Image to Text

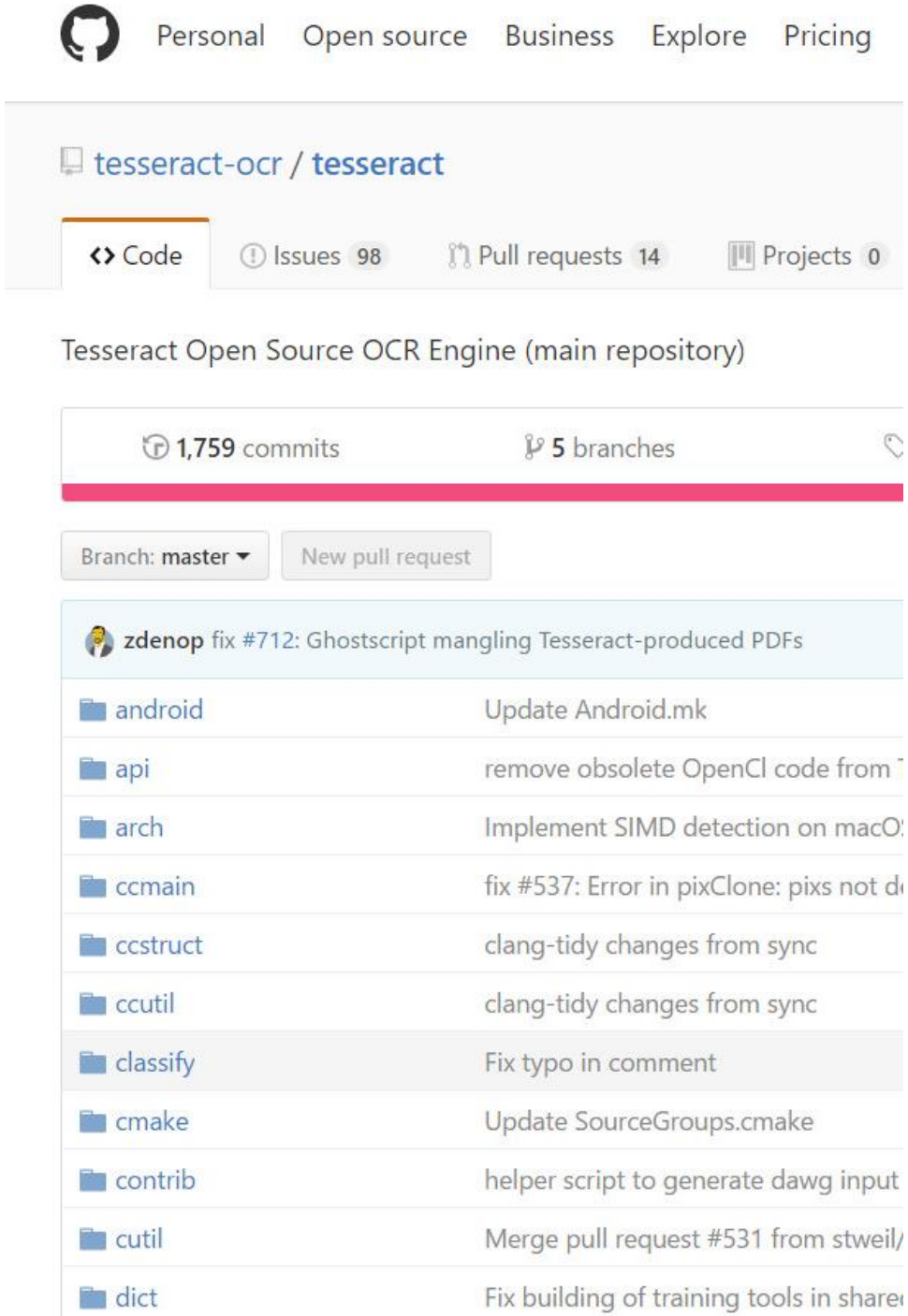
Oswell Chan, Software Developer @ Acronis

WHAT IS

Optical Character Recognition

the mechanical or electronic conversion of images of typed, handwritten or printed text into machine-encoded text

TESSERACT OCR
PACKAGE



Developers

Lead Developer: Ray Smith

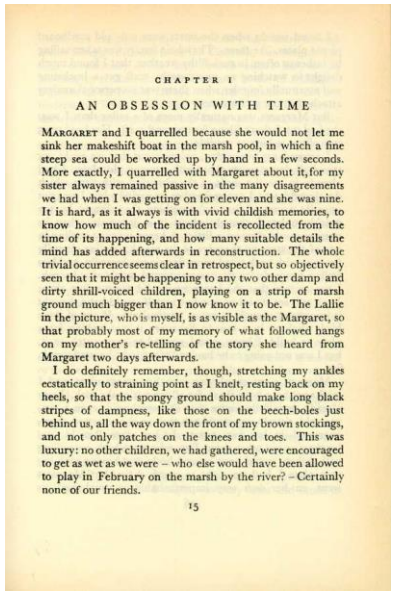
Maintainer: Zdenko Podobny

Package

libtesseract (OCR Engine)

tesseract (cmd-line program)

TESSERACT OCR
CMD-LINE



CHAPTER I

AN OBSESSION WITH TIME

MARGARET and I quarrelled because she would not let me sink her makeshift boat in the marsh pool, in which a fine steep sea could be worked up by hand in a few seconds. More exactly, I quarrelled with Margaret about it, for my sister always remained passive in the many disagreements we had when I was getting on for eleven and she was nine. It is hard, as it always is with vivid childish memories, to know how much of the incident is recollected from the time of its happening, and how many suitable details the mind has added afterwards in reconstruction. The whole trivial occurrence seems clear in retrospect, but so objectively seen that it might be happening to any two other damp and dirty shrill-voiced children, playing on a strip of marsh ground much bigger than I now know it to be. The Lallie in the picture, who is myself, is as visible as the Margaret, so that probably most of my memory of what followed hangs on my mother's re-telling of the story she heard from Margaret two days afterwards.

I do definitely remember, though, stretching my ankles ecstatically to straining point as I knelt, resting back on my heels, so that the spongy ground should make long black stripes of dampness, like those on the beech-boles just behind us, all the way down the front of my brown stockings, and not only patches on the knees and toes. This was luxury: no other children, we had gathered, were encouraged to get as wet as we were – who else would have been allowed to play in February on the marsh by the river? – Certainly none of our friends.

15

CHAPTER I
AN OBSESSION WITH TIME

WRGARET and I quarrelled because she would not let me sink her makeshift boat in the marsh pool, in which a fine steep sea could he worked up by hand in a Few seconds, More exactly, I quarrelled with Margaret about it,ft>r my sister always remained passive in the many disagreements we had when I was getting on for eleven and she was nine It is hard, as it always is with vivid childish memories, to know how much of the incident is recollected from the time ofits happening, and how many suitahle details the mind has added aftermrds in recanstrucfion. The whole trivial oeenrrenceseems clear in retrospect, but so objectively seen that it might be happening to my two other damp and dirty shrill-voiced children, playing on a strip or marsh ground much bigger than I now know it to be. The Lallic in the picture. who is mysell, is as visihle as the Margaret, so dint probably most of my memory or what fallowed hangs on my mod-ler's re-telling of the story she heard from Margaret two days afterwards.

I do definitely remember, though, suetching my ankiu ccstatieally to straining point as I knelt, resting back on my heels, so that the spongy ground should make long black stripes of dampness, like those on the heech-holes just behind us, all the way down the front ofmy hrewn stoeltings, and not only patches on the knees and toes. This was luxury: no other children, we had gathered, were encouraged to get as wet as we were — who else would have heen allowed to play in February on the marsh by Ihe river? -Certainiy mm: at" our (riends.

15

```
tesseract robertson.jpg stdout
```

TO CLEAN OR NOT TO CLEAN

CON"! [Ni/III) FROIH PAGE [52

that the other person won't pick up his socks or hang up her coat," she says. "It's often that it isn't happening as quickly as the other person thinks is acceptable."

Experience has taught the Weisses the futility of trying to convert each other. "She has come around to accepting things, and now she will ask before she throws anything out," says Richardjonathan Parker, who's been back with his girlfriend for a year, has "given up on retraining. Now I try to see things from her point of view. But I try not to pick up after her as much because I don't want to encourage the behavior."

Getting frustrated, getting mad —what's the point? Getting a housekeeper—that's the answer. "For years I did matrimonial law," says Parker, "and I became convinced that one of the best ways to avoid divorce was to hire someone to come in and clean. It's working for us. It really is."

fonnne Kaufman it a frequent contributor In The Wall Streetjournal and The New York Times.

pretend. Remaining attached to some peo-ple is like slaving over a withered garden without realizing all the plants are dead. And letting the negative relationships suck up time and energy only deprives us of the opportunity to nurture and appreciate those friendships that actually do work.

"The interchange between us captured my mind: conversations and joking, doing favors for each other, reading together good books, being foolish and being serious to—gether, disagreeing without hatred almost as though I was debating with myself," wrote St. Augustine of Hippo, that ancient arbiter of relationships, in Confriiom, 1,6oo years ago. "These and so many other like signs coming from the hearts of friends are shown through their eyes and mouths and speech and a thousand little gestures. All of these expressions of friendship brought our hearts together like bundled kindling, mak—ing one out of many"

The truth never fails to show itself in those "thousand little gestures." True friends bear each other's burdens and ac—cept each other's weaknesses, because without reciprocity, there's nothing. Friendship is about collaboration, not dom-ination. Because we should be stewards of each other's rooms, I am happy to help you keep yours clean, but life is too fleeting to let you continue trashing mine.

Paige Ilfilliam: it nfieelnnce writer and the executive editor o f B oston magazine.

```
tesseract opera-scan.jpg out.txt
```

TESSERACT OCR
CMD-LINE

```
tesseract --print-parameters
```

editor_image_xpos 590 Editor image X Pos
editor_image_ypos 10 Editor image Y Pos
editor_image_menuheight 50 Add to image height for menu bar
editor_image_word_bb_color 7 Word bounding box colour

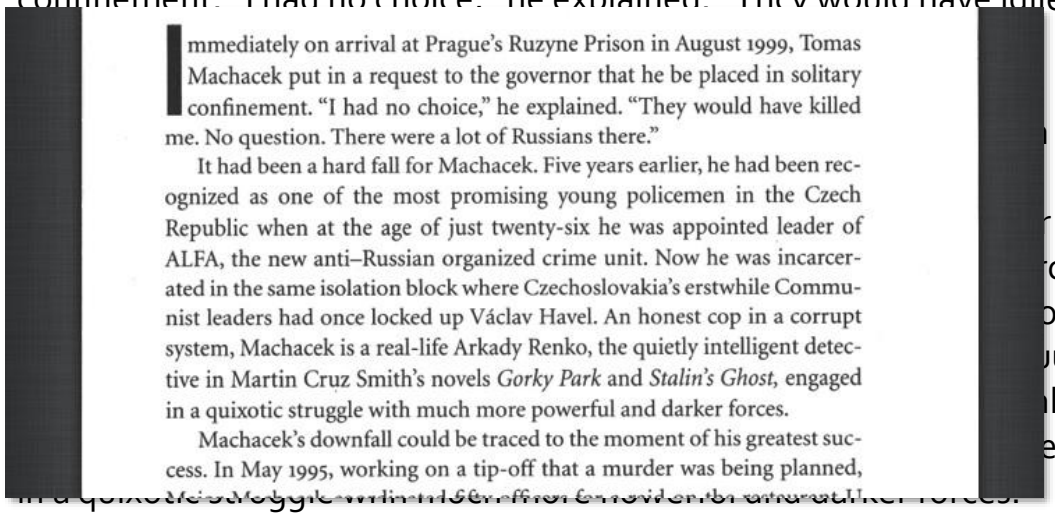
. . .

textord_noise_rowratio 6 Dot to norm ratio for deletion
textord_blshift_maxshift 0 Max baseline shift
textord_blshift_xfraction 9.99 Min size of baseline shift

678 parameters!

```
Usage:
  tesseract --help | --help-psm | --version
  tesseract  --list-langs [--tessdata-dir PATH]
  tesseract  --print-parameters [options...] [configfile...]
  tesseract  imagename|stdin outputbase|stdout [options...] [configfile...]

OCR options:
  --tessdata-dir PATH   Specify the location of tessdata path.
  --user-words PATH     Specify the location of user words file.
  --user-patterns PATH  Specify the location of user patterns file.
  -l LANG[+LANG]        Specify language(s) used for OCR.
  -c VAR=VALUE          Set value for config variables.
                        Multiple -c arguments are allowed.
  -psm NUM              Specify page segmentation mode.
NOTE: These options must occur before any configfile.

Page segmentation modes:
  0    Orientation and script detection (OSD) only.
  1    Automatic page segmentation with OSD.
  2    Automatic page segmentation, but no OSD, or OCR.
  3    Fully automatic page segmentation, but no OSD. (Default)
  4    Assume a single column of text of variable sizes.
  5    Assume a single uniform block of vertically aligned text.
  6    Assume a single uniform block of text.
  7    Treat the image as a single text line.
  8    Treat the image as a single word.
  9    Treat the image as a single word in a circle.
 10    Treat the image as a single character.

Single options:
  -h, --help            Show this help message.
  --help-psm            Show page segmentation modes.
  -v, --version         Show version information.
  --list-langs          List available languages for tesseract engine.
  --print-parameters    Print tesseract parameters to stdout.
```

TESSERACT OCR
CMD-LINE



mmediately on arrival at Prague's Ruzyne Prison in August 1999, Tomas Machacek put in a request to the governor that he be placed in solitary confinement. "I had no choice," he explained. "They would have killed me. No question. There were a lot of Russians there."

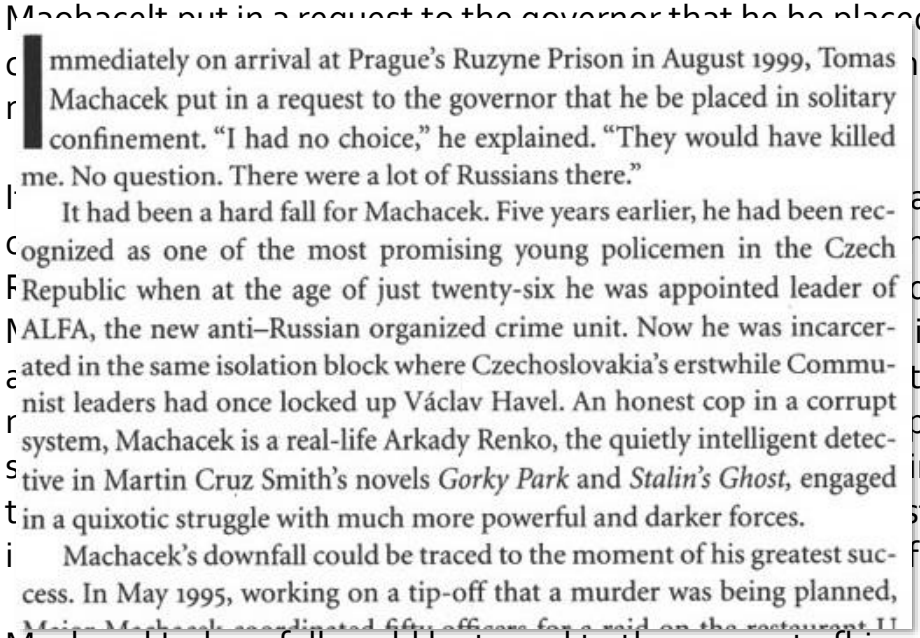
It had been a hard fall for Machacek. Five years earlier, he had been recognized as one of the most promising young policemen in the Czech Republic when at the age of just twenty-six he was appointed leader of ALFA, the new anti–Russian organized crime unit. Now he was incarcerated in the same isolation block where Czechoslovakia's erstwhile Communist leaders had once locked up Václav Havel. An honest cop in a corrupt system, Machacek is a real-life Arkady Renko, the quietly intelligent detective in Martin Cruz Smith's novels *Gorky Park* and *Stalin's Ghost*, engaged in a quixotic struggle with much more powerful and darker forces.

Machacek's downfall could be traced to the moment of his greatest success. In May 1995, working on a tip-off that a murder was being planned,

mmediately on arrival al Prague's Ruzy'nc Prlsun m Augusr 1999, Tomas Machacck pul in a requesl to lhe governor rhal he he placed ln selllary confinement. "I had no choice," he explained. "Thcy would have ldlled

Maehaeele's downfall could be rraeed lo lhe moment ofhis grearesl sue.

eess. ln May ms, worklng on a tipeoff ehal a murder was hung planned
a la



mmediately on arrival at Prague's Ruzyne Prison in August 1999, Tomas Machacek put in a request to the governor that he be placed in solitary confinement. "I had no choice," he explained. "They would have killed me. No question. There were a lot of Russians there."

It had been a hard fall for Machacek. Five years earlier, he had been recognized as one of the most promising young policemen in the Czech Republic when at the age of just twenty-six he was appointed leader of ALFA, the new anti–Russian organized crime unit. Now he was incarcerated in the same isolation block where Czechoslovakia's erstwhile Communist leaders had once locked up Václav Havel. An honest cop in a corrupt system, Machacek is a real-life Arkady Renko, the quietly intelligent detective in Martin Cruz Smith's novels *Gorky Park* and *Stalin's Ghost*, engaged in a quixotic struggle with much more powerful and darker forces.

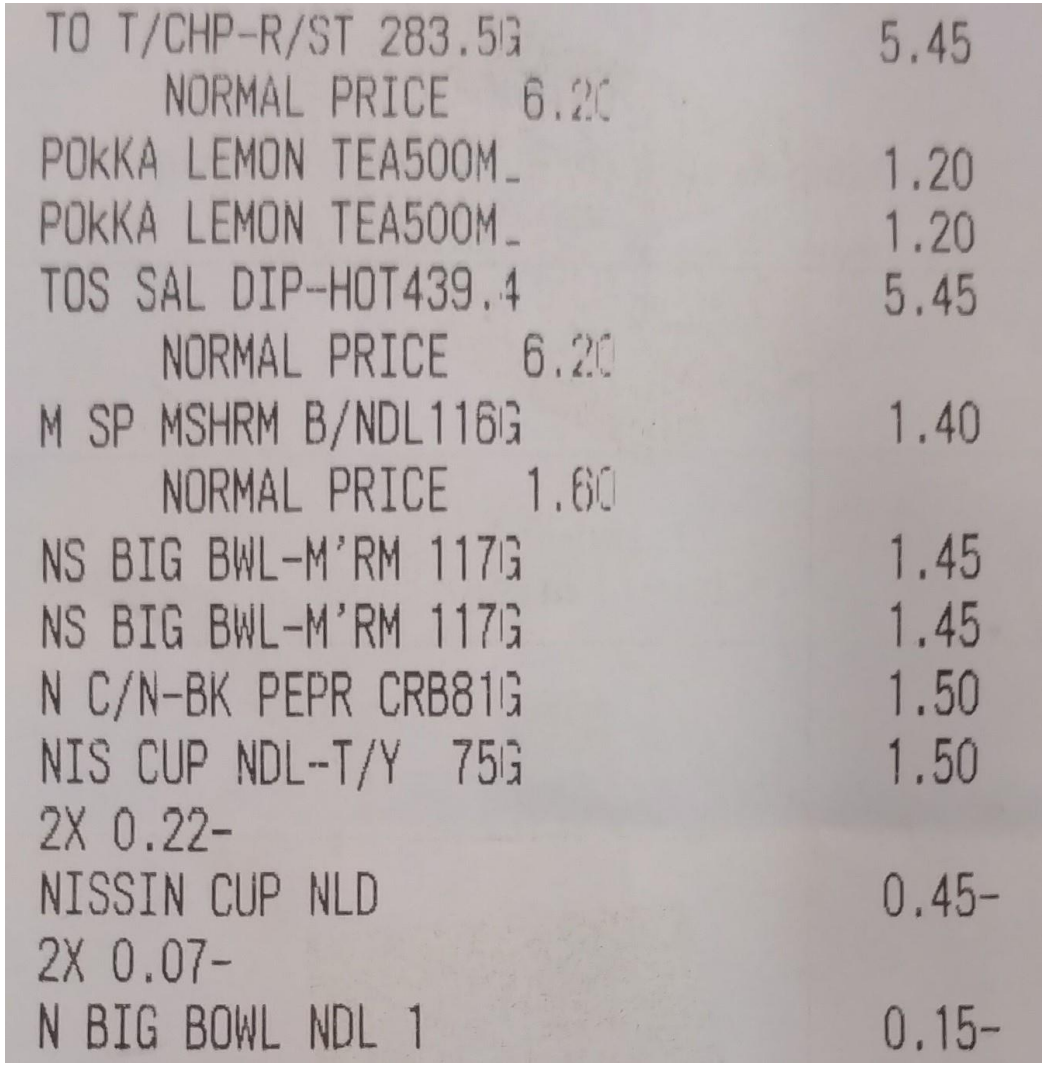
Machacek's downfall could be traced to the moment of his greatest success. In May 1995, working on a tip-off that a murder was being planned,

mmediately on arrival at Praguz's Ruzyne Prison in August iooo, Tomas

Machacek's downfall eonld be traced to the mommt ofhis greatest sue.

aess. in May ms, working on a tipotf (ha! a murder was being planned,
.. a

TESSERACT OCR
CMD-LINE



TO T/CHP—R/ST 283.5G
NORMAL PRICE 6.2C
POkKA LEMON
TEASOOM-
POkKA LEMON
TEASOOM-
TOS SAL DIP~HOT439.4
NORMAL PRICE 6.2K
M SP MSHRM
B/NDL116G
NORMAL PRICE 1.66
NS BIG BNL-M'RM 1176
NS BIG BNL-M'RM 1175
N C/N—BK PEPR CR881G
NIS CUP NDL~T/Y 75G
2X 0.22-
NISSIN CUP NLD
2X 0.07-
N BIG BONL NDL 1

TESSERACT OCR
TESSEROCR

Tesserocr (2.1.3)

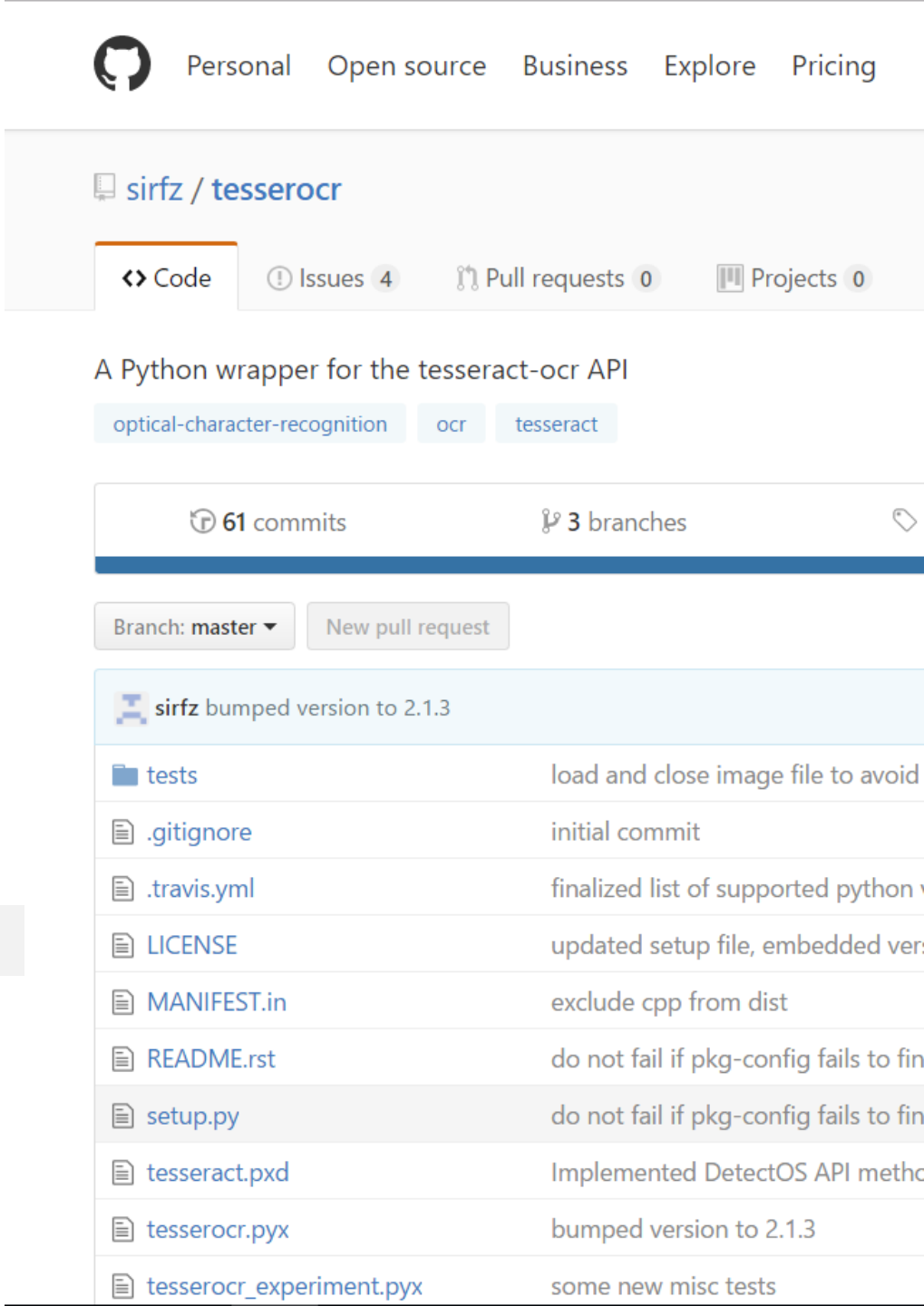
i. libtesseract (>=3.04)

ii. libleptonica (>=1.71)

iii. Cython (0.25.2)

```
$ CPPFLAGS=-I/usr/local/include pip install tesserocr
```

```
$ python setup.py build_ext -I/usr/local/include
```



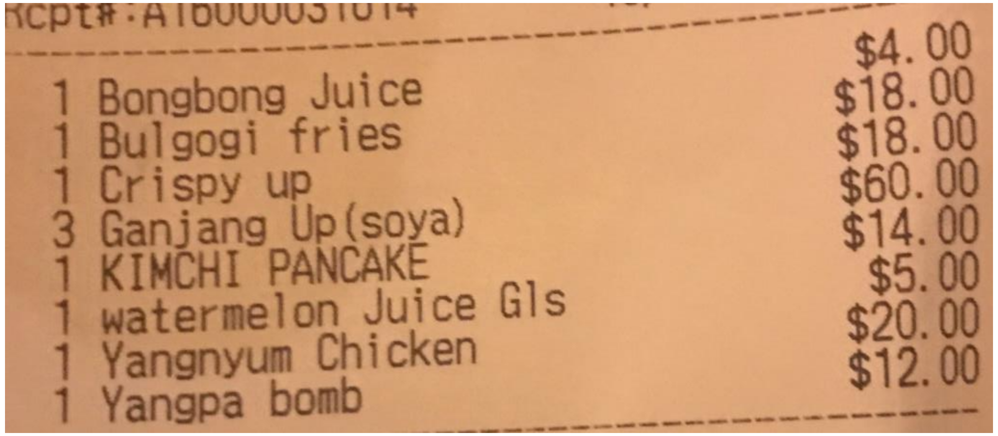
TESSERACT OCR
TESSEROCR

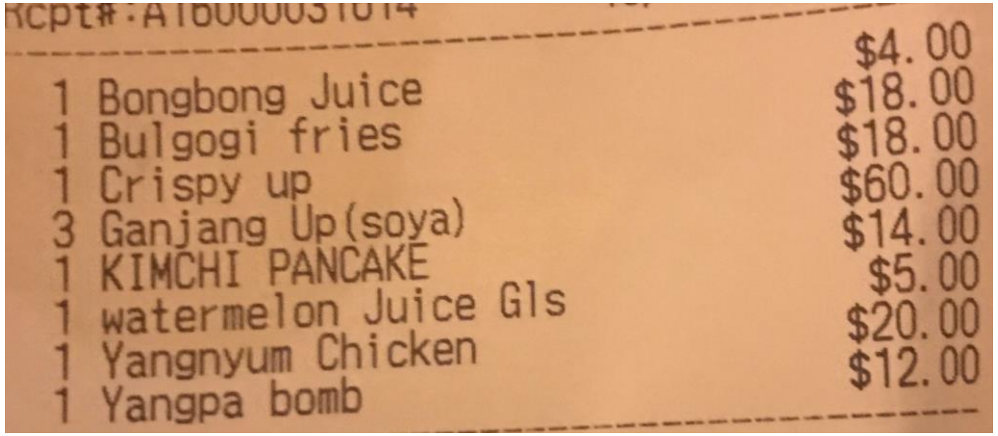
```
from PIL import Image
from tesserocr import PyTessBaseAPI, RIL

image = Image.open('./test/cropped/test21.jpg')
with PyTessBaseAPI() as api:
    api.SetImage(image)
    boxes = api.GetComponentImages(RIL.TEXTLINE, True)
    print('Found {} textline image components.'.format(len(boxes)))
    for i, (im, box, _, _) in enumerate(boxes):
        # im is a PIL image object
        # box is a dict with x, y, w and h keys
        api.SetRectangle(box['x'], box['y'], box['w'], box['h'])
        ocrResult = api.GetUTF8Text()
        conf = api.MeanTextConf()
        print((u"Box[{0}]: x={x}, y={y}, w={w}, h={h}, "
               "confidence: {1}, text: {2}").format(i, conf, ocrResult, **box))
```

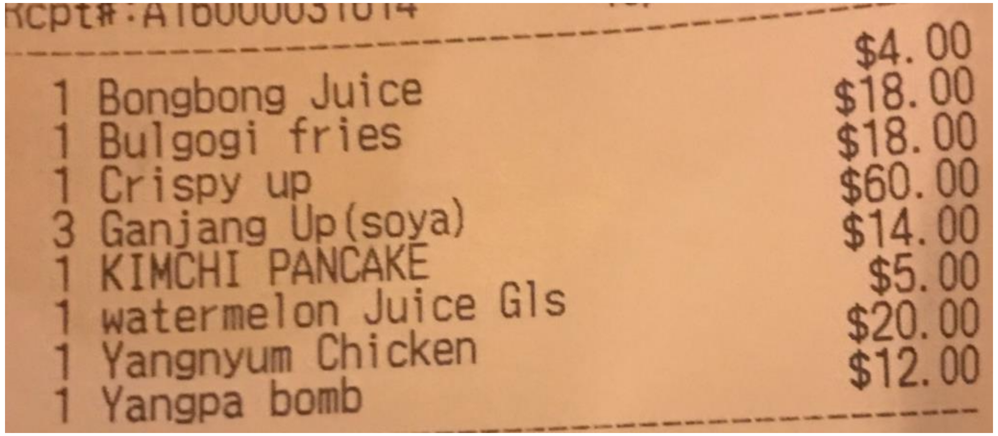
IMPROVING RESULTS
SETTING UP

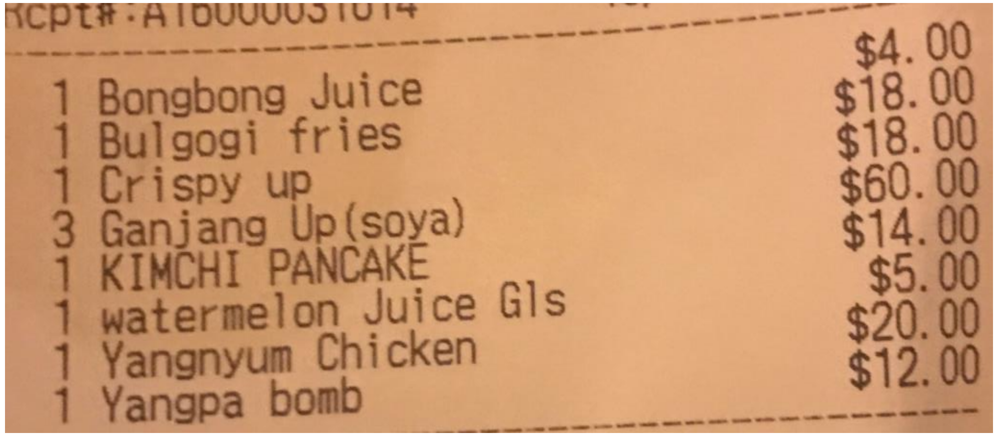
1. Pillow (4.0.0)
2. NumPy (1.12)
3. opencv-python (3.2.0.6)
4. SciPy (0.18.1)

IMPROVING RESULTS
ORIENTATION

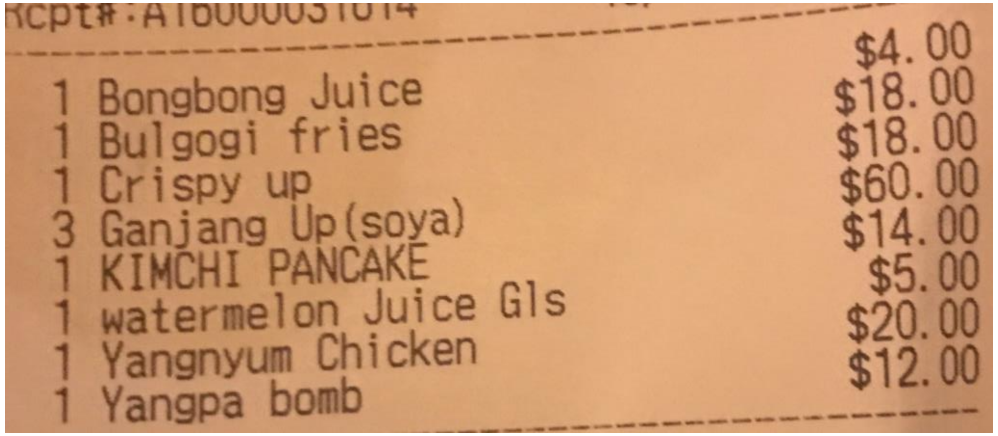


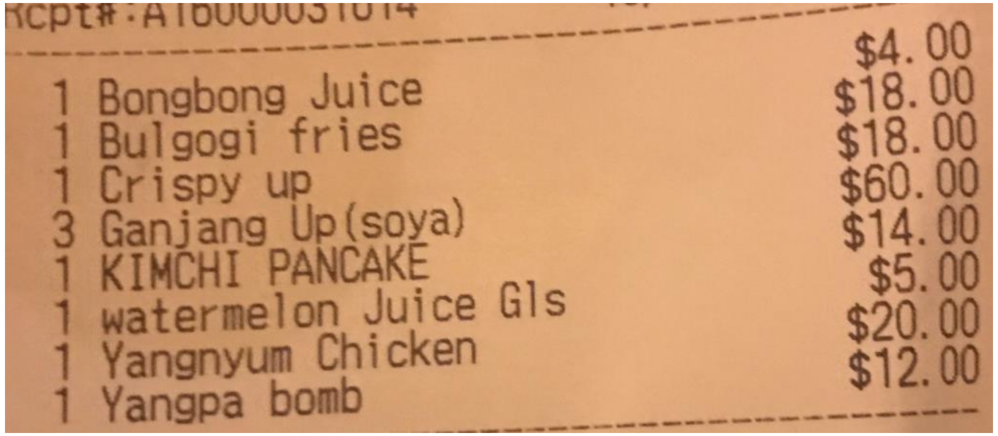


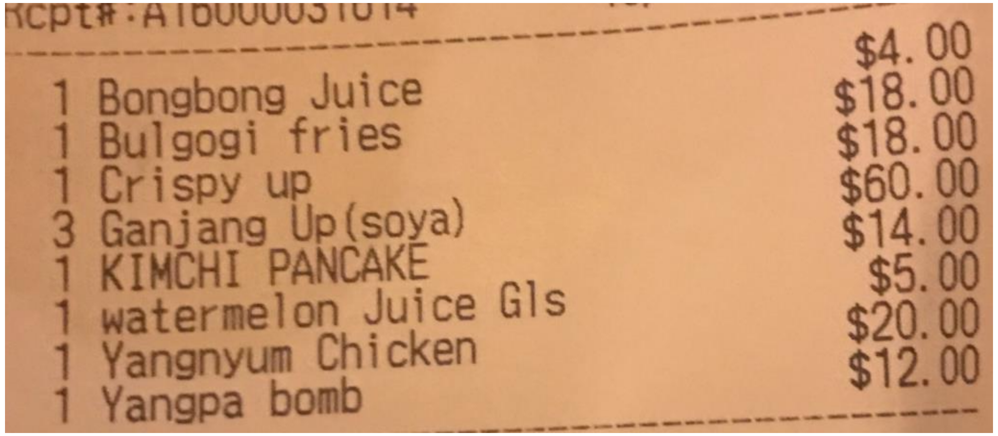


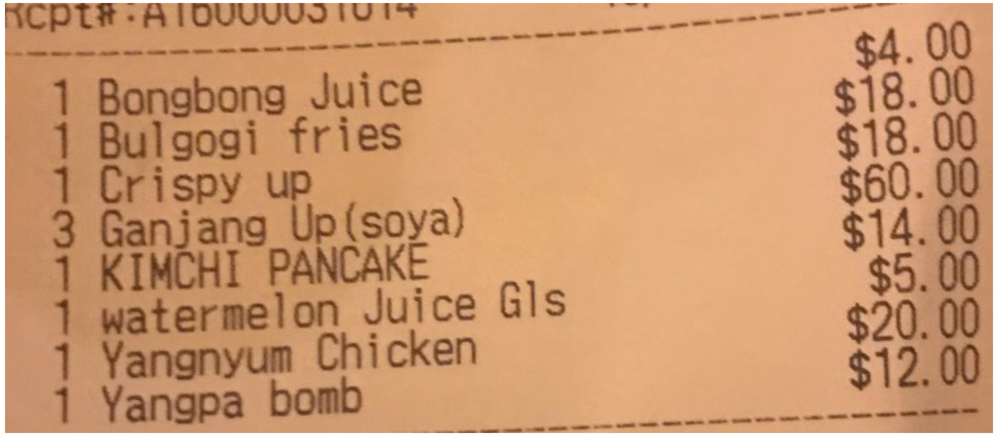


IMPROVING RESULTS
ORIENTATION

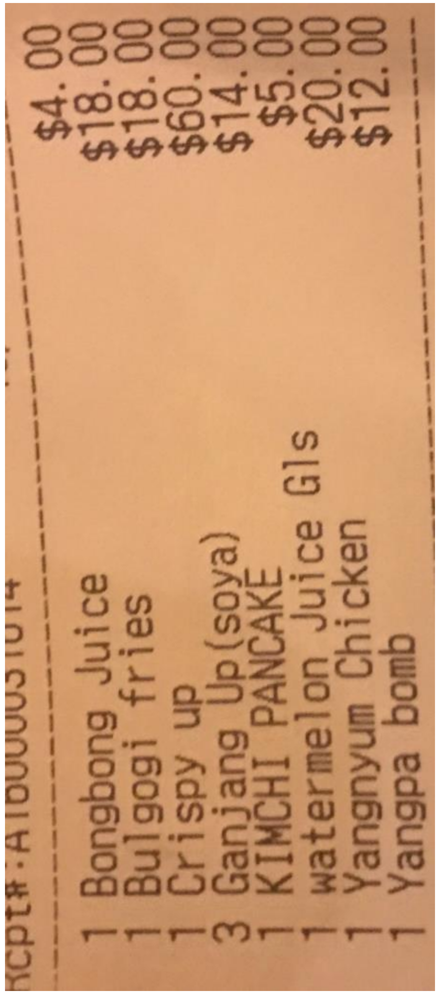


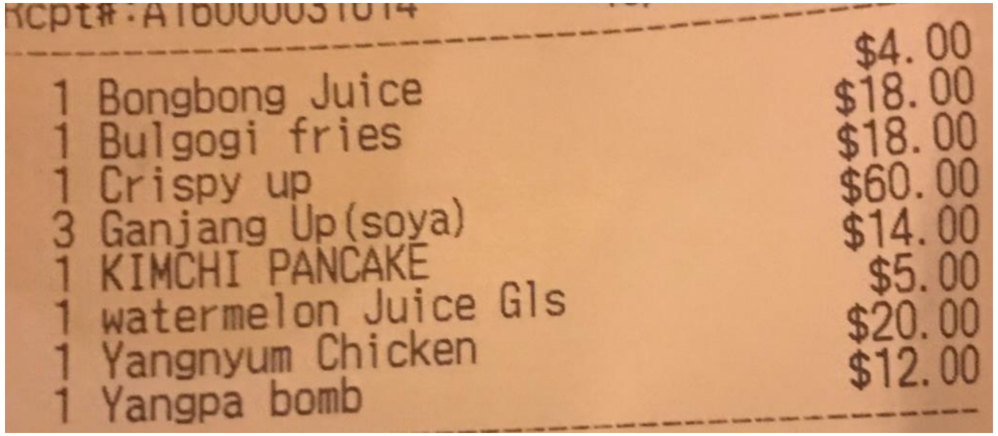


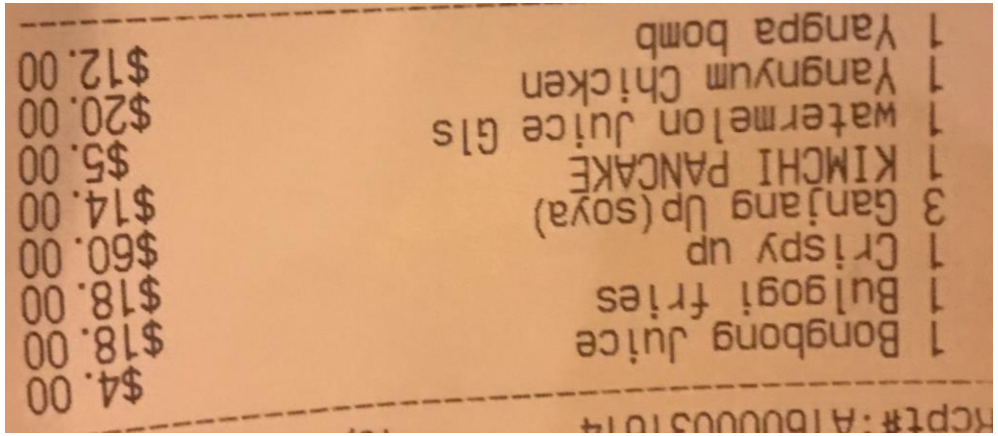


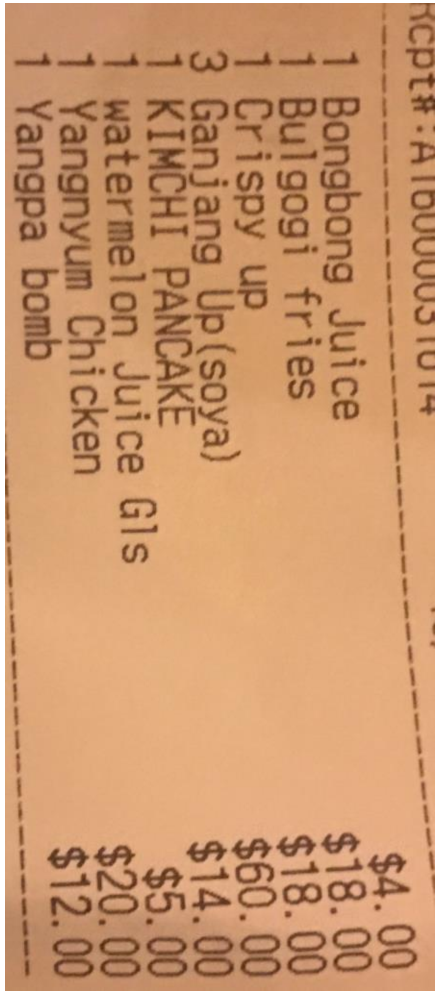


IMPROVING RESULTS
ORIENTATION







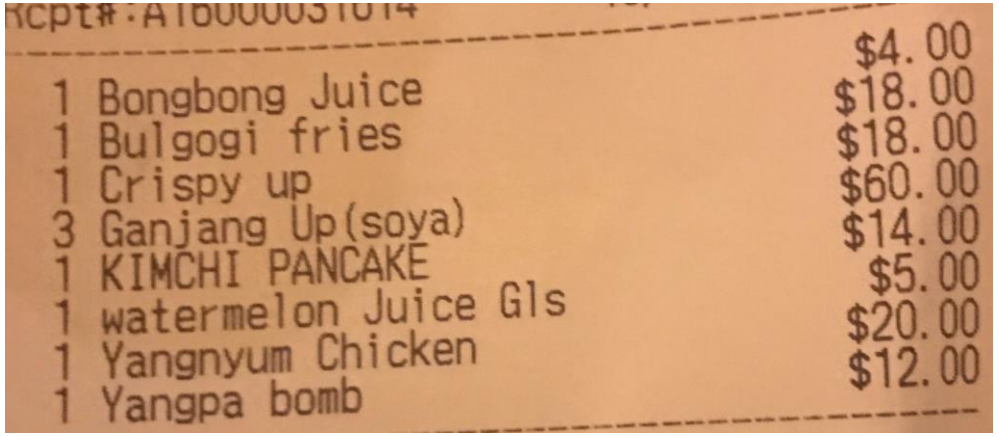


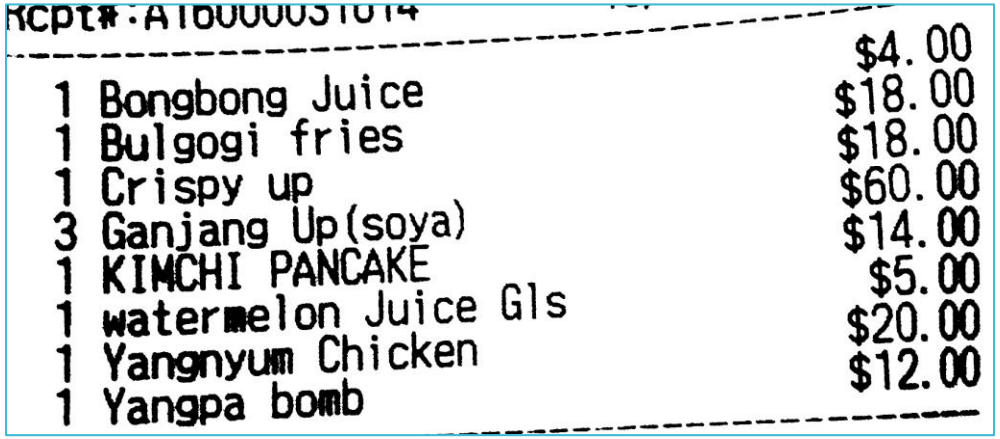
IMPROVING RESULTS
ORIENTATION

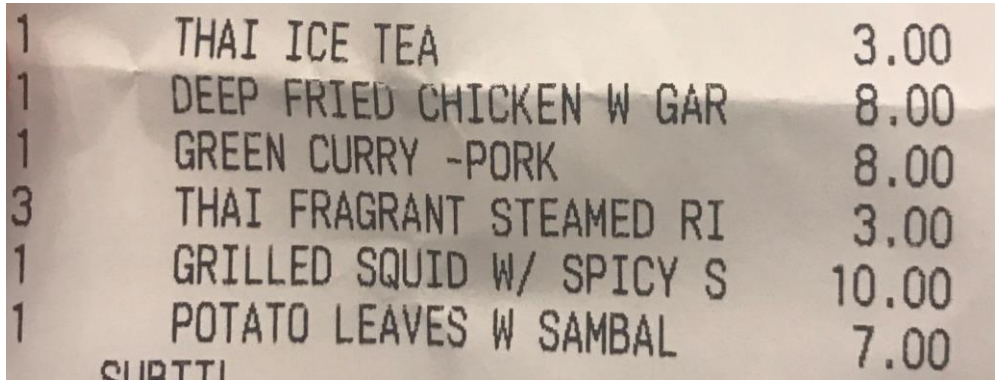
```
def rotate_to_upright(image):
    with PyTessBaseAPI(psm=PSM.OSD_ONLY) as api:
        api.SetImage(image)

        os = api.DetectOS()
        if os:
            if os['orientation'] == Orientation.PAGE_RIGHT:
                image = image.rotate(90, expand=True)

            if os['orientation'] == Orientation.PAGE_LEFT:
                image = image.rotate(270, expand=True)

            if os['orientation'] == Orientation.PAGE_DOWN:
                image = image.rotate(180, expand=True)

    return image
```

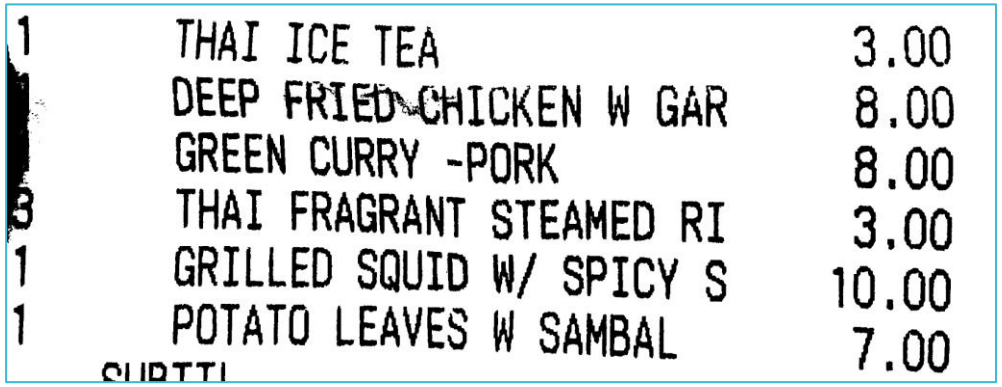
IMPROVING RESULTS
BINARISATION PT 1

```
binarised = api.GetThresholdedImage()
```



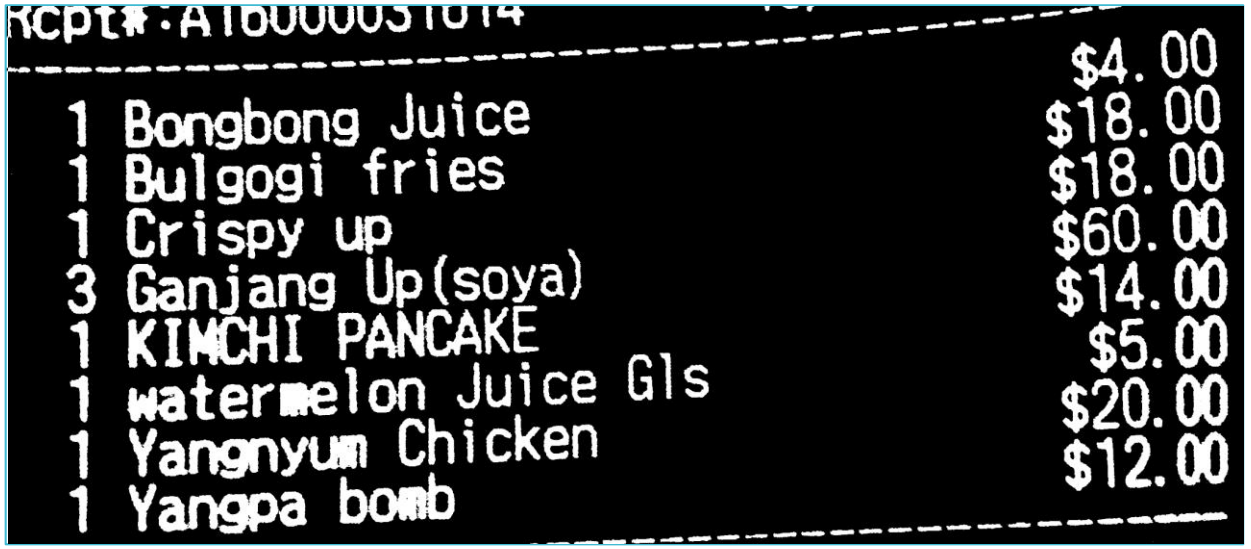






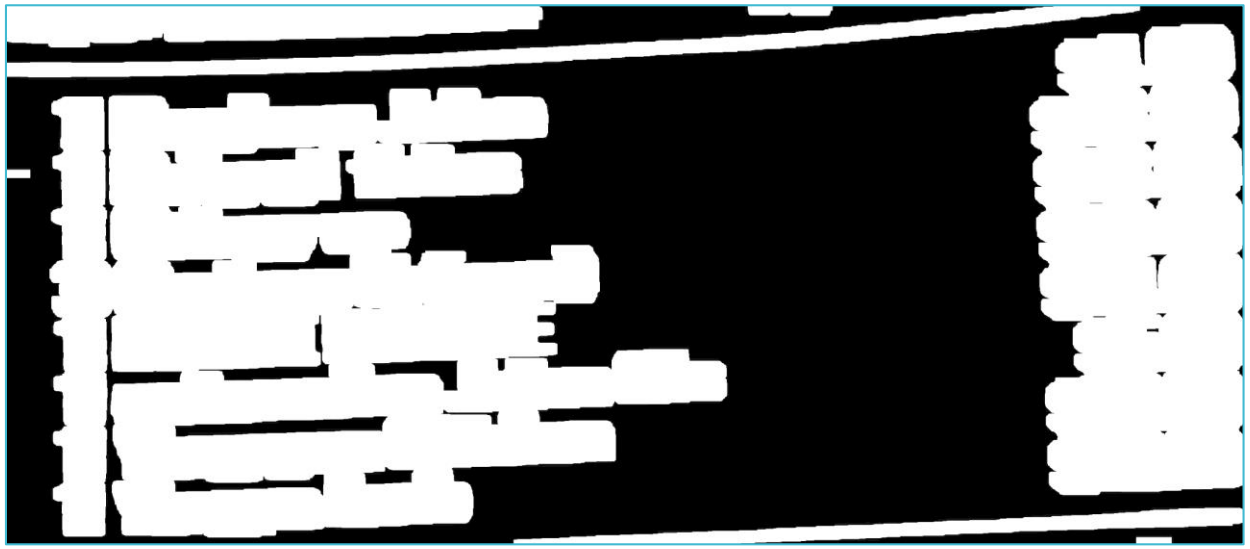
IMPROVING RESULTS
EXTRACTING TEXT

```
binarised = cv2.bitwise_not(binarised)
```



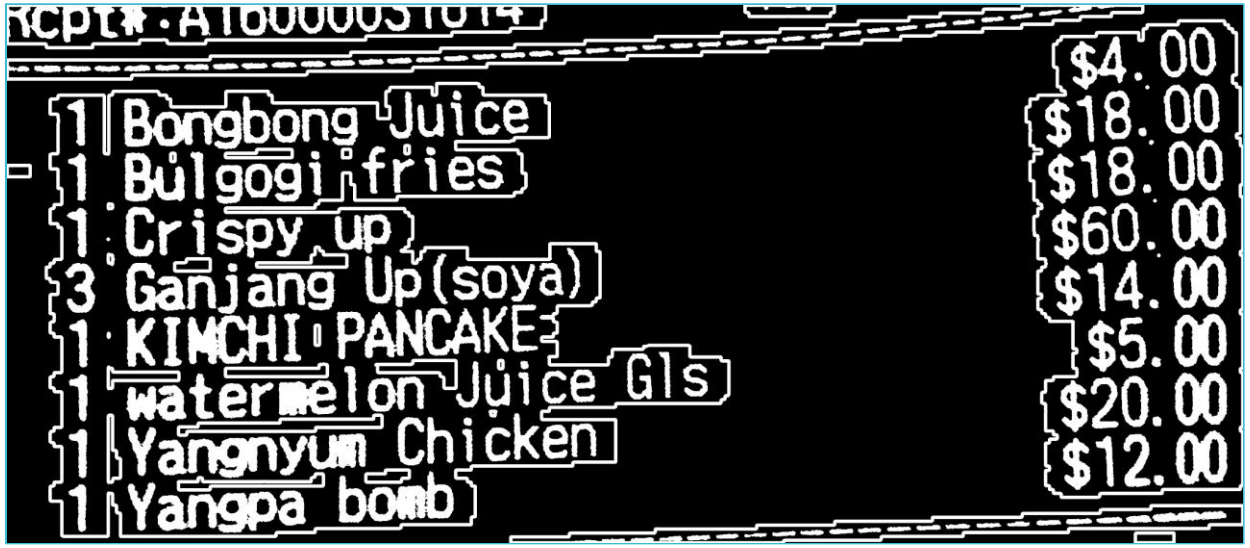
IMPROVING RESULTS
EXTRACTING TEXT

```
dilation = cv2.dilate(image, kernel, iterations=8)
```

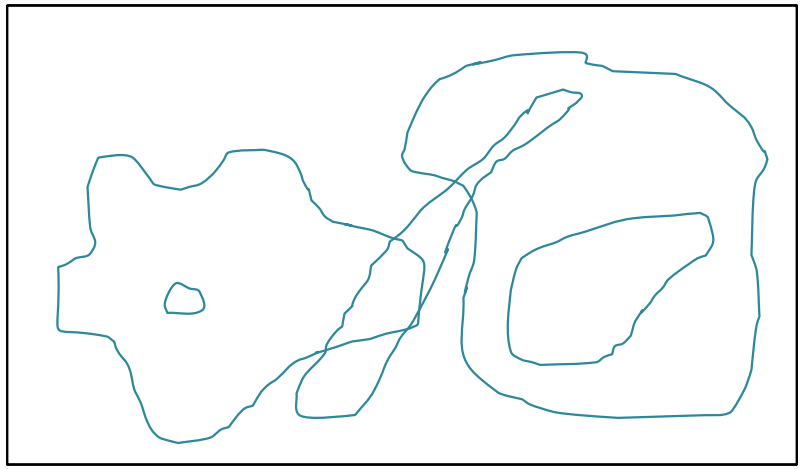


IMPROVING RESULTS
EXTRACTING TEXT

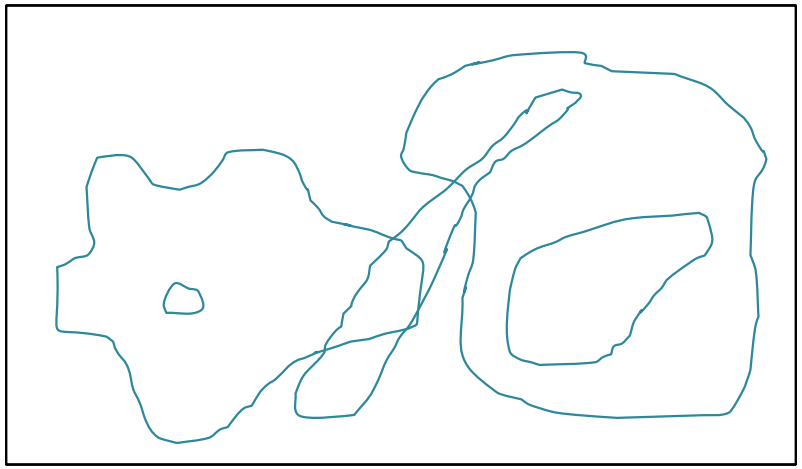
```
img, cnts, __ = cv2.findContours(dilation, cv2.RETR_TREE, cv2.CHAIN_APPROX_SIMPLE)
```



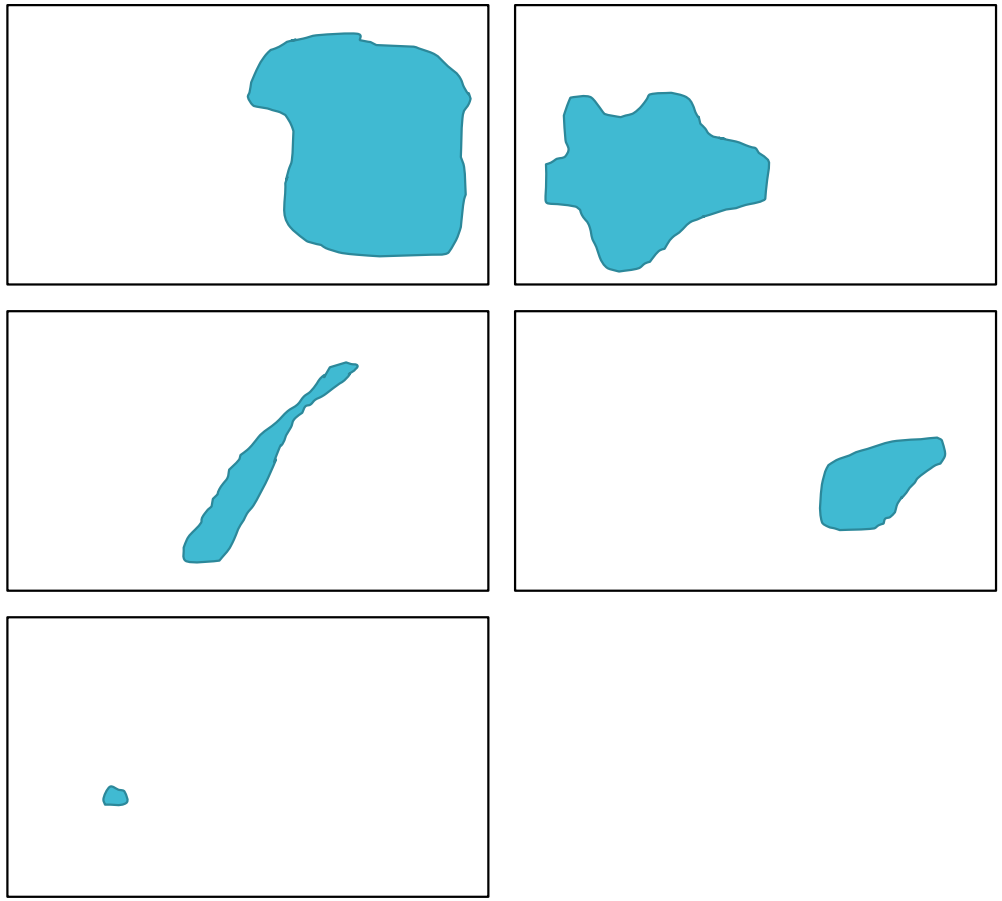
IMPROVING RESULTS
EXTRACTING TEXT



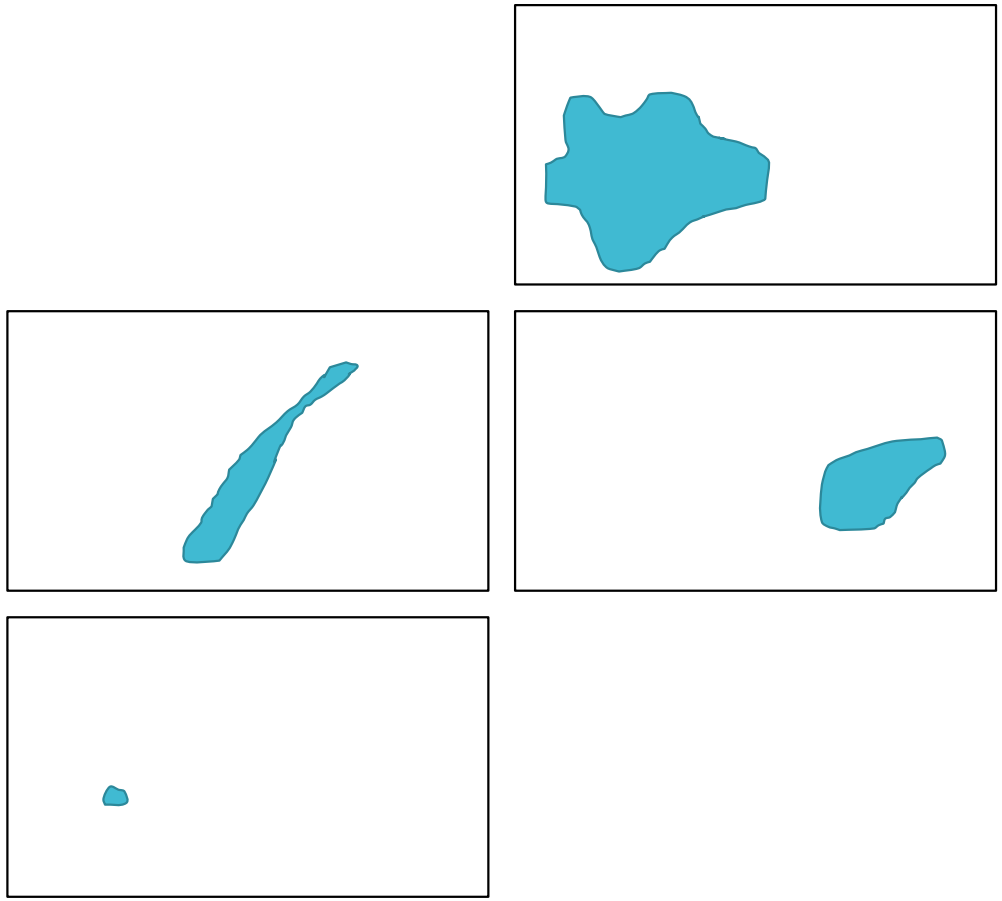
IMPROVING RESULTS
EXTRACTING TEXT

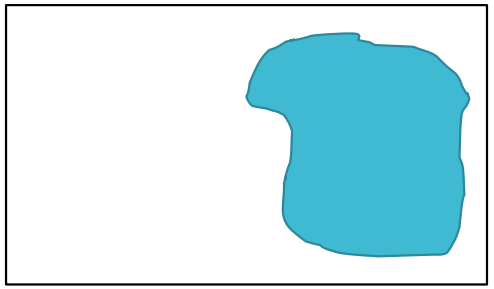


IMPROVING RESULTS
EXTRACTING TEXT

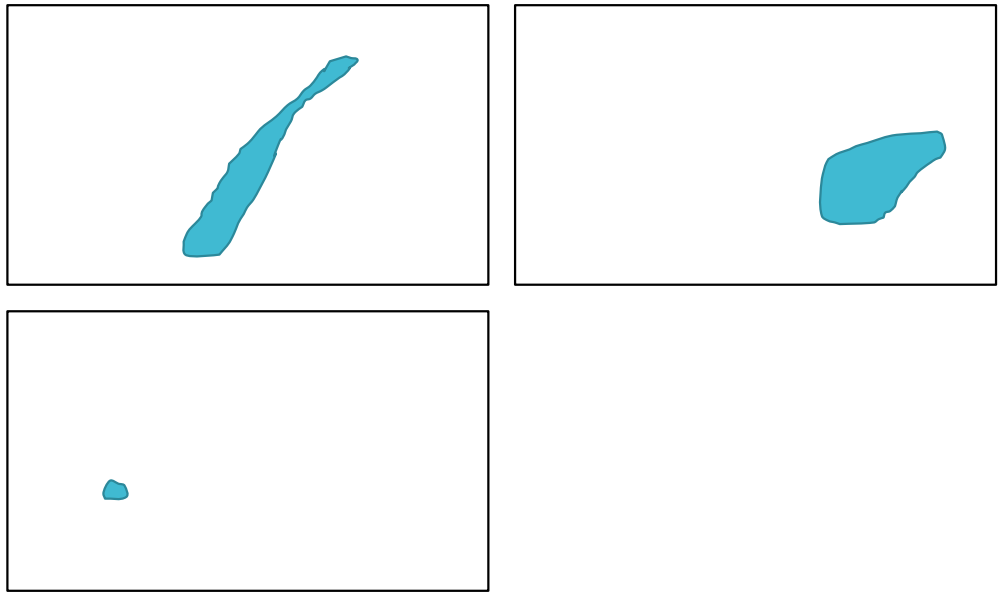


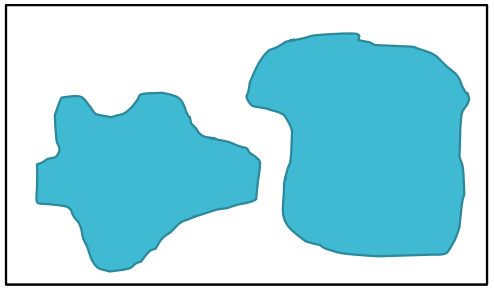
IMPROVING RESULTS
EXTRACTING TEXT



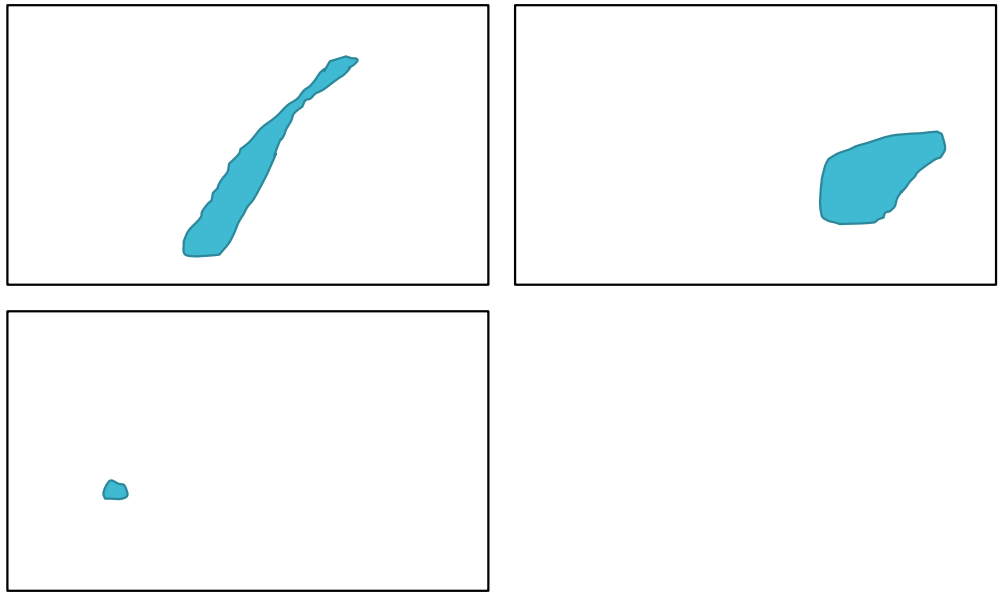


IMPROVING RESULTS
EXTRACTING TEXT

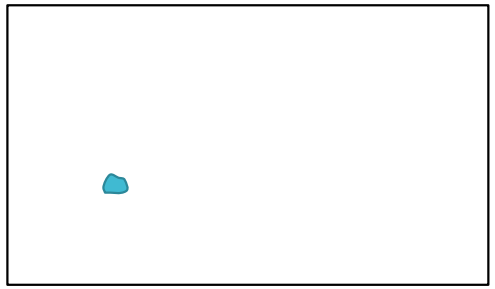


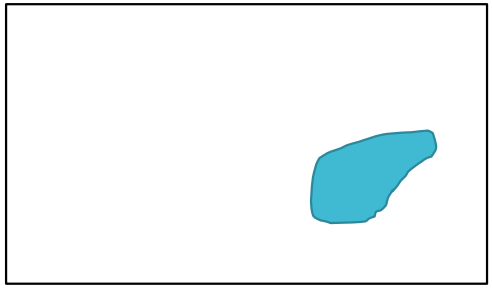


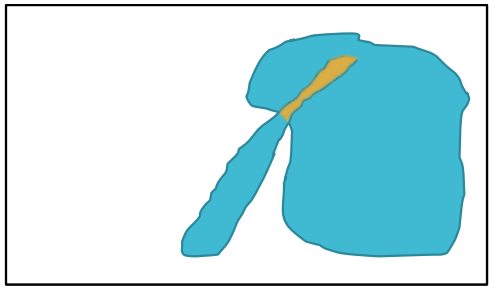
IMPROVING RESULTS
EXTRACTING TEXT

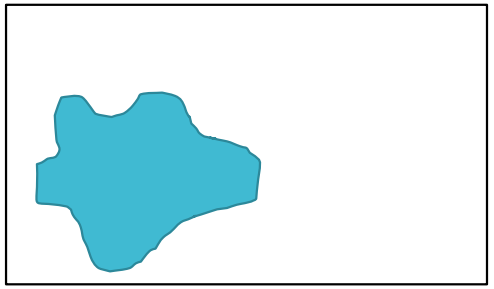


IMPROVING RESULTS
EXTRACTING TEXT

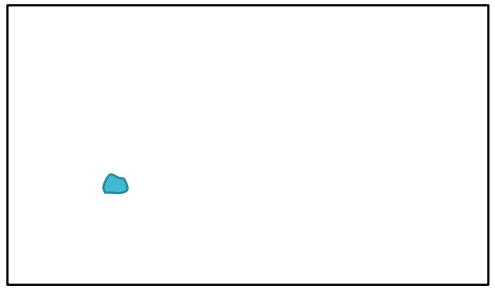


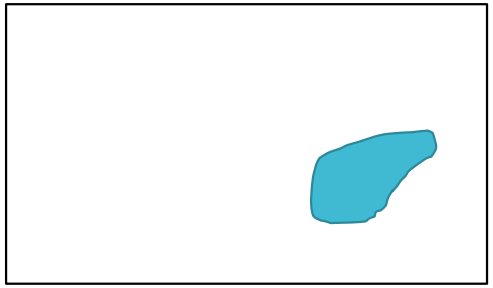


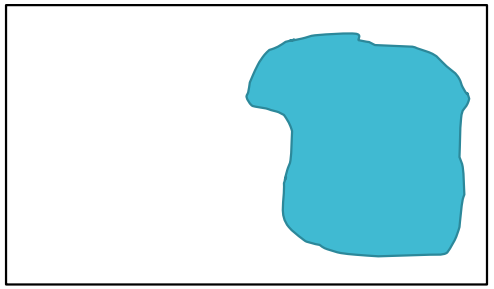


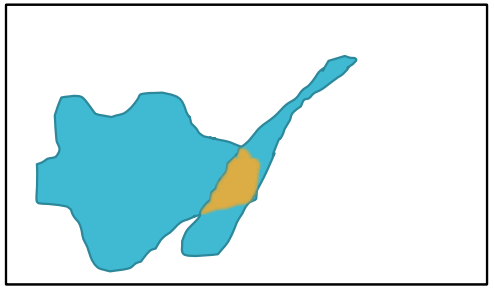


IMPROVING RESULTS
EXTRACTING TEXT

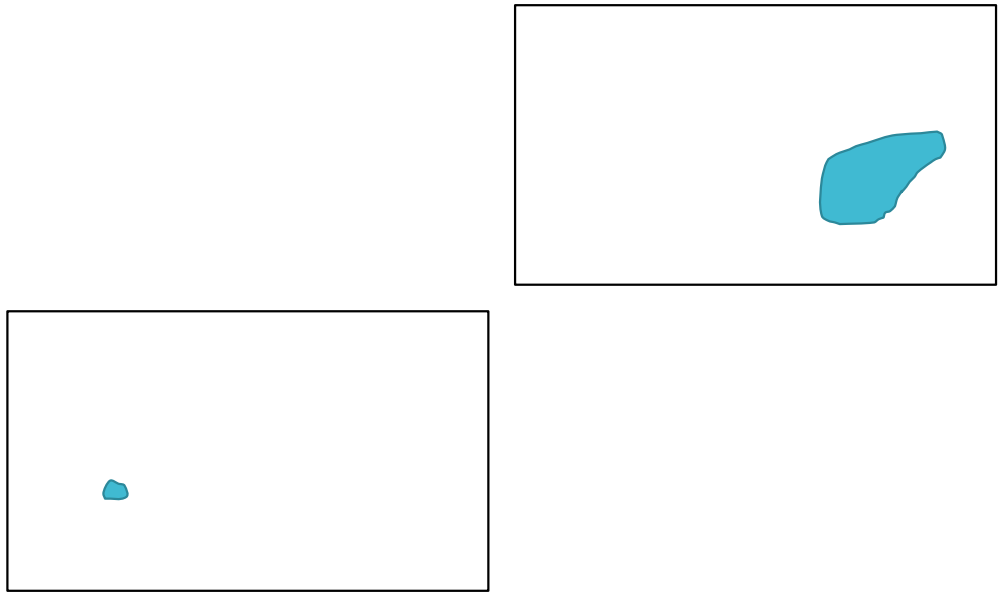


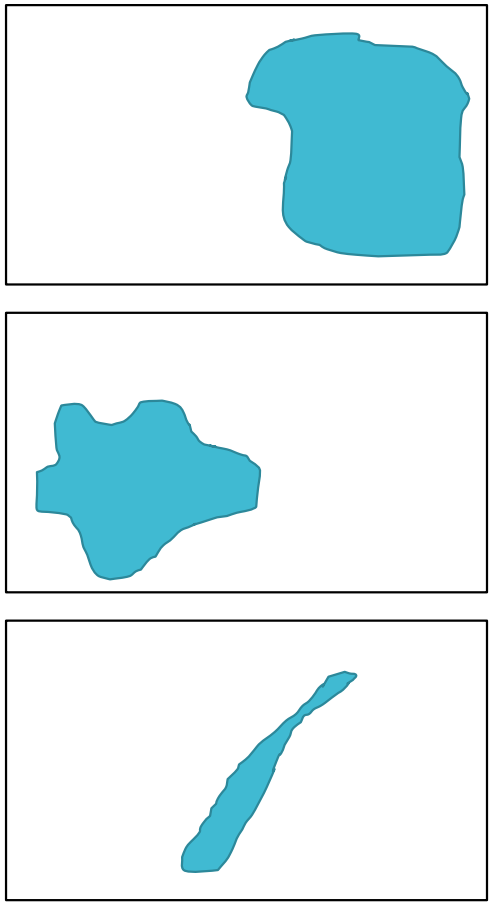




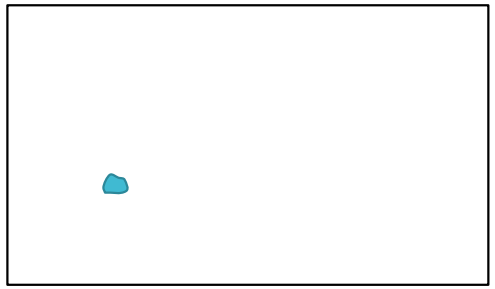


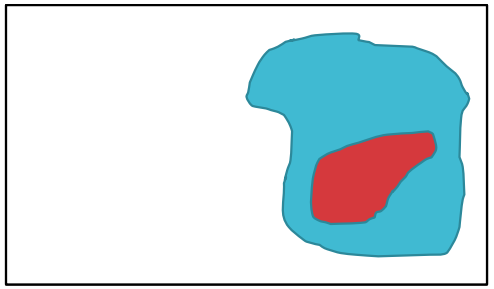
IMPROVING RESULTS
EXTRACTING TEXT

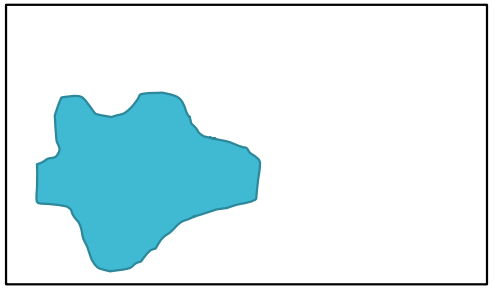


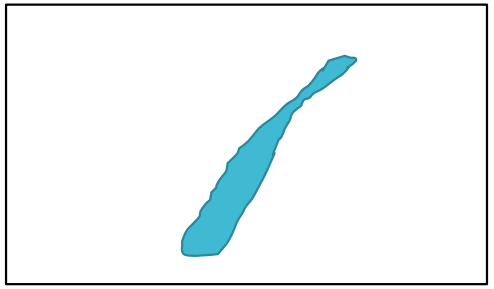


IMPROVING RESULTS
EXTRACTING TEXT

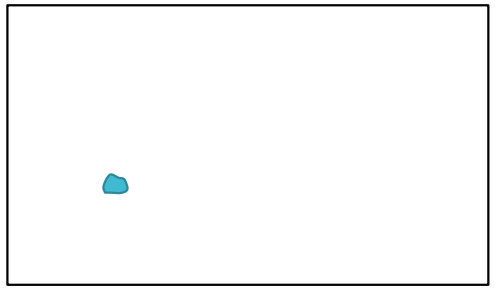


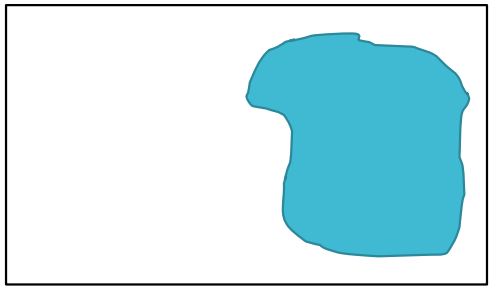


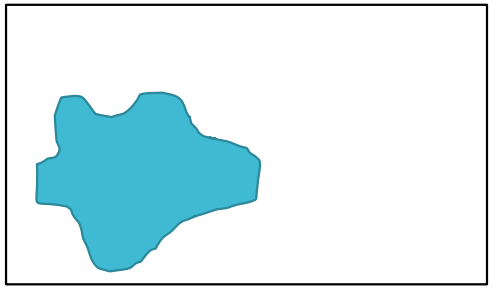


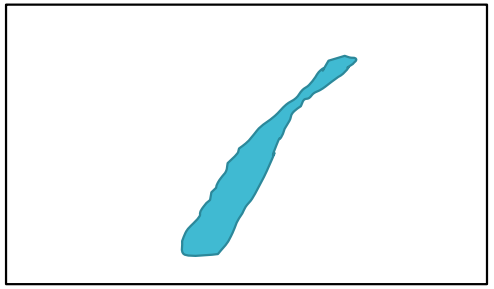


IMPROVING RESULTS
EXTRACTING TEXT

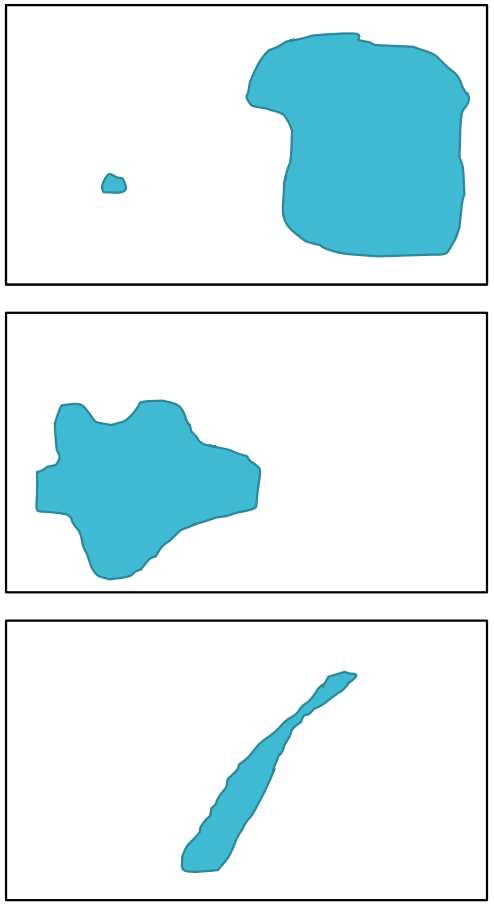




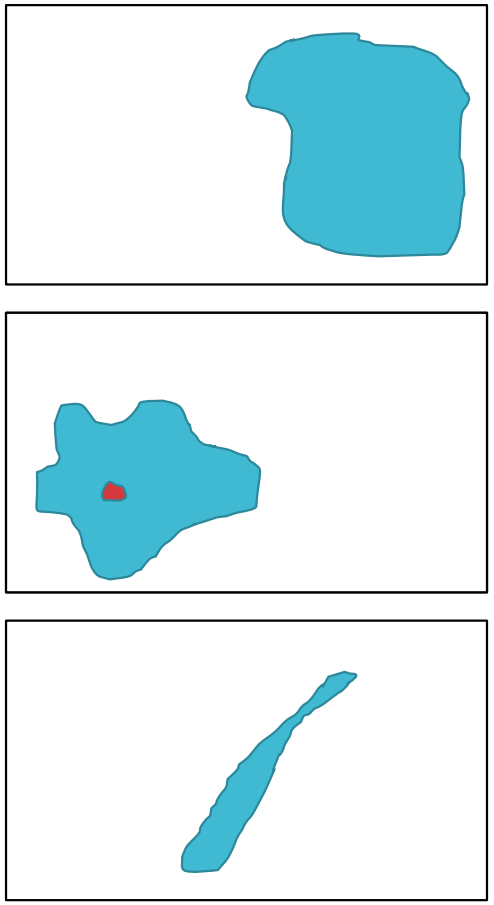




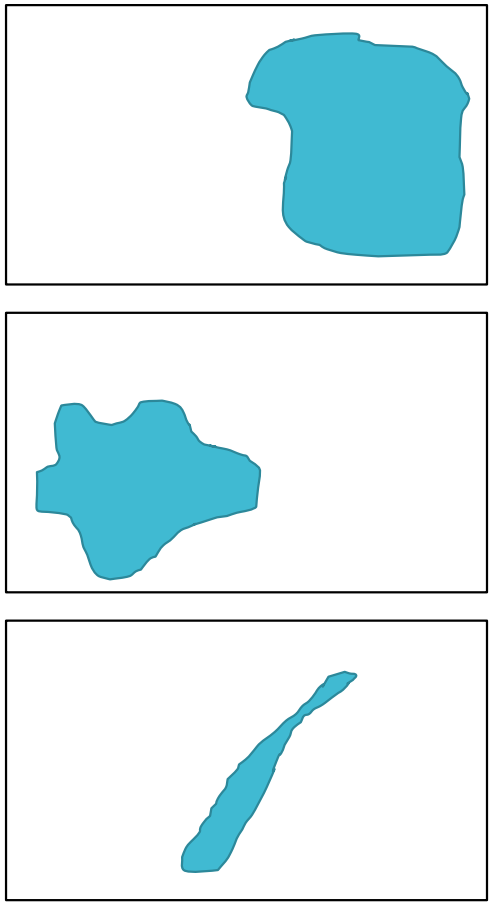
IMPROVING RESULTS
EXTRACTING TEXT



IMPROVING RESULTS
EXTRACTING TEXT

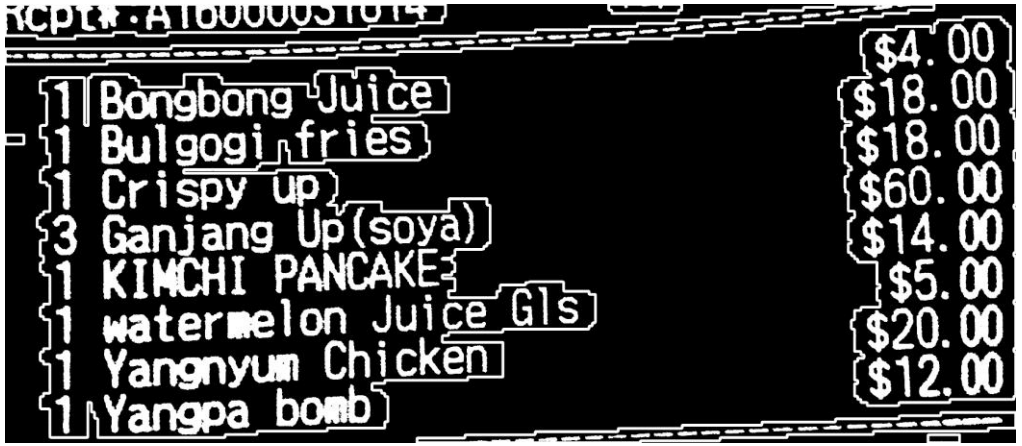


IMPROVING RESULTS
EXTRACTING TEXT



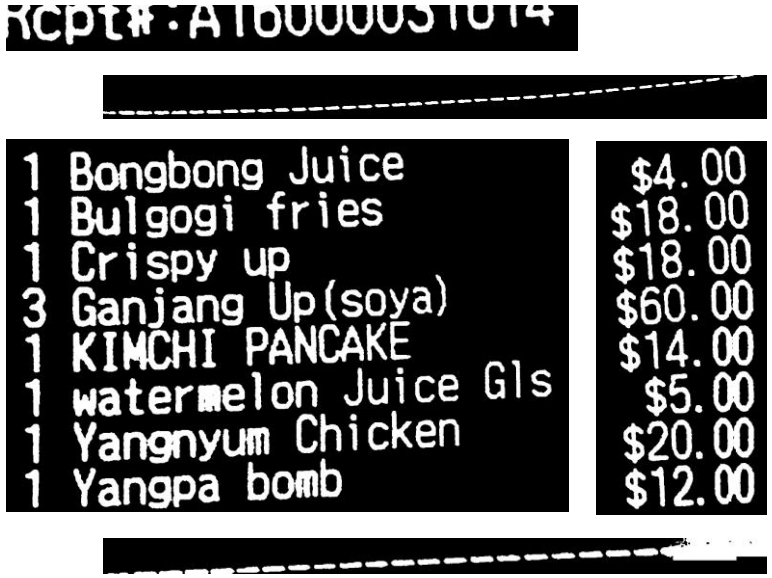
IMPROVING RESULTS
EXTRACTING TEXT

```
cnt_areas = [(cnt, create_mask(image.shape, cnt))
                  for cnt in cnts]
cnt_areas.sort(key=lambda x: cv2.contourArea(x[0]), reverse=True)

superset_areas = [cnt_areas[0]]
for i in range(1, len(cnt_areas)):
    cnt_area = cnt_areas[i]
    area1_is_within = False
    for __, superset_mask in superset_areas:
        if is_within(cnt_area[1], superset_mask):
            area1_is_within = True
            break
    if not area1_is_within:
        superset_areas.append(cnt_area)
```



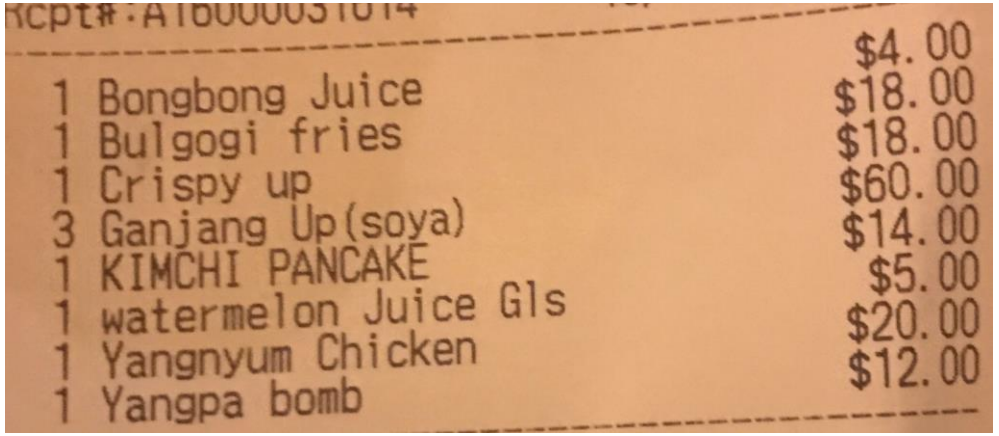
IMPROVING RESULTS
EXTRACTING TEXT

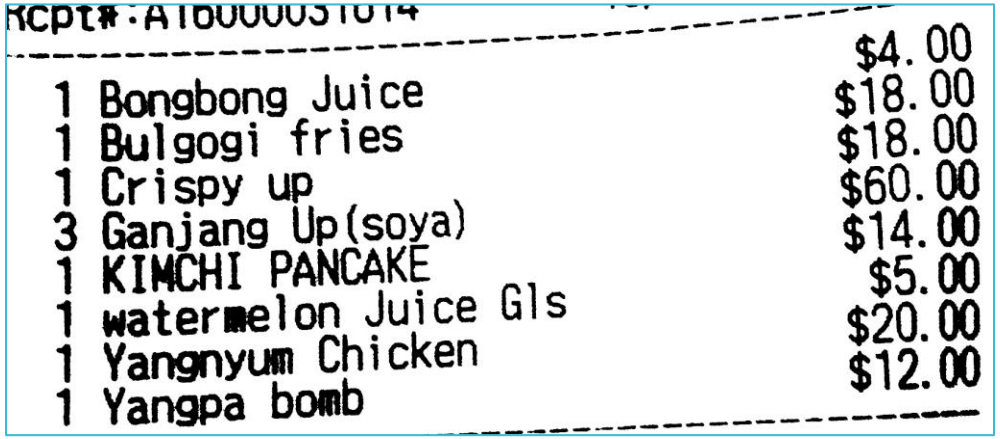
```
text_areas = []
for cnt, mask in superset_areas:
    x, y, w, h = cv2.boundingRect(cnt)

    if w < 50 or h < 50:
        continue

    if x < 0:
        x = 0
    if y < 0:
        y = 0
    crop = original[y: y + h, x: x + w]
    masked_image = cv2.bitwise_and(crop, mask[y: y + h, x: x + w])
    text_areas.append((masked_image, x, y))
```

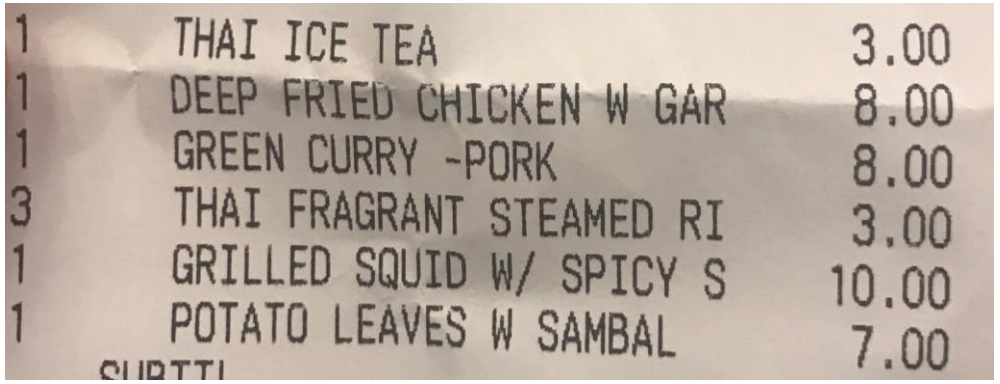


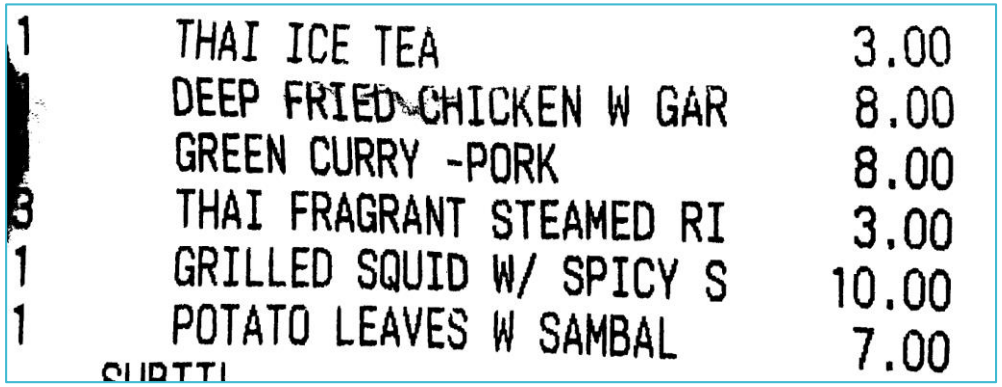
IMPROVING RESULTS
BINARISATION PT 1

```
binarised = api.GetThresholdedImage()
```



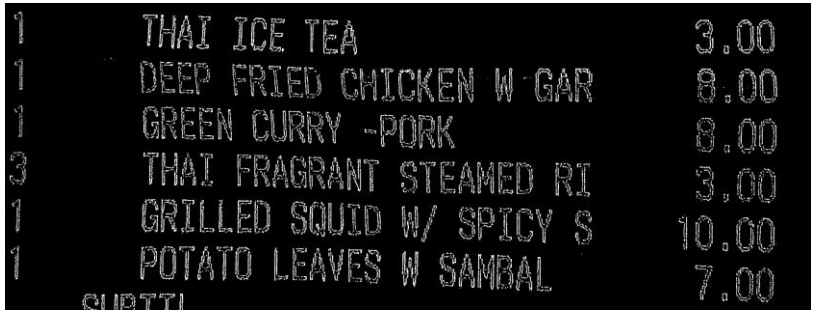




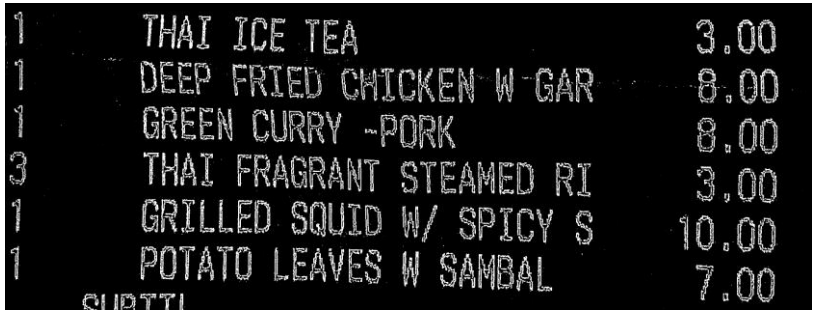


IMPROVING RESULTS
BINARISATION PT 2

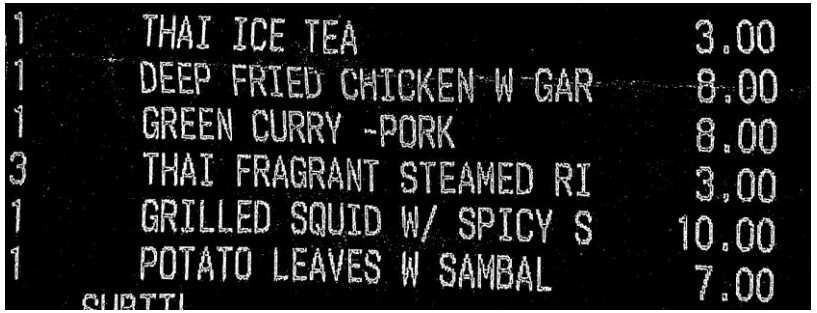
```
cv2.adaptiveThreshold(image, 255, cv2.ADAPTIVE_THRESH_GAUSSIAN_C,
                      cv2.THRESH_BINARY, i, 2)
```



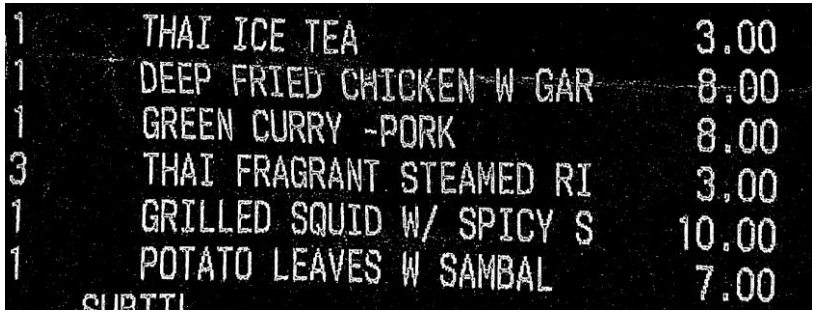
i = 3



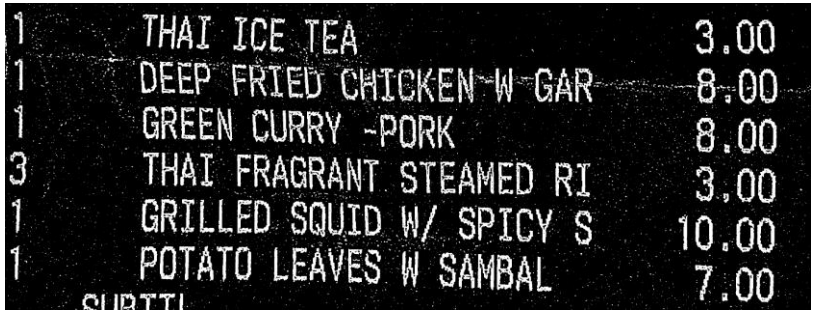
i = 5



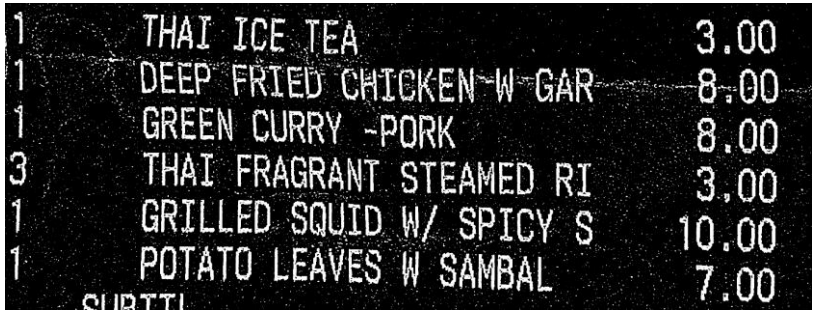
i = 7



i = 9



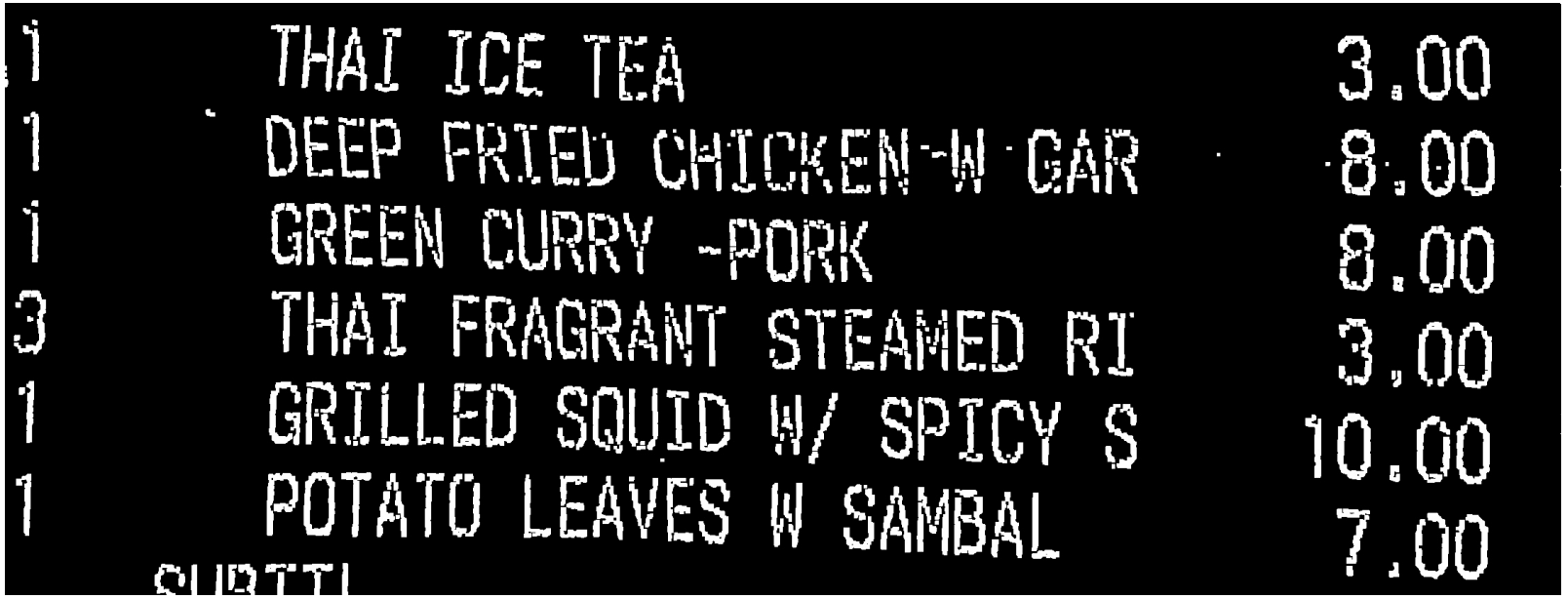
i = 11



i = 13

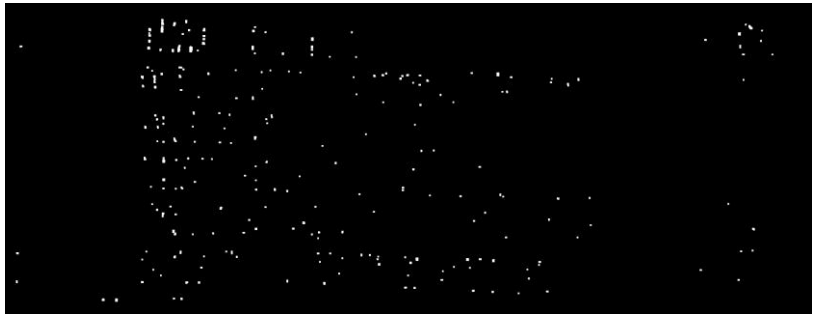
IMPROVING RESULTS
BINARISATION PT 2

```
kernel = np.ones((5, 5), np.uint8)
image = cv2.erode(image, kernel, iterations=1)
image = cv2.dilate(image, kernel, iterations=1)
```

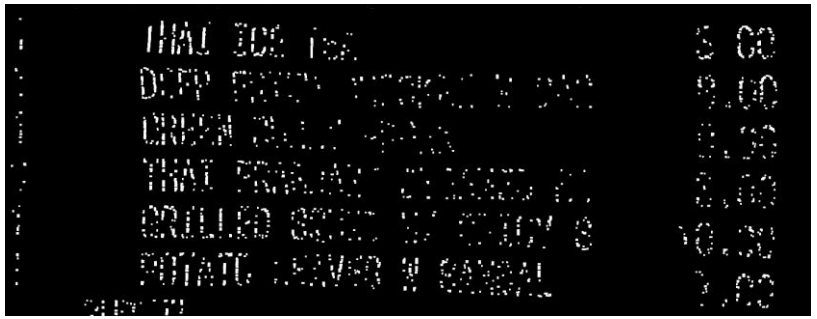


IMPROVING RESULTS
BINARISATION PT 2

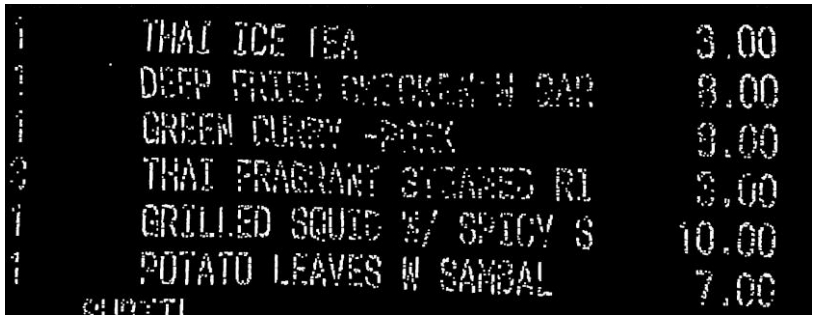
i = 3



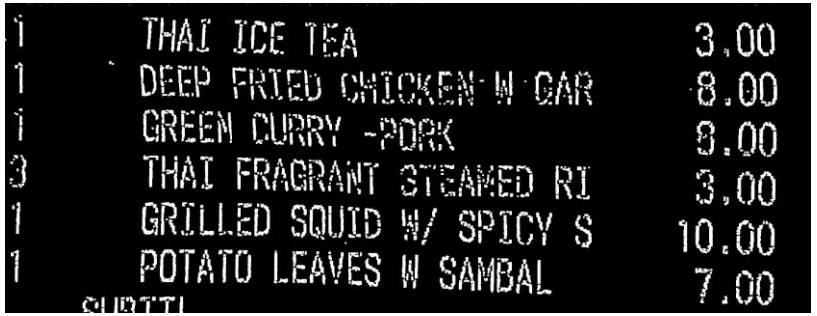
i = 5



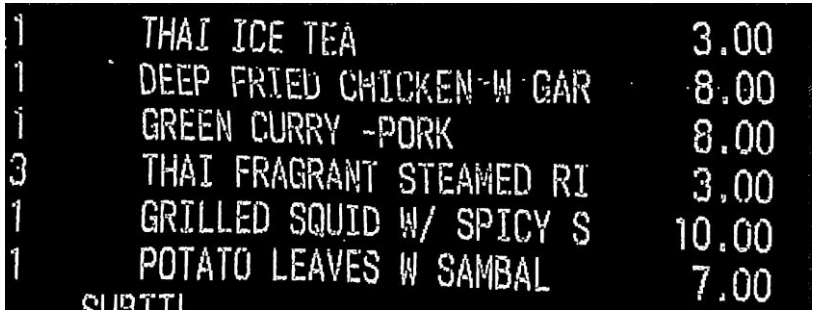
i = 7



i = 9



i = 11

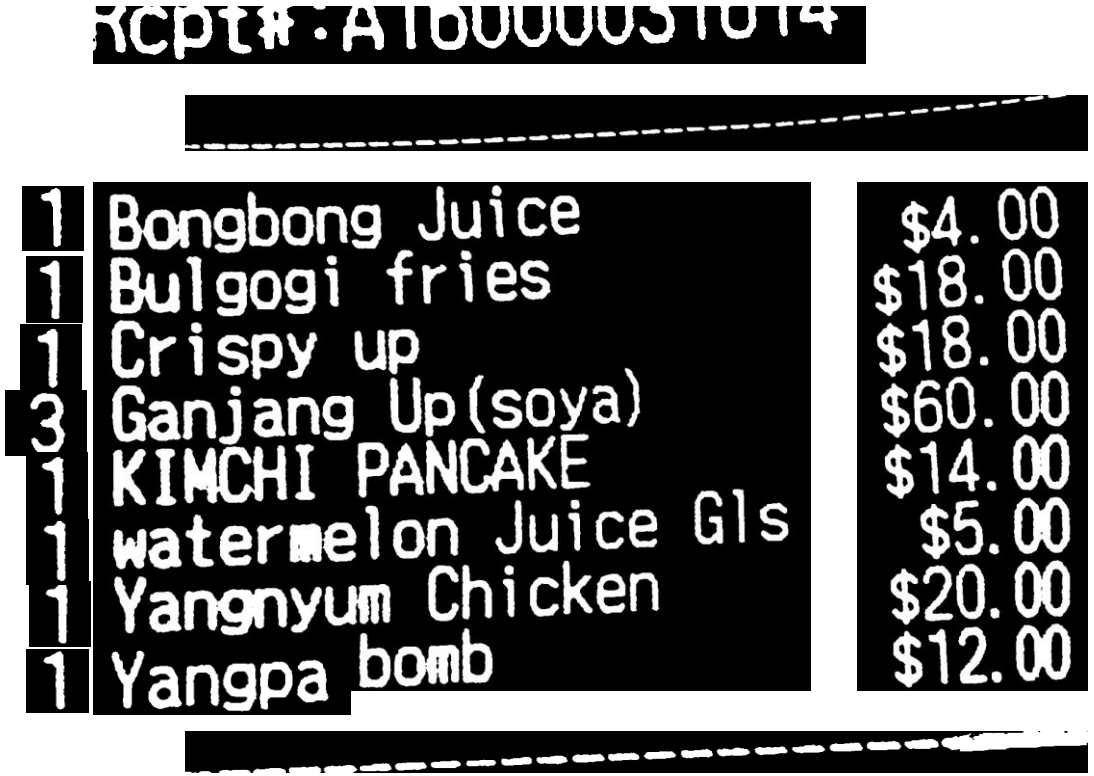


i = 13

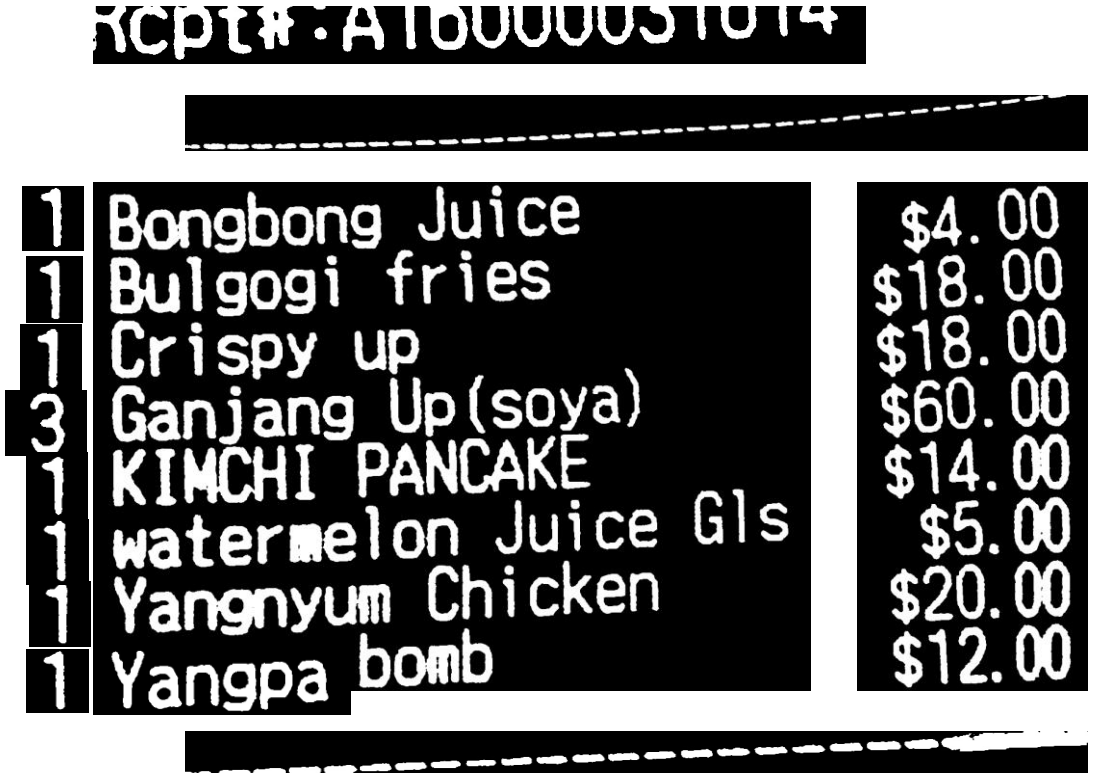
IMPROVING RESULTS
BINARISATION PT 2

```
text_areas = []
for cnt, mask in superset_areas:
    x, y, w, h = cv2.boundingRect(cnt)

    if w < 50 or h < 50:
        continue

    if x < 0:
        x = 0
    if y < 0:
        y = 0
    crop = binarise_image_otsu(original[y: y + h, x: x + w])
    masked_image = cv2.bitwise_and(crop, mask[y: y + h, x: x + w])
    text_areas.append((masked_image, x, y))
```

IMPROVING RESULTS
DESKEWING

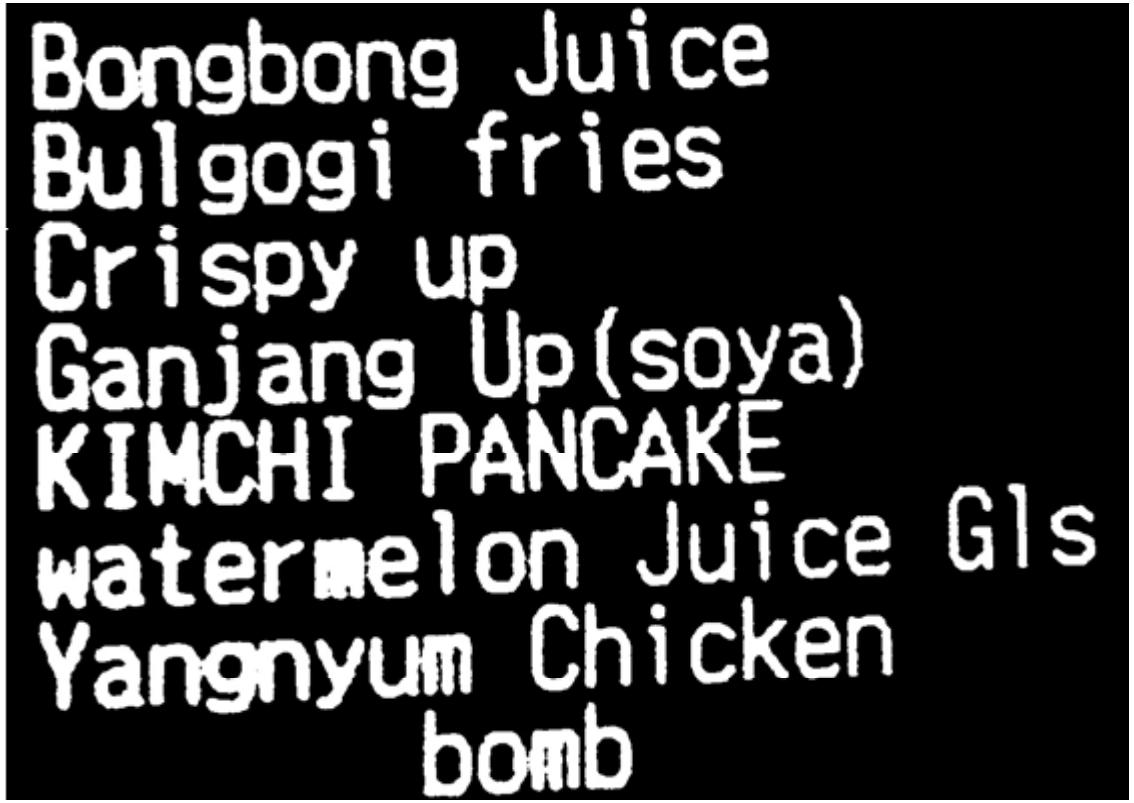


IMPROVING RESULTS
DESKEWING



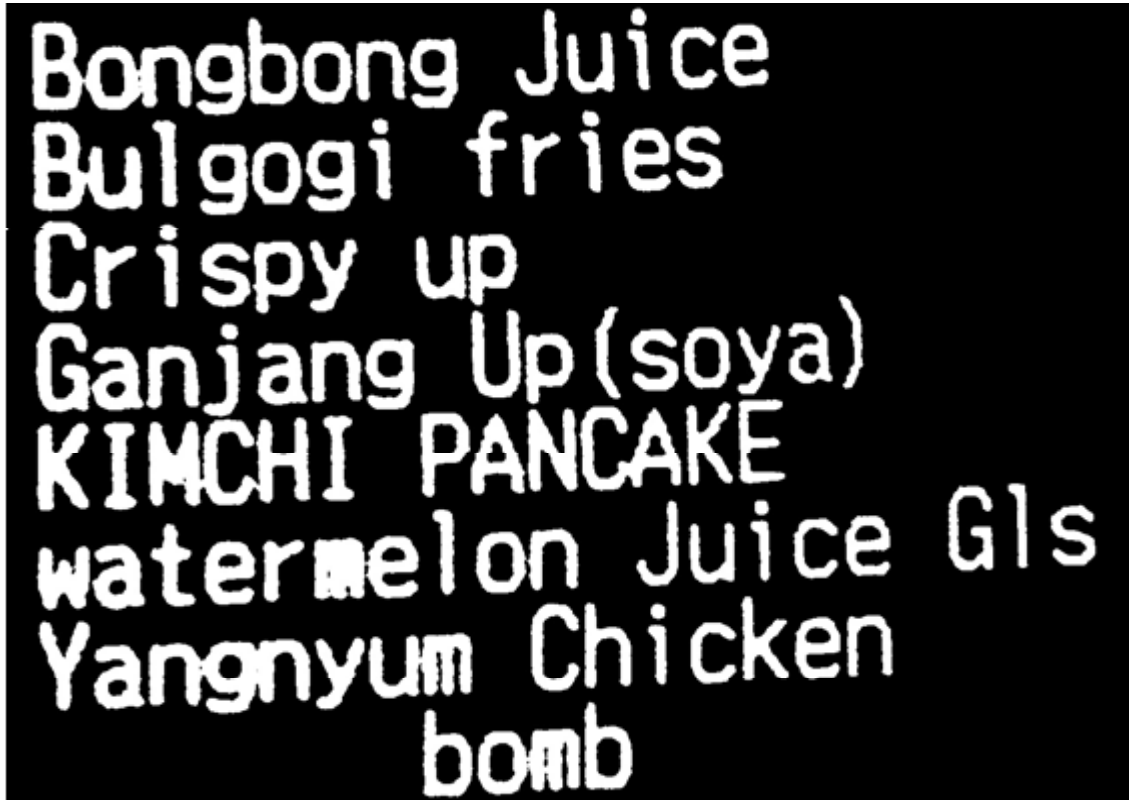
IMPROVING RESULTS
DESKEWING

```
lines = cv2.HoughLinesP(img_segment, 1, np.pi / 180, 100,
                        minLineLength=width // 3, maxLineGap=100)
```



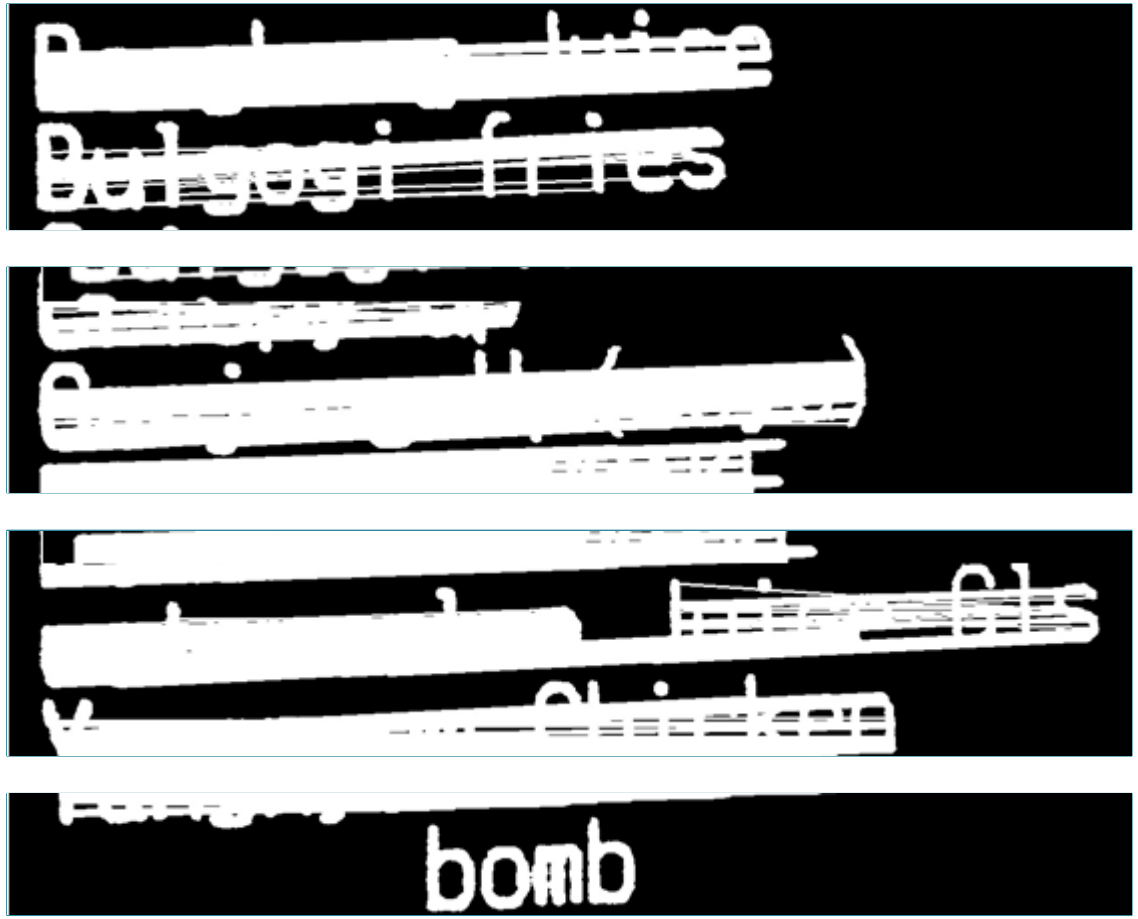
IMPROVING RESULTS
DESKEWING

```
lines = cv2.HoughLinesP(img_segment, 1, np.pi / 180, 100,
                        minLineLength=width // 3, maxLineGap=100)
```

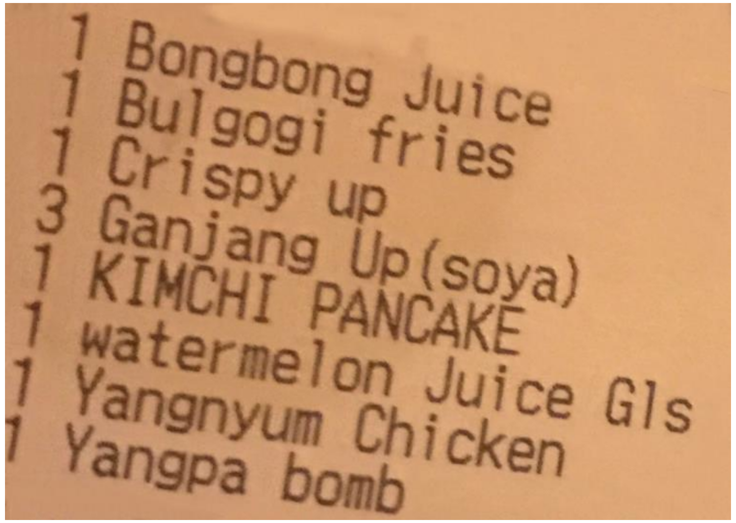


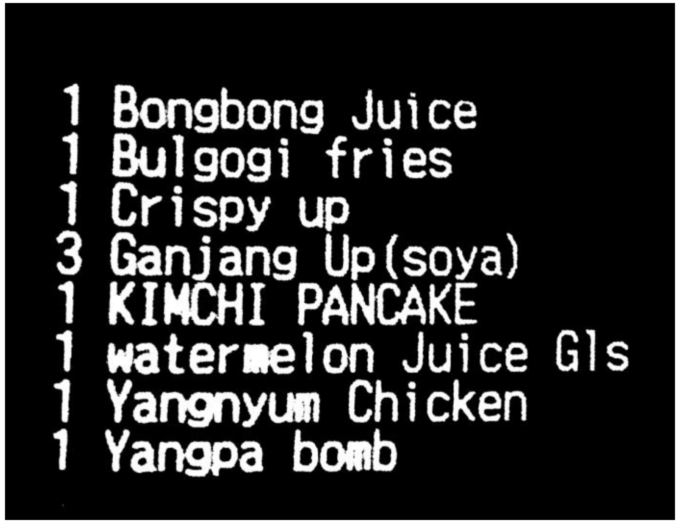
IMPROVING RESULTS
DESKEWING

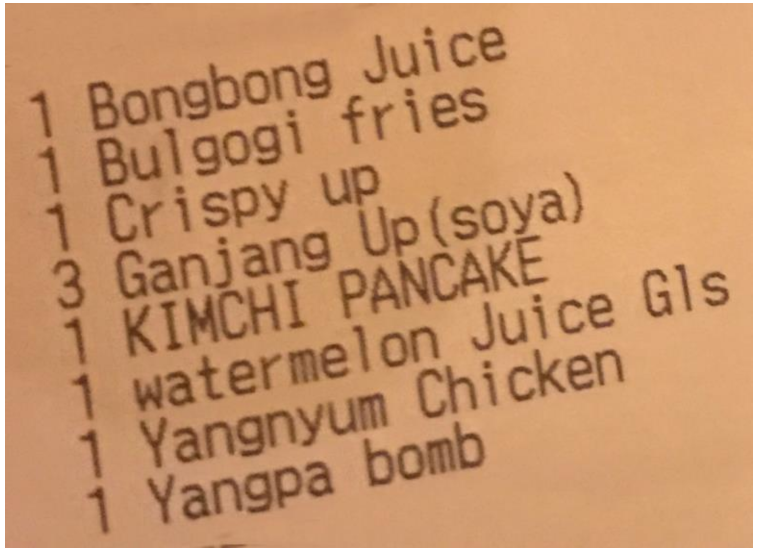
```
lines = cv2.HoughLinesP(img_segment, 1, np.pi / 180, 100,
                        minLineLength=width // 3, maxLineGap=100)
```

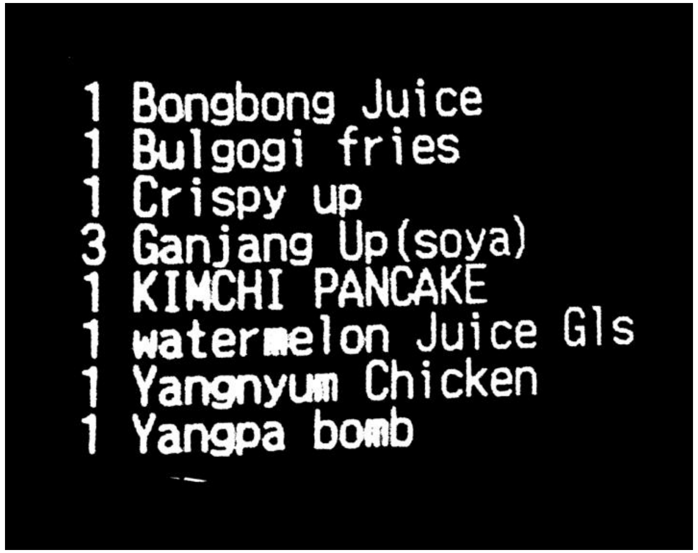


IMPROVING RESULTS
DESKEWING



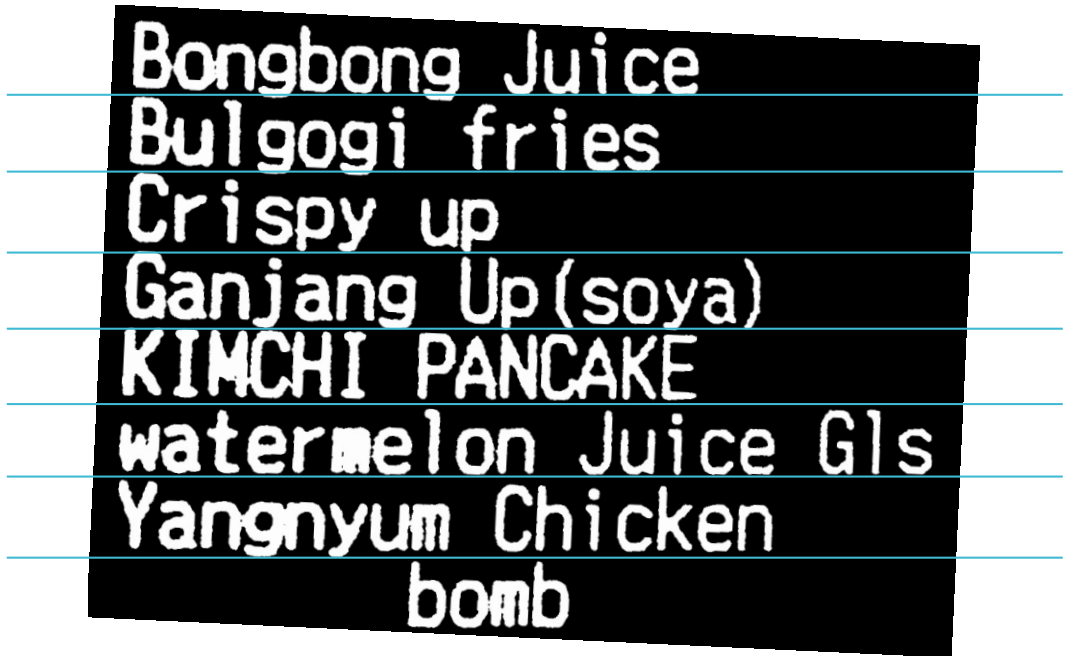






IMPROVING RESULTS
DESKEWING

```
variances = []
threads = []
for angle in np.linspace(-2.0, 2.0, num=32):
    thread = threading.Thread(
        target=cal_img_variance,
        args=(image, angle, variances)
    )
    thread.start()
    threads.append(thread)

for thread in threads:
    thread.join()

best_angle = 0
max_variance = 0
for angle, variance in variances:
    if variance > max_variance:
        max_variance = variance
        best_angle = angle

return best_angle
```



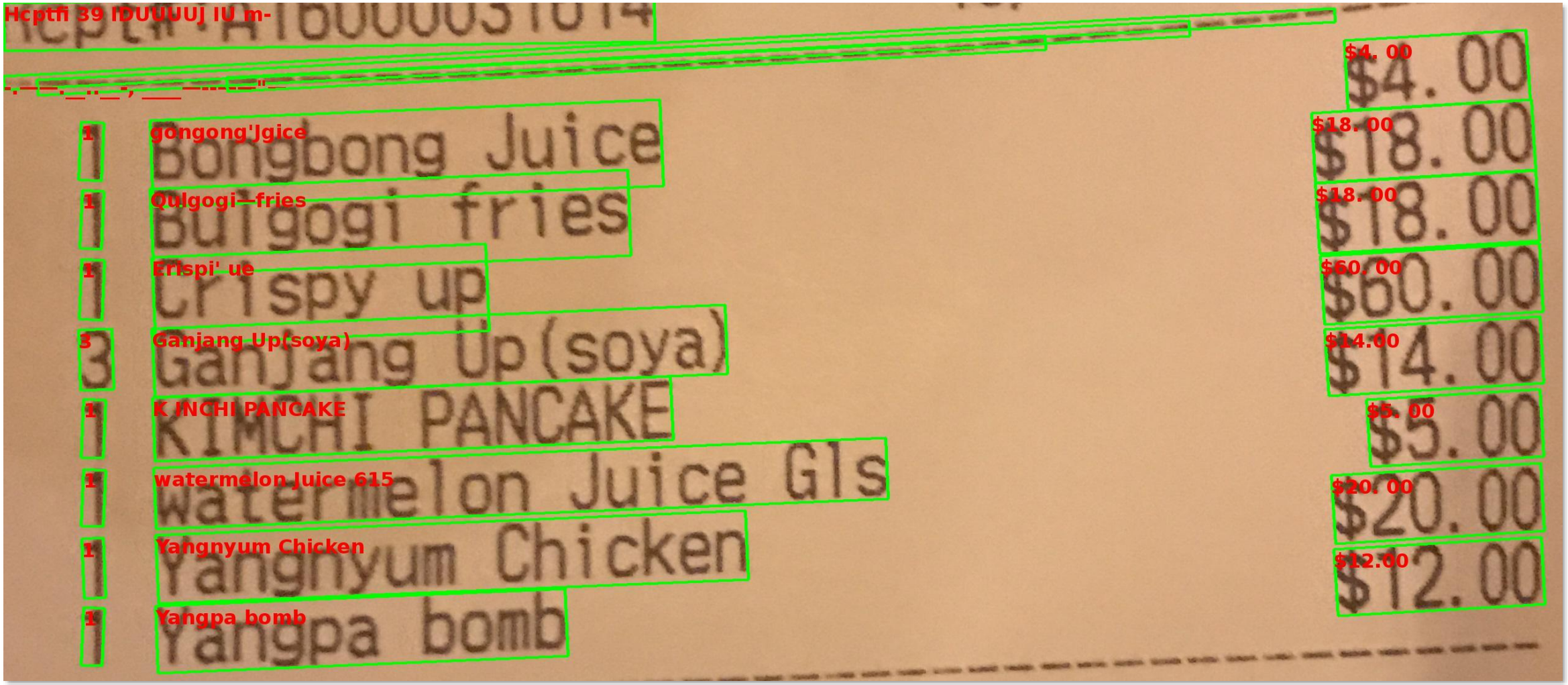
IMPROVING RESULTS
EVALUATING

```
with PyTessBaseAPI(psm=PSM.SINGLE_BLOCK, lang='eng') as api:
    api.SetImage(image)
    boxes = api.GetComponentImages(RIL.TEXTLINE, True)
    for im, box, __, __ in boxes:
        x = box['x']
        y = box['y']
        w = box['w']
        h = box['h']

        api.SetRectangle(x, y, w, h)
        ocr_result = api.GetUTF8Text()
```

```
results = Queue()
processes = []
for text_area in text_areas:
    process = Process(
        target=evaluate_text_area,
        args=(text_area, results)
    )
    process.start()
    processes.append(process)

for process in processes:
    process.join()
```

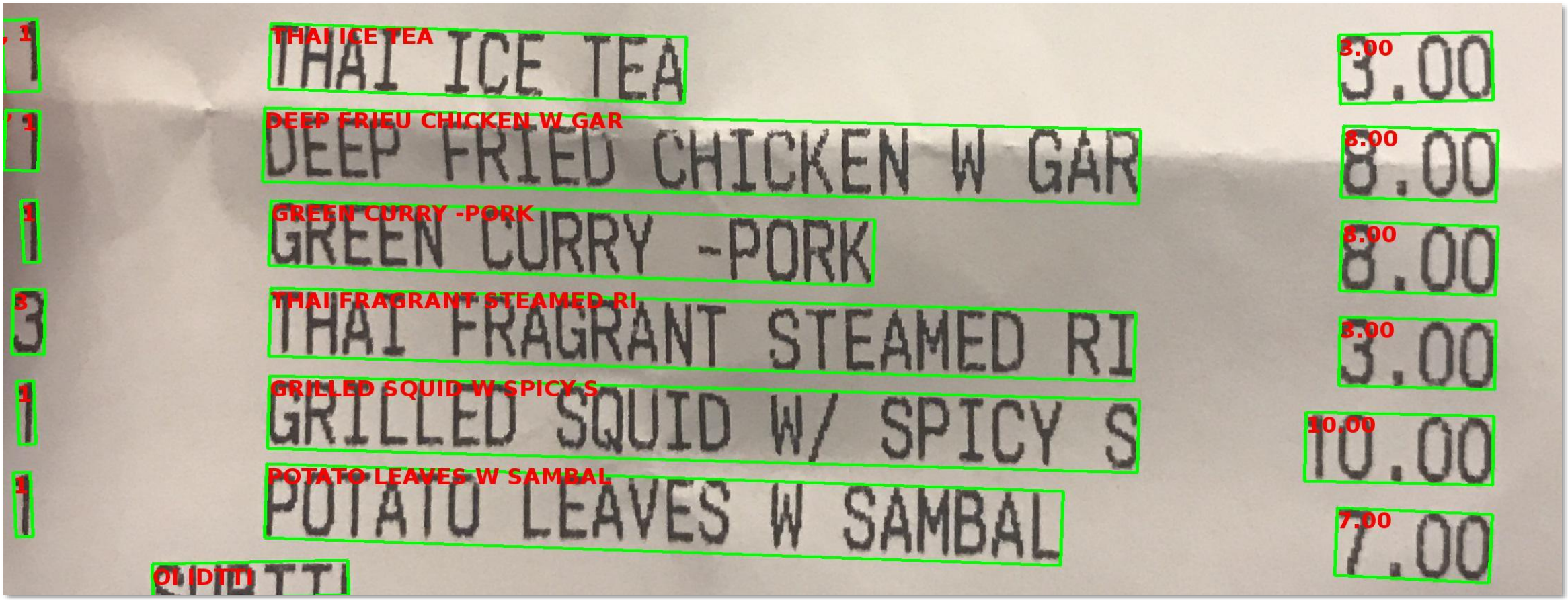
IMPROVING RESULTS

RESULTS



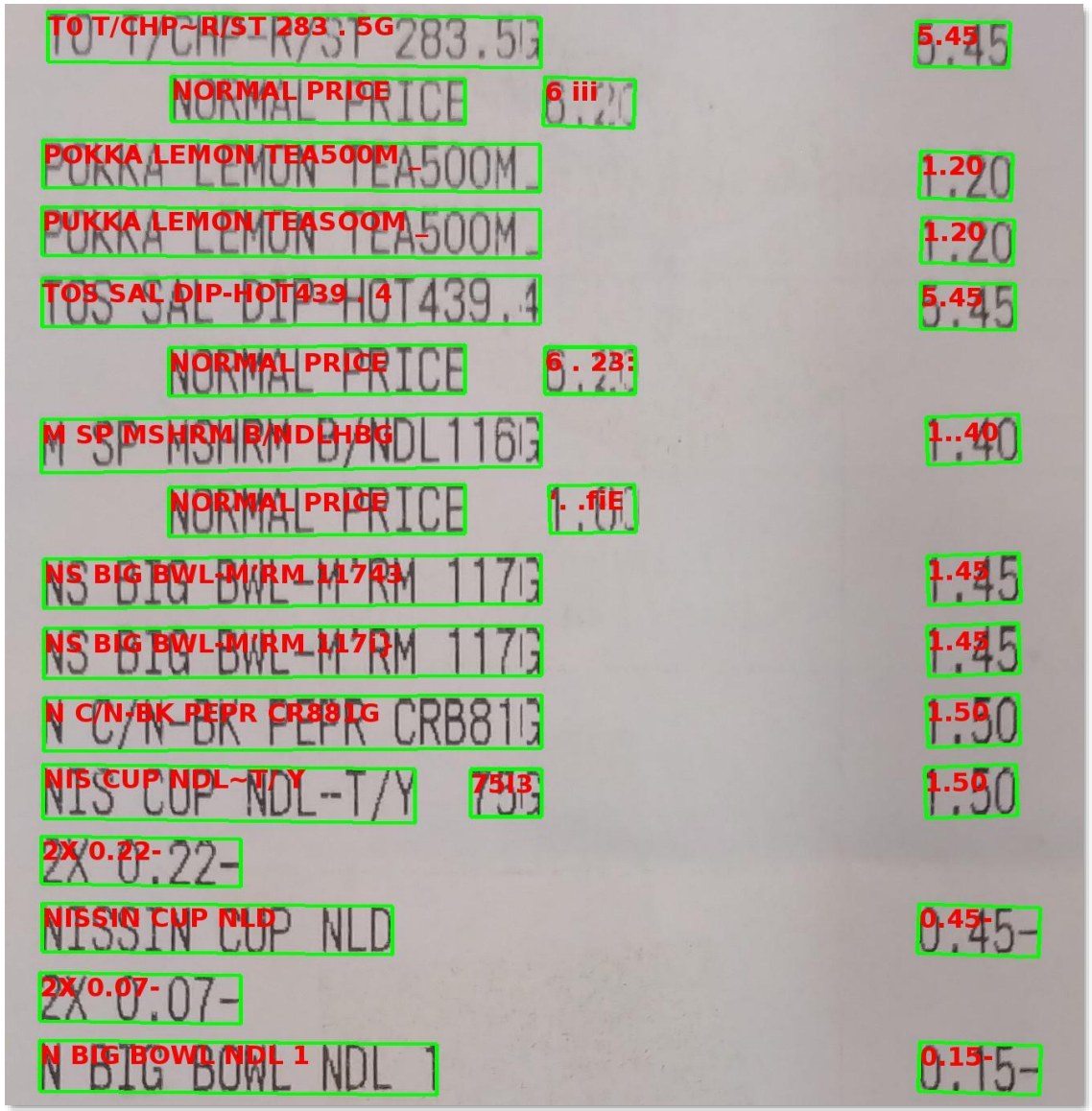
IMPROVING RESULTS

RESULTS



IMPROVING RESULTS

RESULTS



THE END
LESSONS LEARNT

OCR is hard
Nigh impossible to intuit

Constrained images
Tesseract works best with
smaller constrained images

Constrain user
Hard to engineer solution
that works for every scenario

THE END
WHAT'S NEXT

Performance
Can we do better?

Bug Fixing
Identify the reason for "random" crashes

Accuracy
Tesseract 4.0 w/ LSTM

THE END
CREDITS

1. https://github.com/tesseract-ocr/tesseract
2. https://github.com/tesseract-ocr/docs/tree/master/das_tutorial2016
3. https://github.com/sirfz/tesserocr
4. http://www.danvk.org/2015/01/07/finding-blocks-of-text-in-an-image-using-python-opencv-and-numpy.html
5. http://felix.abecassis.me/2011/09/opencv-detect-skew-angle/
6. http://stackoverflow.com/questions/37239872/search-if-a-contour-exists-inside-another-contour

Repo: https://github.com/oswellchan/tesserocr-talk