KID
Handling Scanned Files

Fabian Blasch

10.03.2022

Tesseract

Tesseract is an optical character recognition engine, it can be used to convert pictures into machine readable characters. The sourcecode is available on GitHub and fortunately the API was implemented in an R package.

First Steps

Tesseract offers the possibility to alter the engine based on the problem at hand. In the case of KIDs, we want to extract the position of the scale of each SRRI entry. Thus one parameter of great importance is the character white list. It allows us to narrow the classification into the characters or in this case digits handed over to tesseract via the white list.

To eliminate noise we first look at a snippet taken from one of Security's KIDs that only contains the scale and no text, this allows for the identification of problems related to the extraction of the scale without the surrounding noise.

```
# load KiDs
setwd("./../KIDs")
devtools::load_all()
```

```
## i Loading KIDs
```

```
# example pdf
setwd("./../../../KIDs/Security")
pd <- list.files(pattern = ".pdf")[1]

# first convert to png so we can use tesseract directly
pdftools::pdf_convert(pd, pages = 1, dpi = 600) -> img_f
```

```
## Converting page 1 to KID_Apollo_2_Global_Bond_3_1.png... done!
```

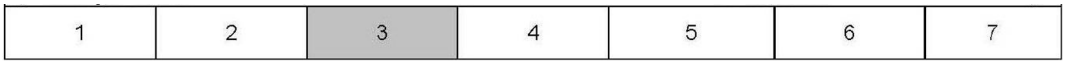
```
# set whitelist to digits 1-7 and whitespace
eng_spec <- tesseract::tesseract(options = list(tessedit_char_whitelist = " 1234567"))

# image that only contains the scale and no text
tesseract::ocr_data("Scale_notext.jpeg", engine = eng_spec) |> knitr::kable()
```

word	confidence	bbox
1	0.000000	97,5,165,84
2	88.425575	342,5,402,85
4	7.978073	816,5,869,84
5	60.284004	1061,5,1121,84
6	77.367195	1313,5,1366,85
7	0.000000	1528,5,1574,85

The readout worked good for all digits on white background, next we may use other engine parameters as well as image pre-processing to obtain the scale in its entirety. First off, we will try to convert to black and white to correctly identify the entire scale. The R package magick offers powerful tools for image processing.

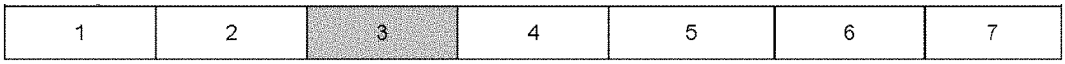
```
# reset
setwd("./../../../KIDs/Security")

# image
img_m <- magick::image_read("Scale_notext.jpeg")

# open image for a closer look
```

```
# magick::image_browse(img_m)

# plot image
par(mar = c(1, 4, 1, 2))
plot(img_m)
```



Now to switch from grayscale to a true black and white there are two different options. We can either use the function image_convert() or image_threshold().

```
# convert
img_m_conv <- magick::image_convert(img_m, type = "Bilevel")

# display
# plot image
par(mar = c(1, 4, 1, 2))
plot(img_m_conv)
```




```
# repeat ocr
tesseract::ocr_data(img_m_conv, engine = eng_spec)|> knitr::kable()
```

word	confidence	bbox
1	44.12727	106,5,159,85
2	88.78406	338,5,406,85
3	37.03163	578,5,631,85
4	53.32304	818,5,871,85
5	73.43266	1058,5,1119,85
6	82.94523	1313,5,1366,85
7	0.00000	1523,5,1569,85

As visible from the table we are now able to detect the entire scale, on top of that the classification confidence increased significantly. Lets see if we can increase the confidence by using different thresholds when converting to black and white.

```
# alter with varying thresholds
img_m_thresh <- lapply(paste0(seq(10, 90, 10), "%"),
                       \(x) magick::image_threshold(img_m, threshold = x, type = "white"))
# align
par(mfrow = c(9, 1), mar = c(1, 4, 1, 2))

# plot
lapply(img_m_thresh, \(x) plot(x)) |> invisible()
```

For the plots above the threshold starts art 10% and increases to 90% in increments of 10%.In theory we want to choose a threshold that keeps the digits as legible as possible while converting the shade to white. By observing the plots we know that this threshold has to lie somewhere between 70% and 80%. We have to keep in mind however, that this will vary across different KIDs, thus this parameter will have to be determined via cross validation at a later stage, should it perform better than bi-level processing.

```
lapply(img_m_thresh,
       \(x) tesseract::ocr_data(x, engine = eng_spec) |> knitr::kable())
```

```
## [[1]]
##
##
## |word | confidence|bbox |
## |:----|----------:|:----|
##
## [[2]]
##
##
## |word | confidence|bbox           |
## |:----|----------:|:--------------|
## |1    |   37.47584|98,4,159,85    |
## |2    |   90.69081|338,5,399,85   |
## |3    |   66.06633|571,4,624,85   |
## |4    |   71.54774|811,4,871,85   |
## |5    |   67.67908|1058,4,1119,85 |
## |6    |    0.00000|1306,4,1366,85 |
##
## [[3]]
##
##
## |word | confidence|bbox           |
## |:----|----------:|:--------------|
## |1    |    0.00000|90,5,157,85    |
## |2    |   88.71307|334,5,402,85   |
## |3    |   63.00784|579,5,632,85   |
## |4    |   86.19589|816,5,877,85   |
## |6    |   67.12558|1061,5,1121,85 |
## |6    |   90.62632|1313,5,1366,85 |
## |7    |    0.00000|1521,5,1574,85 |
##
## [[4]]
##
##
## |word | confidence|bbox           |
## |:----|----------:|:--------------|
## |1    |   9.509064|97,5,165,85    |
## |2    |  77.638237|342,5,402,85   |
## |3    |  74.381256|579,5,632,85   |
## |4    |  71.474670|816,5,877,85   |
## |6    |  47.728615|1061,5,1121,85 |
## |6    |  82.772720|1313,5,1366,85 |
## |7    |   0.000000|1484,5,1574,85 |
##
## [[5]]
##
##
## |word | confidence|bbox          |
## |:----|----------:|:-------------|
## |1    |   11.63204|97,5,165,85   |
## |2    |   59.56064|342,5,402,85  |
## |3    |    0.00000|572,5,639,84  |
##
## [[6]]
```

```
##
##
## |word | confidence|bbox           |
## |:----|----------:|:--------------|
## |1    |    0.00000|97,5,157,84    |
## |2    |   90.05758|334,5,402,85   |
## |3    |   82.89348|572,5,632,84   |
## |4    |   83.86511|816,5,877,84   |
## |6    |   72.44370|1061,5,1121,84 |
## |6    |   90.63850|1313,5,1366,85 |
## |7    |    0.00000|1521,5,1581,85 |
##
## [[7]]
##
##
## |word | confidence|bbox           |
## |:----|----------:|:--------------|
## |1    |   36.78156|98,5,166,84    |
## |2    |   57.74221|328,5,403,85   |
## |3    |   22.47987|565,5,633,84   |
## |4    |    0.00000|810,5,870,84   |
## |6    |   13.63799|1307,5,1367,84 |
##
## [[8]]
##
##
## |word | confidence|bbox           |
## |:----|----------:|:--------------|
## |1    |    0.00000|90,5,157,84    |
## |2    |   90.30231|334,5,402,85   |
## |3    |   80.03242|572,5,632,84   |
## |4    |   54.35625|816,5,869,84   |
## |56   |   63.68980|1061,1,1114,90 |
## |6    |   75.67296|1313,5,1366,85 |
## |7    |    0.00000|1521,5,1581,85 |
##
## [[9]]
##
##
## |word | confidence|bbox           |
## |:----|----------:|:--------------|
## |1    |   39.07653|97,5,157,84    |
## |2    |   91.19086|334,5,402,85   |
## |3    |   58.06713|579,5,632,84   |
## |4    |   56.51816|816,5,869,84   |
## |56   |   59.64135|1061,1,1114,90 |
## |6    |   81.46667|1313,5,1366,85 |
## |7    |    0.00000|1521,5,1581,85 |
```

Even though the threshold approach did not significantly improve the confidence in identification, it offers a parameter for optimization at a later stage. Accordingly, to be able to tell for sure whether an additional CV parameter is worth the computational complexity will be determined at a later stage.

Performance in untrimmed files

Cropping an image drastically reduces noise, accordingly removing less of the noise will most likely come with a few problems. Before evaluating which options for cropping are realistically executable, we will explore the performance without cropping.

```
# reset
setwd("./../../../KIDs/Security")

# image
img_fm <- magick::image_read(img_f)

# unaltered ocr
ocr_full_u <- tesseract::ocr_data(img_f, engine = eng_spec)

# bi-level
biL_img_f <- magick::image_convert(img_fm, type = "Bilevel")

# bi-level altered ocr
ocr_full_a <- tesseract::ocr_data(biL_img_f, engine = eng_spec)

# return
ocr_full_a |> knitr::kable()
```

word	confidence	bbox
6	0	483,557,763,639
2	0	1273,1480,1313,1540
74612	0	478,1622,1107,1680
56711	0	1131,1623,1575,1672
16	0	2237,1750,2609,1812
1	0	2655,1750,2797,1799
2	0	473,2052,657,2113
2	0	1583,2194,1858,2244
2	0	3828,2462,3930,2499
2	0	1426,2971,1493,3008
2	0	2562,3239,2641,3287
3	0	2822,3239,2901,3288
2	0	1042,3315,1596,3378
2	0	2368,3316,2438,3365
112	0	2464,3316,2600,3364
265	0	2018,3442,2448,3492
4	0	499,3783,635,3839
4	0	495,3865,635,3921
2	0	1844,4473,1954,4522
1	0	3305,4473,3323,4521
2	0	2292,4676,2678,4725
2	0	474,4880,682,4929
6	0	1067,4955,1442,5005
6	0	1728,5083,1939,5132
6	0	924,5209,1637,5272
2	0	3916,5336,4024,5384