Aquaforest OCR
SDK for .Net

Reference Notes Version 1.20
Build 100817

August 2010

http://www.aquaforest.com/

New features included in version 1.20 of the SDK are listed below.

1 LINE REMOVAL

Line removal has been added to improve the handling of tables and lines through or touching characters. Note, lines are removed from the copy of the image that is used in the OCR process and not from any image that is used to generate PDF output.

2 NEW STATUS EVENT –

New event “void StatusUpdate(object sender, StatusUpdateEventArgs statusUpdateEventArgs)” The StatusUpdateEventArgs provides information relating to status of the page processed including page number, image processing outcome, whether text was extracted, whether the page was detected as blank and the orientation used.

3 OCR CONFIDENCE SCORING

The new status event described above also includes a Confidence Score which gives an indication of the quality of the OCR output and can be used to detect bad scans or other problem pages.

4 BLANK PAGE DETECTION AND REMOVAL

BlankPageThreshold can be used to set the minimum number of “on” pixels that must be present on a page image. Any page for which the image that has fewer “on” pixels than the limit will be excluded from the output. A value of 100 produced reasonable blank page detection in testing, but the validity of this should be confirmed using “typical” source documents.

5 PDF STAMP SUPPORT

ConfigurePDFStamp has been added to the PreProcessor class and using this method stamps can be configured to be added to each page of the PDF output. The stamps contain one or more of the following:

Prefix – a string to be added to the beginning of the stamp, before the number section.
Start – the value that the number portion of the stamp should start at. The number portion will be incremented by 1 each page.
Digits – a value indicating the minimum length that the number portion of the stamp should be displayed as. Preceding 0’s will be used to pad any numbers less than this whilst numbers greater than this will be displayed in full.
Suffix - a string to be added to the end of the stamp, after the number section.

Thus a stamp with Prefix = “Beginning”, Start = “1”, Digits = “4” and Suffix = “End” would produce the text “Beginning0001End” on the first page. Any one of these can be set to null resulting in the exclusion of that part from the final text.

Additionally the stamp can be added either as visible searchable text or as an image and can be positioned in one of the following:

Top Left
Top Right
Centre
Bottom Left
Bottom Right
Top Centre
Centre Left
Centre Right
Bottom Centre

6 IMPROVED TEMPORARY SPACE CONTROL

DeleteTemporaryFilesOnPageCompletion. When set to true the temporary files generated for each page during OCR processing will be removed as soon as the OCR engine has finished with them which can save a significant amount of disk space when processing large multi-page files*.

*Note: the OCR engine is finished with the temporary files for a page as soon as the output for that page is added to the overall output. If you wish to use functionality such as ReadPageWords, GetPageImage, etc then this will require that the temporary files are available for the page requested and so will fail if DeleteTemporaryFilesOnPageCompletion is true.

7 DOT MATRIX FONT SUPPORT

Dotmatrix. This property on the OCR object can be set to true to improve recognition for dot-matrix fonts. If set to true when processing fonts other than dot-matrix the recognition quality can be poor. Default value is false.

8 PDF FILE MERGING

PdfMerger class has been added to allow the creation of PDFs form multiple sources. This might be useful if you wish to merge multiple single page TIFFs into a single searchable PDF or OCR particular pages from a source document into an output document.

9 ENHANCED OPEN PDF SUPPORT

The support for opening PDF as a source type has been extending to include a greater range of image formats stored within the PDF and improve handling of the variety of ways in which the image data can be defined and stored.