.

4. No Warranty. The Software is being delivered to you AS IS and Amyuni makes no warranty as to its use or performance. AMYUNI AND ITS SUPPLIERS DO NOT AND CANNOT WARRANT THE PERFORMANCE OR RESULTS YOU MAY OBTAIN BY USING THE SOFTWARE OR DOCUMENTATION. AMYUNI AND ITS SUPPLIERS MAKE NO WARRANTIES, EXPRESS OR IMPLIED, AS TO MERCHANTABILITY, OR FITNESS FOR ANY PARTICULAR PURPOSE. IN NO EVENT WILL AMYUNI OR ITS SUPPLIERS BE LIABLE TO YOU FOR ANY CONSEQUENTIAL, INCIDENTAL OR SPECIAL DAMAGES, INCLUDING ANY LOST PROFITS OR LOST SAVINGS, EVEN IF AN AMYUNI REPRESENTATIVE HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES, OR FOR ANY CLAIM BY ANY THIRD PARTY. Some states or jurisdictions do not allow the exclusion or limitation of incidental, consequential or special damages, or the exclusion of implied warranties or limitations on how long an implied warranty may last, so the above limitations may not apply to you.

BY INSTALLING, USING OR DISTRIBUTING THIS PRODUCT, YOU INDICATE ACCEPTANCE OF THE FOREGOING AGREEMENT.

OCR Module Description

The new Amyuni OCR Module is introduced with the release of version 4.5 of the Amyuni PDF Components.

The OCR module enables developers to:

- Convert non-searchable PDF files into searchable PDFs
- Create searchable PDF documents out of various image formats such as multi-page TIFF, JPEG or PNG while applying text recognition on the images
- Compress image based PDF documents using high compression JBIG2 or more standard CCITT, JPEG and PNG compression formats

The OCR module can be either licensed independently of our other PDF Components or as an add-on to existing licenses.

The Amyuni OCR module is based on the Tesseract Open Source project with the Amyuni PDF technology being used to process and create the PDF documents. The Tesseract library provides high reliability at a low cost and avoids developers the annoyances related to licensing commercial OCR tools which are often licensed on a per-page basis or at a ridiculously high cost to the developer.

Features

- Open multi-page TIFF files directly into PDF Creator for OCR (Optical Character Recognition) and save the documents in PDF format
- Convert image based or non-searchable PDF documents into searchable PDFs
- Apply JBIG2 Compression which heavily reduces size of scanned documents. Other standard compression formats such as CCITT, JPEG or PNG can also be used
- Support for multiple languages such as English, French, Italian, German, Portuguese, Spanish, Dutch and Vietnamese
- Obtain up to 98% accuracy on English language documents
- Extracted text can be either visible or hidden inside the PDF document. In both cases, the text is positioned as close as possible to the original text
- Extracted text can be saved to a regular text file rather than to a PDF file
- Rasterize any PDF document to convert it into an image based searchable or non-searchable PDF

- Benefit from a robust PDF library that can create highly optimized and well structured PDF documents that can be emailed and viewed by any PDF compatible viewer

Supported Platforms

- Support for all Windows platforms include:
    - **32-bit ver.: Windows 7, Windows 2008, Vista, XP, 2003, 2000**
    - **64-bit ver.: Windows 7, Windows 2008, Vista, XP, 2003**

Distributable Files

PDFCreactiveX.dll

This is the main ActiveX control that hosts the Amyuni PDF Library and the interface to the OCR engine.

acPDFCreatorLib.Net.dll

This is the .NET class library that is equivalent to the PDFCreactiveX.dll ActiveX control. Developers can either use the ActiveX or .NET but do not need to include both.

Tessdll.dll

This file contains the Tesseract OCR engine. This DLL should be located in the main folder as PDFCreactiveX.dll.

Tessdata Folder

This folder contains all the dictionaries used by the OCR engine. Each language is supported by 8 dictionary files prefixed with the language name, e.g deu for German. If not all languages are needed, then only the required languages can be distributed, e.g. only the eng and fra prefixed files can be distributed for English and French only support.

Important Note

All the samples that are provided in this documentation assume that the developer is using the ActiveX version (PDFCreactiveX.dll.) When using the .NET version (acPDFCreatorLib.Net.Dll), the functions are very similar although the code slightly different. Rather than duplicating all the documentation and sample code, we have chosen to provide a complete .NET sample at the end of this documentation.

General Operation

The OCR module can be used for one of two purposes:

1. Open an existing PDF and convert all text into searchable text. A number of PDF files contain only images or text that is not searchable. Applying OCR on these PDF files will make the text searchable. After opening the PDF file and applying OCR, the PDF can be resaved as another PDF or the text contents exported into a plain text file. Here is a sample VB script that will illustrate that:

```
Dim pdf

set pdf = CreateObject( "PDFCreactiveX.PDFCreactiveX.4.5" )
pdf.SetLicenseKey "Amyuni", "YQ...lI"

pdf.Open "ocrtest.pdf", ""
pdf.OCRPageRange 1, pdf.PageCount, "eng", 1

pdf.Save "ocred.pdf", 1

Set pdf = Nothing
```

2. Create a searchable PDF file from various image files such as TIFF, JPEG or PNG. The PDF file is creating by loading each image individually, applying OCR and resaving the PDF. Here is a sample VB script that will illustrate that:

```
Dim pdf

set pdf = CreateObject( "PDFCreactiveX.PDFCreactiveX.4.5" )
pdf.SetLicenseKey "Amyuni", "YQ...lI"

pdf.Open "Tif to OCR.tif", ""
pdf.OCRPageRange 1, pdf.PageCount, "eng", 1

pdf.Save "ocred.pdf", 1

Set pdf = Nothing
```

SetLicenseKey Method

The SetLicenseKey method is used to set Licensing Information.

Before any operation can be done on the PDF Creator control, the control should be activated using this method.

Syntax

C++:

```
HRESULT SetLicenseKey (BSTR Company, BSTR LicKey)
```

C#:

```
void SetLicenseKey (string Company, string LicKey)
```

VB:

```
Sub SetLicenseKey (Company As String, LicKey As String)
```

Parameters

Company

Name of the company or private user having licensed the product.

LicKey

License key provided by Amyuni Technologies when downloading or purchasing a product.

Example in VB:

```
Const LicenseTo = "Evaluation Version"
Const ActivationCode = "07EFCDAB010…DF7E8AB55D617055803A"
Private Sub cmdActivateObject _Click ()
     PDF1.SetLicenseKey LicenseTo, ActivationCode
End Sub
```

Important Note

The activation code might contain special characters which are not valid within a string and for the programming language that you are using. The special characters should be replaced by their escape characters. E.g. in VB, a double-quote can be replaced by Chr(34). In C++, special characters such as '\' or '"' can be replaced by '\\' or '\"'.

Open and OpenEx Methods

The Open and OpenEx methods open a PDF, XPS, TIFF or other image format documents. The Open method reads all the file contents into memory and closes the file right-away. The OpenEx method opens each page as it is requested, keeping the file handle open. OpenEx has the advantage of being more efficient but prevents other applications from writing to the file while it is being viewed.

Syntax

C++:

```
HRESULT Open (BSTR FileName, BSTR Password, [out, retval] BOOL *Result)
HRESULT OpenEx (BSTR FileName, BSTR Password, [out, retval] BOOL *Result)
```

C#:

```
bool Open (string fileName, string password )
bool OpenEx (string fileName, string password )
```

VB:

```
Function Open (FileName As String, Password As String) As Long
Function OpenEx (FileName As String, Password As String) As Long
```

Parameters

FileName

Name of the PDF or TIFF file to open.

Password

Password to use if document is protected (PDF only.)

Return Value

Result

True if document opened successfully, False otherwise.

Example in C++:

```
long  result = m_pdf->OpenEx( _bstr_t("pwd.pdf"), _bstr_t("") );
if ( !result ) {
      //This call is only useful after the first attempt to open the file
      if ( m_pdf->Protected ) {
            // ask user for password and try again...
      }
}
```

OCRPageRange Method

The OCRPageRange method performs OCR on a range of pages or the complete document. The document is first opened using the Open or OpenEx methods and should be resaved afterwards, this method by itself will not resave the document.

Syntax

C++:

```
HRESULT OCRPageRange (long StartPage, long EndPage, BSTR Language, acOCROptions Options)
```

C#:

```
void OCRPageRange(int startPage, int EndPage, string Language, acOCROptions Options)
```

VB:

```
Sub OCRPageRange (StartPage As Long, EndPage As Long, Language As String, Options As Long)
```

Parameters

StartPage, EndPage

Start and end page numbers to OCR. Page numbers start with page 1.

Language

3 letter ISO_639_Language_Code indicates which dictionary to use during OCR. OCR accuracy is greatly improved by indicating to the OCR engine which is the main document language. Support values are:

eng (English), fra (French), ita (Italian), deu (German), por (Portuguese), spa (Spanish), vie (Vietnamese), nld (Dutch)

Options

Only one option is currently supported:

acOCROptionVisibleText = 1

By default, the text that is retrieved from the OCR engine is hidden and lies on top of the original document contents. This makes the document searchable without the text hiding the original document contents. When this option is set to 1, the text is visible. This option should be set to 1 in order to extract the text to a separate TXT or RTF file.

Return Value

None

This method launches an exception when an error occurs. Exceptions should be handled properly by the calling application.

```
Dim pdf

set pdf = CreateObject( "PDFCreactiveX.PDFCreactiveX.4.5" )
pdf.SetLicenseKey "Amyuni", "YQ...lI"

pdf.Open "Tif to OCR.tif", ""
pdf.OCRPageRange 1, pdf.PageCount, "eng", 1

pdf.Save "ocred.pdf", 1

Set pdf = Nothing
```

RasterizePageRange Method

The RasterizePageRange method converts page contents into a color or grey scale image. When archiving documents or performing OCR, it is sometimes preferable for all pages to be stored as images rather than complex text and graphic operations. The document is first opened using the Open or OpenEx methods and should be resaved afterwards, this method by itself will not resave the document.

Syntax

C++:

```
HRESULT RasterizePageRange (long StartPage, long EndPage, long Resolution,
acRasterizeColorOption ColorOption, acImageCompressionConstants compression)
```

C#:

```
void OCRPageRange(int startPage, int EndPage, int Resolution, acRasterizeColorOption
ColorOption, acImageCompressionConstants compression)
```

VB:

```
Sub OCRPageRange (StartPage As Long, EndPage As Long, Resolution As Long, ColorOption
As Long, Compression As Long)
```

Parameters

StartPage, EndPage

Start and end page numbers to OCR. Page numbers start with page 1.

Resolution

Resolution in dots per inch (DPI) at which the documents are to be rasterized. A DPI of 150 is recommended in most situations. Higher resolutions will produce very large PDF files.

ColorOption

- acColorOptionBW = 1 Black and White images
- acColorOptionGray = 2 256 Grey level images
- acColorOptionRGB = 3 RGB color images

Compression

- acCompression256Colors = 1 256 Color level images
- acCompressionJPegLow = 2 Low quality JPeg Compression
- acCompressionJPegMedium = 7 Medium quality JPeg Compression
- acCompressionJPegHigh = 9 High quality JPeg Compression
- acCompressionCCITTFax = 10 CCITT Level 6 Black and White Compression
- acCompressionPNG = 11 PNG (Flate) Compression
- acCompressionJPG2000 = 12 JPeg 2000 Compression
- acCompressionJBIG2 = 13 JBIG2 Compression

Return Value

None

This method launches an exception when an error occurs. Exceptions should be handled properly by the calling application.

```
Dim pdf

set pdf = CreateObject( "PDFCreactiveX.PDFCreactiveX.4.5" )
pdf.SetLicenseKey "Amyuni", "YQ...lI"

pdf.Open "ocrtest.pdf", ""
pdf.RasterizePageRange 1, pdf.PageCount, "eng", 1
pdf.OCRPageRange 1, pdf.PageCount, "eng", 1

pdf.Save "rasterized.pdf", 1

Set pdf = Nothing
```

Save Method

The Save method saves the current PDF document to a file.

Syntax

C++:

```
HRESULT Save (BSTR FileName, FileSaveOptionConstants SaveOption)
```

C#:

```
void Save (string FileName, FileSaveOptionConstants SaveOption)
```

VB:

```
Sub Save (FileName As String, SaveOption As FileSaveOptionConstants)
```

Parameters

FileName

Name of the file where to save the document.

SaveOption

This option specifies what data to write in the saved document. The SaveOption is of type FileSaveOptionConstants which is defined as follows:

Option	Value	Description
acFileSaveView	1	Save as a regular PDF file.
acFileSavePDFA 4	4	Save document in PDF/A format, PDF Specifications Version 8.
acFileSavePDF14	5	Save document PDF Specifications Version 1.4.

Please refer to the other methods described in this document for sample code.

ExportToRTF Method

The ExportToRTF method is used to export a PDF document to RTF or Text format after text recognition has been applied to it.

Syntax

C++:

```
HRESULT ExportToRTF (BSTR FileName, acRtfExportOptions Option, long UseTabs)
```

C#:

```
void ExportToRTF (string fileName, acRtfExportOptions Option, long UseTabs)
```

VB:

```
Sub ExportToRTF (FileName As String, Option As acRtfExportOptions, UseTabs As Long)
```

Parameters

FileName

Name of the file with .rtf or .txt extension to export to

Options

Option	Value	Description
acRtfExportOptionAdvancedRTF	0	Advanced RTF: using frames to position objects.
acRtfExportOptionFullRTF	1	Full RTF: Text, Graphics and images with no frames.
acRtfExportOptionRTFText	2	Formatted Text only.
acRtfExportOptionText	3	Simple text, non-formatted.

UseTabs

Set this parameter True to enable tabs in the document, False (or 0) to replace tabs with spaces (Effective only for non-formatted simple text).

Example 1 in VB:

```
With PDFCreactiveX1
      .Open "test.pdf", ""
      'optimize the document before exporting
      .OptimizeDocument 1 'recommended Line optimization
      'export the PDF file to simple text, non-formatted RTF
      'with tabs enabled
      .ExportToRTF "exportRTF.rtf", 3, True
End With
```

Example 2 in C#:

```
try
{
      axPDF.Open("temp.pdf", "");
      axPDF.OptimizeDocument(1);
```

```
        axPDF.ExportToRTF("rtf_export_test_ver2.txt",
        acRtfExportOptions.acRtfExportOptionText, 0);
    }
    catch (Exception ex)
    {
        MessageBox.Show(ex.Message);
    }
```

Commented Sample using the .NET Library

The acPDFCreatorLib.Net.dll should first be added to the project's references.

```
using Amyuni.PDFCreator;

// instantiate the main PDF document object
IacDocument pdf = new IacDocument();

// set the license key before any operation can be done on the object
pdf.SetLicenseKey("Amyuni", "07EFC...546");

// create a file stream and open the file from the stream
FileStream fIn = new FileStream("ocrtest.pdf", FileMode.Open, FileAccess.Read);
pdf.Open(fIn, "");

// rasterize the first 3 pages (Optional)
pdf.RasterizePageRange(1, 3, 120, IacRasterizeColorOption.acColorOptionGray,
IacImageCompressionConstants.acCompressionDefault);

// apply OCR on the first 3 pages
pdf.OCRPageRange(1, 3, "eng", IacOCROptions.acOCROptionVisibleText);

// save the resulting PDF to a stream
FileStream fOut = new FileStream("rasterized.pdf", FileMode.OpenOrCreate,
FileAccess.ReadWrite);
pdf.Save(fOut, IacFileSaveOption.acFileSaveView);

// cleanup
pdf.Dispose();
fIn.Close();
fOut.Close();
```

Links to Support and Documentation:

If you have any questions or problems with our products, the following resources are available to you through our web site:

Online Documentation:
http://www.amyuni.com/WebHelp/Developer_Documentation.htm#index.htm

Frequently Asked Questions:
http://www.amyuni.com/forum/viewforum.php?f=18

Technical Notes:
http://www.amyuni.com/en/resources/technicalnotes/

User forum:
http://www.amyuni.com/forum/index.php

Posting questions to our technical support staff:
http://www.amyuni.com/en/support/getsupport/

We also provide some additional tools that can be downloaded free of charge and used with the PDF Creator product. These tools are available at:
http://www.amyuni.com/en/resources/freetools/