Aquaforest

Aquaforest SDK
Release Notes

Aquaforest SDK

Release Notes

Version 3.2
March 2024

Content

1 Version 3.2.2402.15

1.1 Bug fixes

- `[SDK-210]` When "Remove Hidden Text" is set to "True", pages with "visible text" are re-OCRed

2 Version 3.2.2308.09

2.1 Bug fixes

- `[SDK-199]` Visible text is removed from document after OCR process.

3 Version 3.1.2206.08

3.1 Bug fixes

- `[SDK-183]` Fixes `System.Runtime.InteropServices.SEHException` in PDF rasterizer
- `[SDK-184]` Fixes exception when using ConvertToTiff and RetainBookmarks
- `[SDK-186]` Fixes page size setting which was being ignored when creating a PDF page using the PDF Toolkit

3.2 Enhancements

- [SDK-185] Improves performance of PDF rasterizer
- [SDK-187] Adds support for subscription license

4 Version 3.0.2110.27

4.1 Bug fixes

- `[SDK-181]` Extended OCR engine adds multiple instances of words in the output PDF document when OCR’ing in Native mode.

5 Version 3.0.2110.07

5.1 Bug fixes

- `[SDK-171]` Cloud OCR generates 1KB when no pages are processed in `ConvertToTiff` mode.
- `[SDK-176]` Searchability check returns wrong result for image-only files when using a trial license.
- `[SDK-177]` OCR damages PDF files containing both visible and hidden text in a page when `RemoveExisitingPDFText = true` and `PdfToImageIncludeText = false`.
- `[SDK-180]` Bug fixes and improvements related to PDF page rasterization.

6 Version 3.0.2104.29

6.1 Enhancements

- Updated Extended OCR engine to version 15.6.8.3176.

6.2 Bug fixes

- `[SDK-166]` Fixed issue with mirroring that occurs when processing certain TIFF files.

7 Version 3.0.2104.21

7.1 Bug fixes

- `[SDK-169]` SDK fails with an exception when processing large PDF documents (2GB+).
- `[SDK-170]` Microsoft Cloud OCR fails when processing in Handwritten mode.

8 Version 3.0.2103.26

8.1 Bug fixes

- `[SDK-168]` SDK does not release cores when there is an exception.

9 Version 3.0.2103.12

9.1 Bug fixes

- `[SDK-164]` Text is removed from PDF forms when OCRing PDFs in Native mode.

10 Version 3.0.2102.25

10.1 New Features

- Added properties.xml for the Cloud OCR engine.
- Added `EmbedFontSubset` setting in the Standard OCR engine.

10.2 Improvements

- `[SDK-162]` Improved font handling and character encoding when OCRing documents (for all 3 OCR engines).
  Implemented 3 different ways of handling documents that contain characters that cannot be encoded with available fonts through the `FontEncodingErrorMode` setting that can be set to one of the following:
    - `ReplaceCharacters` - Replace the specific character that could not be encoded with the value set in `FontEncodingErrorReplacementString` and continue processing. Default is the tilde character '~' if not set.
    - `IgnoreCharacterOrWord` - Ignore the character or the word containing the character that could not be encoded and continue processing.
    - `FailDocument` - Stop processing the whole document and exit with error.

10.3 Bug fixes

- `[SDK-161]` Output document is not generated when `CreationDate` is set in Extended OCR.

11 Version 3.0.2101.19

11.1 Improvements

- `[SDK-159]` Add new setting in aquaforestimage to not fallback to libtiff if leptonica fails
  Sometimes, if there is an image with an unusual combination of bits per pixel and compression, a completely black image is generated during pre-processing. The black image is then added to the PDF files instead of the original image. A new setting (`LibTiffErrorHandling`) has been added, such that, when set to `false`, processing of the whole document stops completely when this happens instead of adding the black pages and continuing processing.

11.2 Bug fixes

- `[SDK-158]` Processing of whole document stops in non-native mode if a page fails during aquaforestimage pre-processing, instead of processing the page as image-only
- `[SDK-157]` Processing certain files in Native mode in the Extended OCR engine 'breaks' the document

12 Version 3.0.2010.23

12.1 New Features

- SDK 3.0 has the following new components
  - Data Extraction engine to automatically extract name-value pairs from PDF documents
  - PDF Toolkit to manipulate PDF files
  - Cloud OCR engine that supports OCRing documents using the Google or the Microsoft OCR engines
- Added a new extensible logger. Previous versions only supported outputting log information to the console or a file. A new `IAquaforestLogger` interface has now been added that can be extended to output anywhere you want. It also has options to set the log level to either Debug, Information, Warning, Error to filter out logs. The logger can be injected into the constructor of all SDK 3.0 components. As a result of this change, `EnableConsoleOuput` and `EnableDebugOutput` settings have been removed as they can now be controlled by the logger object. A reference to `Aquforest.Logging` must be added to use the `IAquaforestLogger` interface

12.1.1 OCR

- `[SDK-143]` Add new iDRS engine in Extended OCR
  - PDF/A-3a and PDF/A-3b output is now supported
  - Vietnamese and Thai language is now supported
  - Added new `BlankPageDetectionMode` property that allows choosing which algorithm to use for blank page detection
- `[SDK-151]` Add options to stop processing when page(s) fail(s) to process in Native mode. Previously, if the OCR engines failed to OCR all pages of a document in Native, the original source document was outputted. However, this made it difficult to mark and identify the document as failed.
  The new `NativeProcessingErrorMode` property can be used to indicate to the OCR engine how to respond if there are issues processing page(s) in Native mode. The available options are:
  - Do not stop processing if a page fails to process
  - Stop processing only if all pages fail to process
  - Stop processing if at least one page fails to process
- `[SDK-152]` Add a method or property to check if PDF is secured
- `[SDK-153]` Implement mobile capture extension in Extended OCR. This is controlled by the `PerpectiveCorrection` setting

12.1.2 PDF Toolkit

- `[PTK-2]` Enable access to co-ordinates for words in extracted text
- `[PTK-8]` Ability to create/retain bookmarks
- `[PTK-15]` Extract Text from areas
- `[PTK-23]` Add the ability to flatten PDF Form (XFA Forms)
- `[PTK-27]` Image To PDF Converter
- Read XFA form data

12.2 Improvements

12.2.1 OCR

- Updated PDF tool used in the Extended OCR engine to process PDFs in Native mode
- `[SDK-128]` Check validity of "Temp" Folder to provide additional debug information to customers
- `[SDK-130]` Implement iDRS 15.4.5 OR higher to address using ForceTableZones in conjunction with Arabic Language

- `[SDK-155]` Improve memory usage when processing large files
- `[SDK-123]` Changed the way the Extended OCR engine is initialised. Previously, the engine was initalised when a new `Ocr` object was created.

```
using (Ocr ocr = new Ocr(resourceFolder))
{
    […]
}
```

Now it needs to be initialised on application start and unloaded at the end.

```
ExtendedOcrEngine.SetupOcr(LICENSE, RESOURCES_FOLDER);

using (Ocr ocr = new Ocr(logger))
{
    […]
}

ExtendedOcrEngine.UnloadOcr();
```

12.3 Changes

- SDK 3.0 is built against .NET Framework 4.7.2
- The license key, resource folder and an optional logger must now be specified in the constructor of OCR, Barcode and Data Extraction engines. As a result, `License` and `ResourceFolder` properties have been removed from these products
- `DecodeResult.BarcodeResult` has been removed from the Barcode engine. To access barcode results, use `DecodeResult.BarcodeResults` instead
- `[SDK-129]` Removed `EmbedFonts` property and replaced it with `EmbedFontsSubset` in Extended OCR engine
- `AsianOCREngine` is now obsolete and do not need to be set in Extended OCR engine

12.4 Bug fixes

12.4.1 OCR

- `[SDK-112]` Processing PDFs that have already been OCRed with another product in Native mode with `RemoveExistingPDFText` set to `True` causes double text layers when OCRed by Aquaforest OCR engines
- `[SDK-122]` [Extended OCR] Running multi-threaded OCR through the SDK API throws System.AccessViolationException
- `[SDK-124]` Standard OCR and Extended OCR engines cannot be used in the same project
- `[SDK-125]` Very poor performance when processing certain PDF files
- `[SDK-131]` Pages in certain PDF document go blank when processing with Standard OCR engine
- `[SDK-135]` OCR SDK Hangs when processing in Native Mode
- `[SDK-138]` [Extended OCR]: Ocr.DeleteTemporaryFiles() is not thread-safe
- `[SDK-140]` Assembly load error when using both OCR engines in one application
- `[SDK-141]` OCRing an image with Extended OCR and 'no OCR = true' causes "Invalid call to method" error
- `[SDK-142]` Barcode is not detected if PerformPreprocessing is true or BlankPageThreshold > -1
- `[SDK-150]` Visible text is rendered as unknown symbols when processing in PDF with custom encoding in Native mode
- `[SDK-154]` Retrieving "Bits Per Component" from images in PDF pages throws exception

12.4.2 PDF Toolkit

- `[PTK-1]` Provide Word to PDF feature
- `[PTK-9]` License Pop Up issue on single page
- `[PTK-10]` Trial version should limit text extraction to 3 pages
- `[PTK-16]` Displays “Did not close PDF file” even if the document was not open in the first place
- `[PTK-17]` Warning message 'You did not close a PDF Document' being displayed when processing secure PDF
- `[PTK-25]` Not extracting the last item of text on a page
- `[PTK-30]` Convert an image file to PDF: "alpha channel not implemented" error when processing JPEG file

12.5 New Sample Projects

12.5.1 OCR

- Sample project to demonstrate how to get text from PDFs with different searchability level. If the PDF is already searchable (i.e., contains text in all the pages), it will return the text without performing any OCR. If the PDF is not searchable, it will OCR it first and then gets the text. This sample is available for all three OCR engines
- `[SDK-127]` Sample project to process color pages with Extended OCR engine
- `[SDK-131]` Sample project to demonstrate retrieving page confidence score with Standard OCR

12.5.2 PDF Toolkit

- Extract PDF Text as HOCR JSON
- Extract Text from Form
- Image Files merged to PDF
- Searchable PDF to Image PDF

12.5.3 Data Extractor

See Welcome page for Data Extractor samples.

12.5.4 Cloud OCR

See Welcome page for Cloud OCR samples.