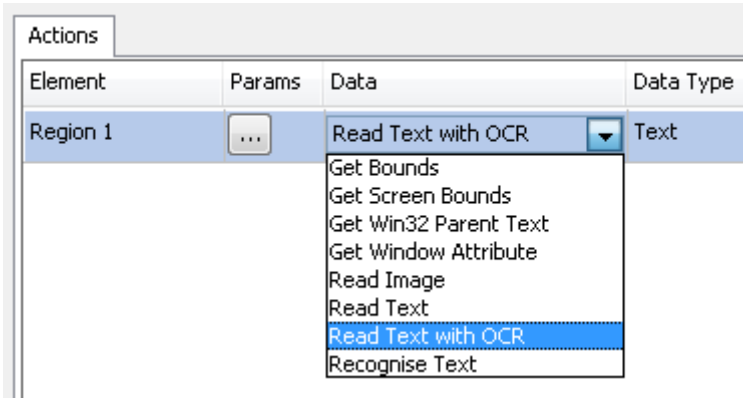
Read Text with OCR

Blue Prism's Read Text with OCR action uses Google's Tesseract open source OCR (Optical Character Recognition) engine to be able to read text without identifying the font or disabling font smoothing.

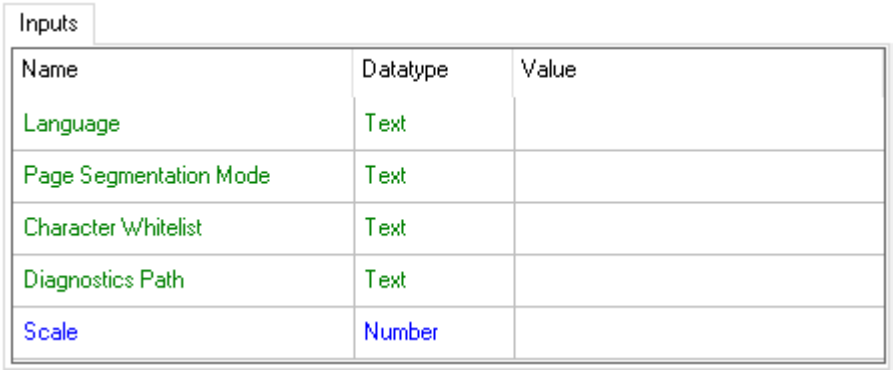
Using Read Text with OCR

To use Read Text with OCR, spy a Region element, drag it into a Read stage and the option will appear in the Data dropdown, as shown below.



Settings

The OCR engine can work well with its default settings, but the Read stage input parameters can be adjusted if necessary.



Language

- This is used to specify a language other than the default, which is English.
- Download language pack for correct version of Tesseract from: https://code.google.com/p/tesseract-ocr/downloads/list
- Extract the language pack to "C:\Program Files\Blue Prism Limited\Blue Prism Automate\Tesseract\tessdata"
- Use the 3 character ISO code to call the language needed within BP, it will be the same as the file extracted, e.g. SPA for Spanish, FRA for French

Page Segmentation Mode

By default the Tesseract engine expects a page of text when it processes an image. If you're just seeking to OCR a small region try a different segmentation mode, using the Page Segmentation Mode input parameter. Note that adding a small white border to text which is too tightly cropped may also help with page segmentation.

The different options available for the Page Segmentation Mode input are as follows:

Page Segmentation Mode settings	Description
OSD	Orientation and script detection (OSD) only.
AutoWithOSD	Automatic page segmentation with OSD.
AutoNoOCR	Automatic page segmentation, but no OSD, or OCR.
Auto	Fully automatic page segmentation, but no OSD. (Default)
Column	Assume a single column of text of variable sizes.
VerticalBlock	Assume a single uniform block of vertically aligned text.
Block	Assume a single uniform block of text.
Line	Treat the image as a single text line.
Word	Treat the image as a single word.
CircledWord	Treat the image as a single word in a circle.
Character	Treat the image as a single character.

If the output quality of 'Read Text With OCR' is not as expected the Page Segmentation should be changed to an appropriate setting for the text area being read. For example, if you are reading a text area that should contain a single line of text, change the setting for this parameter to "Line".

For further information on segmentation modes please consult the official documentation provided by Tesseract on their website.

Character Whitelist

- Used to restrict which characters can be recognised. For example, to ignore all non-numeric characters, enter “1234567890-”
- The order of characters does not matter, “1234567890” works as well as “0987123456”
- Make sure to include any special characters that maybe needed, e.g. . , $ ‘ – ()

Diagnostics Path

- Optional location for the output of what gets OCR’d. This is helpful for diagnostics problem solving if the OCR is not working as expected.
- Files in the output folder will be overwritten with each run

Scale

- This is how much the engine will zoom in to read the image. The default is 4 but a value between 8 and 12 will often provide better results. Going over 14 produces poorer results within a larger region of text.
- It is recommended that some experimentation is done with different values until the scale which returns the best results for your use case is found.

- When trying to get text from multiple columns, the scaling should be set to 10 or higher to maintain the text on a single row, otherwise the data maybe returned in a single column

Details

OCR is not intended as a replacement for Character Matching, and the Recognise Text action is still available. OCR and Character Matching are different recognition techniques and both have advantages and disadvantages.

Tips

- The OCR feature works best when there is a longer string and not one to three words
- Since terminal emulators used by mainframes are mono-spaced, continue using Character Matching and create your own font if necessary.
- Unlike Character Matching, OCR does not need font smoothing to be switched off
- Both methods require a clear view of the application screen

Limitations

- The engine is designed to work from 300dpi images and not screen prints (~100dpi), so this is not a complete replacement of Recognise Text
- OCR can result in a 'false positive' or a 'false negative'. An example of a false positive is when the OCR incorrectly determines that some text value exists on the screen, when in reality it does not. A false negative would be where OCR mistakenly decides that a value does not exist, when in fact it does.
- By contrast Character Matching is more deterministic, either there is a 100% match with the character shape or there is no match.
- Care should always be taken when using any OCR technology. Quality cannot be guaranteed in advance, and only through large scale testing of your specific use case will you know if the technology is suitable for your solution. Where possible Recognise Text should always be used instead.

Blue Prism Group plc, Centrix House, Crow Lane East, Newton-le-Willows, WA12 9UY, United Kingdom
Registered in England: Reg. No. 4260035. Tel: +44 870 879 3000. Web: www.blueprism.com