ADAM MICKIEWICZ UNIVERSITY IN POZNAŃ

Faculty of Mathematics and Computer Science

Extracting text from silly drawings

Text localisation and OCR in unstructured graphical documents

August 5, 2023

Karolin Izabel Boczoń

s434659.students.wmi.amu.edu.pl

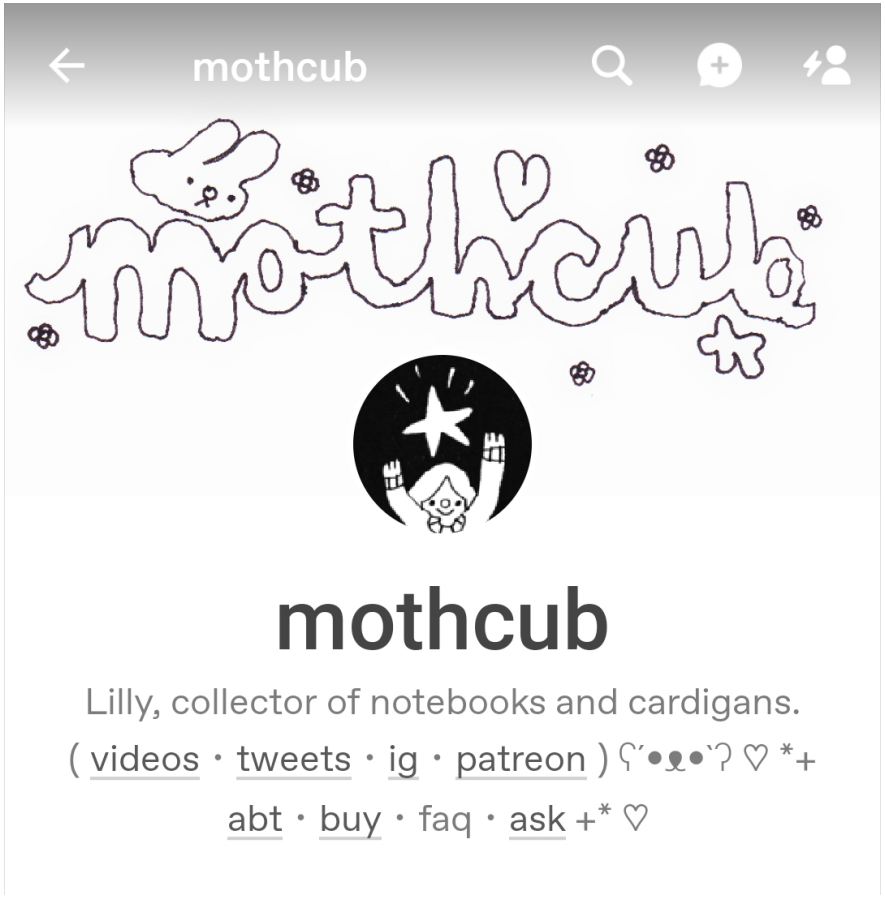
Motivation

The idea was to catalogue silly drawings based on the text message written on it for easier sharing with others.

- fun little project
- image processing
- uniform dataset

Credits

Lilly Ashton, known as *mothcub*, is an Internet artist. She makes artworks, videos and music.

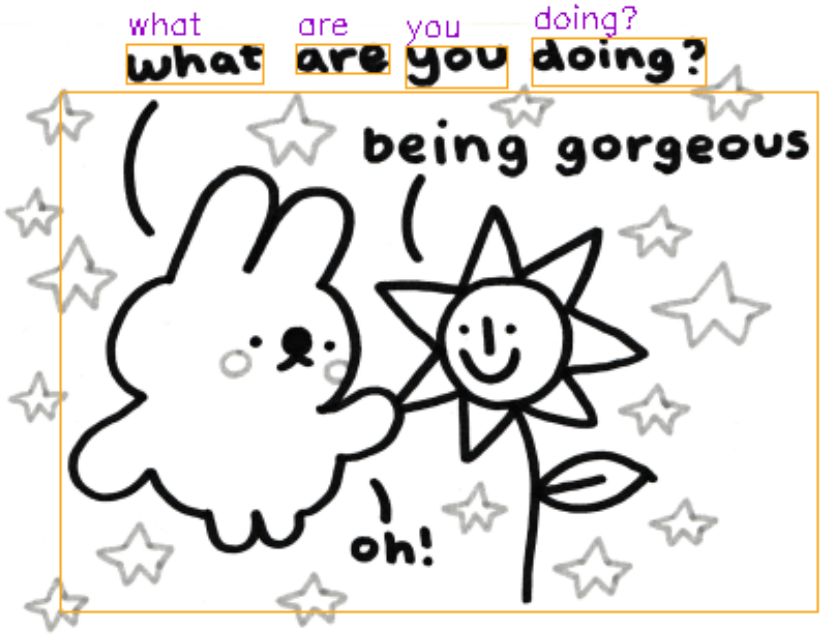


First attempts

OCR alone doesn't do well with graphical documents, mistaking decorative elements for text.

First attempts

Submitting text-only images to the recognition system improves the results immediately, yet it requires additional image processing.



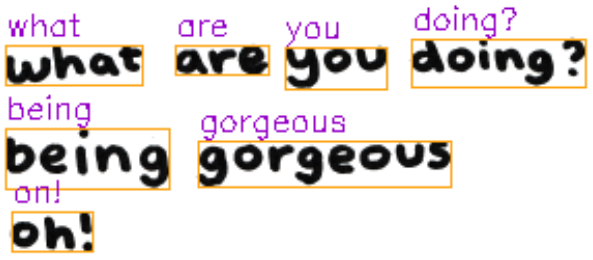
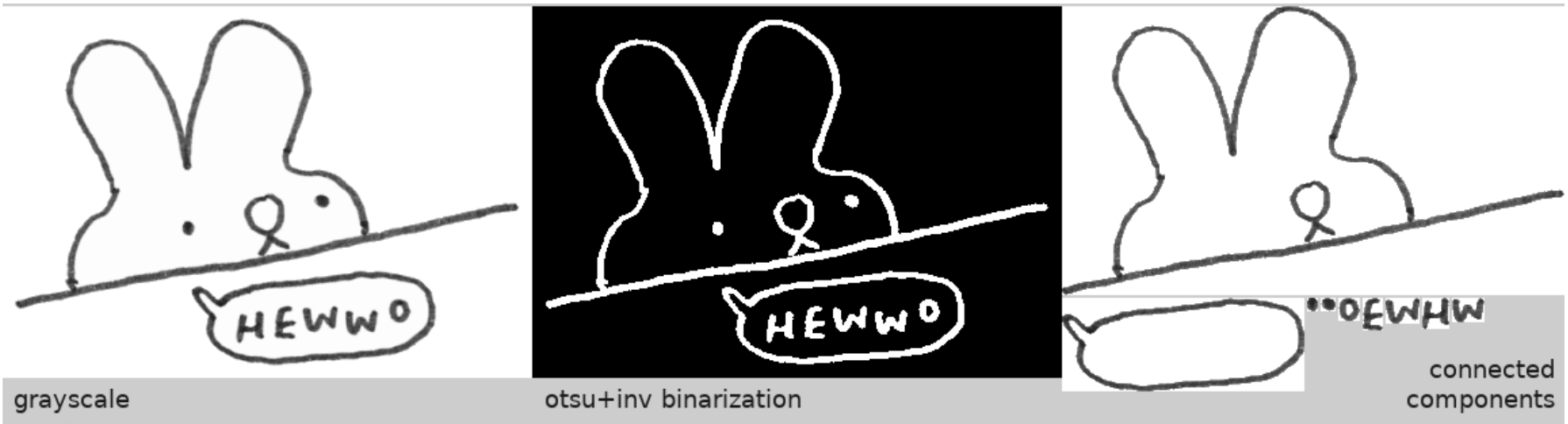


Image segmentation

Since most of the collected images are hand drawn with a felt-tip pen on a white background and the elements never overlap, we can easily separate them.

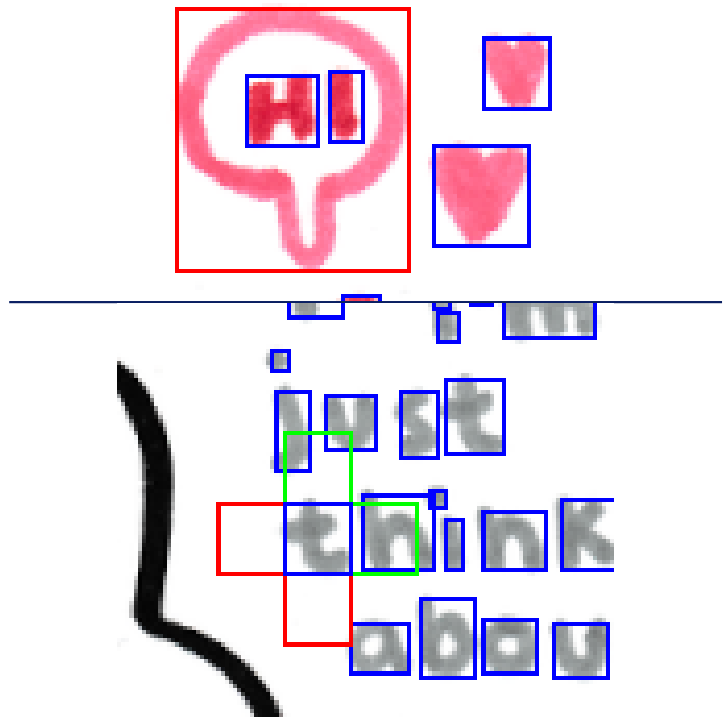


Text/graphic separation

Then we classify each component as either text or non-text CC: letter or decorator.

Rules:

- a letter bounding box can't contain another CC
- each letter is surrounded by other letters - CCs similar in height
- ?

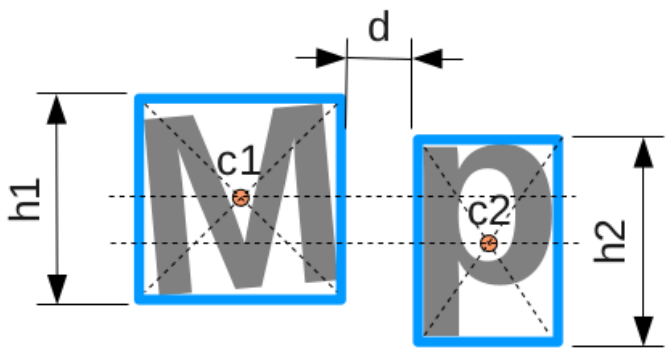


Text localisation

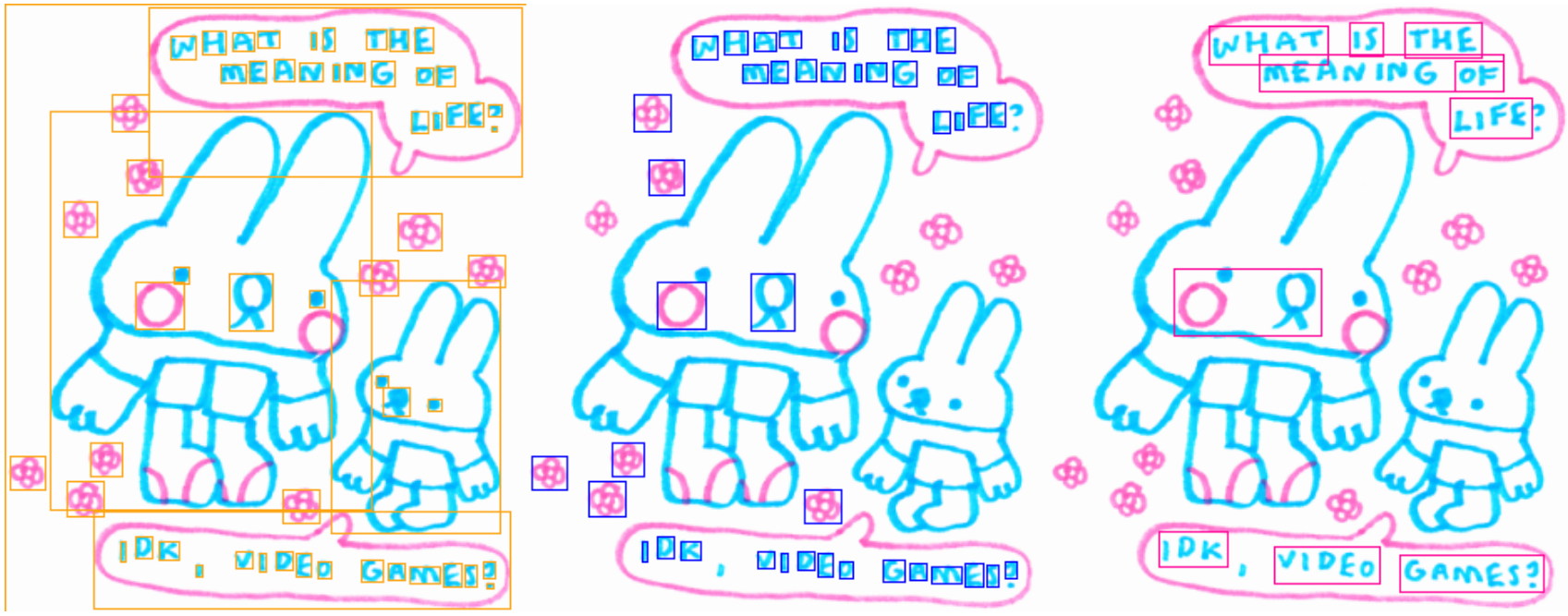
The drawings are unstructured documents so the text can be placed anywhere in the image.

Recipe:

1. begin with the leftmost letter for each line
2. then add letters on the right if
   - $d < max(h1, h2)$
   - $y_{min}(l_1) < c2.y < y_{max}(l_1)$



Did it work?



method	CER	WER
OCR_alone	62,13	90,33
with_text_extr	55,07	77,32
with_text_extr_no_punct	53,70	74,25

What's next?

Improvements:

- better binarization method for problematic images
- masking non-text CCs when cropping ROI
- combining ROIs into one image (?)

Future work:

- concept tagging from text

References

- Rigaud C. et al., Automatic Text Localisation in Scanned Comic Books
- Rosebrock A., Tesseract OCR: Text localization and detection
- mothcub.tumblr.com