Aquaforest OCR
SDK for .Net

Reference Notes Version 1.1
Build 100304

March 2010

http://www.aquaforest.com/

The SDK now supports the use of PDF documents as a source types for the OCR process. The processing will extract images (including TIFF Group 3, Group 4, RLE, LZW, JPEG (type 6), JPEG 2000, JBIG2) from a PDF page and perform OCR on the extracted images. Then, assuming PDF output is enabled, it will insert the recognised text into the final PDF page in the correct position relative to the image.

Two new options are available to support this:

1. ReadPDFSource has been added to allow the opening of PDF source files.

2. RemoveExistingPDFText if set to true will result in the removal of any existing text from the output PDF*.

*Note: when PDF output is generated from a PDF source it is a copy of the PDF that is manipulated rather than generating a new one. This approach offers several advantages such as potential size savings and performance enhancements.

Please note the following limitation of this release of the SDK when using a PDF document as the source for the OCR. Whilst autorotation can be set to true, if pre-existing text is found on a page and the option to remove exclude this from the output is not set to true, then the autorotation will be overridden for that page.

Other enhancements that have been added to this release include:

1. 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.

2. 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 Centre
    - Top Right
    - Centre Left
    - Centre
    - Centre Right
    - Bottom Left
    - Bottom Centre
    - Bottom Right