Aquaforest OCR

SDK for .Net

Release Notes Version 1.30

Build 110905

August 2011

http://www.aquaforest.com/

New features included in version 1.30 of the SDK are listed below.

1 READIMAGESOURCE

The OCR SDK API has been extended to include .NET Image objects to be passed as a source.

2 IMPROVEMENTS TO AUTOROTATION

Dictionary lookup has been added to the checks to determine “good” results for rotated pages as well as flip detect for the initial test on whether a page is rotated. These two measures have greatly improved the detection of rotated pages and the determination of the correct rotation to use.

3 OPTIMISED OCR

When the OptimiseOcr property on the OCR object is set to false the OCR and image processing engines will use the settings in the ImagePreProcessingDefaults section of the file Properties.xml modified by any properties set on the OCR and PreProcessing objects.

Setting OptimiseOcr true will enable the use of these default settings first (without modification by the properties set on the OCR and PreProcessing objects followed by the same defaults modified by the values in the ImagePreProcessing sections from ID="1" to ID="n" where n is the last consecutive set defined in Properties.xml.

Using heuristics and dictionary lookup the quality of the OCR output is then compared in order to determine the optimum set to output. In this way it is possible to define different sets of OCR and pre-processing conditions that are suited to different types of source documents. This approach can also improve the handling of documents that contain different types of pages, e.g. scanned at different qualities, containing different languages, containing standard and dot matrix prints, etc.

4 CHANGES TO PROPERTIES FILE

The following are descriptions of those properties in the file Properties.xml that are most likely to be changed to improve engine performance. If you require further information regarding any properties in the file then please contact Aquaforest via support@aquaforest.com for assistance.

Binarize – This setting determines how the image will be converted into a bitonal one for OCR. The following are valid options:

> -1 – This utilizes a technique whereby those parts of the image that have certain characteristics indicative of characters are extracted from the underlying image. This approach can give the best results on pages such as magazine images, news print, etc and will handle light text on darker backgrounds. This approach can cause an increase in processing time with certain images.
> 0 – This utilizes the binarisation capabilities built into the OCR engine and whilst it can give good results in limited situations it is not generally recommended.
> >0 – A value greater than 0 (the recommended default is 200) will use a simple threshold technique comparing the intensity of the pixel to the threshold value to determine whether it should be set to black or white. This simple approach is the fastest option.

BoxSize – Setting a value above 0 will cause the removal of enclosing boxes from the image used for the OCR processing. The default recommended is 100, i.e. where the box edges are 100 pixels or greater.

BackgroundFactor - Sampling size for the background portion of the image. The higher the number, the larger the size of the image blocks used for averaging which will result in a reduction in size but also quality. Default value is 3

DotMatrix - Set this to True to improve recognition of dot-matrix fonts. Default value is False. If set to true for non dot-matrix fonts then the recognition can be poor

ForegroundFactor - Sampling size for the foreground portion of the image. The higher the number, the larger the size of the image blocks used for averaging which will result in a reduction in size but also quality. Default value is 3

Jbig2EncFlags – These are the flags that will be passed to the application used to generate JBIG2 versions of images used in PDF generation (assuming this compression is enabled). Options are as follows:

-b <basename>: output file root name when using symbol coding
-d --duplicate-line-removal: use TPGD in generic region coder
-p --pdf: produce PDF ready data
-s --symbol-mode: use text region, not generic coder
-t <threshold>: set classification threshold for symbol coder (def: 0.85)
-T <bw threshold>: set 1 bpp threshold (def: 188)
-r --refine: use refinement (requires -s: lossless)
-O <outfile>: dump thresholded image as PNG
-2: upsample 2x before thresholding
-4: upsample 4x before thresholding
-S: remove images from mixed input and save separately
-j --jpeg-output: write images from mixed input as JPEG
-v: be verbose

Language – The acceptable vales are as follows:

0 - English
1 - German
2 - French
3 - Russian
4 - Swedish
5 - Spanish
6 - Italian
7 - Russian English
8 - Ukrainian
9 - Serbian
10 - Croatian
11 - Polish
12 - Danish
13 - Portuguese
14 - Dutch
19 - Czech
20 - Roman
21 - Hungar
22 - Bulgar
23 - Slovenian
24 - Latvian
25 - Lithuanian
26 - Estonian
27 - Turkish

MaxDeskew – Maximum angle by which a page will be deskewed.

Morph – Morphological options that will be applied to the binarized image before OCR. If left blank none is applied. Common options include those listed below but for more options please contact support@aquaforest.com:

> d2.2 – 2x2 dilation applied to all black pixel areas, useful for faint prints.
> e2.2 – 2x2 erosion applied to all black pixel areas, useful for heavy prints.
> c2.2 – closing process that performs a 2x2 dilation followed by a 2x2 erosion with the result that holes and gaps in the characters are filled.

NoPictures - By default, if an area of the document is indentified as a graphic area then no OCR processing is run on that area. However, certain documents may include areas or boxes that are identified as “graphic” or “picture” areas but that actually do contain useful text. Setting NoPictures to True will cause it to ignore areas identified as pictures whilst setting it to False will force OCR of areas identified as pictures.

OneColumn - The default value for this is true which improves the handling of single column text. Better handling of multi-column text such as magazine or news print can be achieved.

PdfToImage – The SDK ships with two engines for the conversion of PDF pages to images for OCR. The default engine is used when this is set to 0 but if certain PDF source documents are proving problematic then the alternate engine can be used by changing this value to 1.

PdfToImageIncludeText – When set to False this will prevent the conversion of real text (i.e. electronically generated as opposed to text that is part of a scanned image) from being rendered in the page images extracted from the PDF. This is because the text is already searchable and so generally does not require OCR. The value can be set to True however if the OCR is required on this real text.

Quality - JPEG quality setting (percentage value 1 - 100) for use in saving the background and foreground images. Default value is 75

RemoveLines – The value used in Line removal. If blank no line removal will occur. The normal value to use to enable line removal is 100.5 but it you are experience difficulties with this value or have any questions then please contact support@aquaforest.com.