Alan Felt
CS6440 Image Processing
Project 5 – OCR

My processing pipeline is as follows:

1. Import image with graphical text to be converted to text
2. Perform a non-local means image denoising using skimage restore.denoise_nl_means(). I use the default values
3. Use the filters.threshold_otsu() function in skimage to perform a threshold operation on the denoised image.
4. Pass the image into the remove_small_holes() function in skimage.morphology to get rid of holes caused by noise.
5. Label the connected components of the image
6. Pass the connected components into the remove_small_objects() function in skimage.morphology to remove small connected components caused by noise.
7. Measure the properties of each label (size, position, image pixels in connected component).
8. Sum the values of x and y for the top left corner of each label. The top left-most character in the text will be the smallest value of the sum.
9. Sort the labels based on their y position. Get all labels whose y position falls above the lower bound of the top-left character. Sort them based on their x position and remove them from the set of labels and properties.
10. Scale the connected component down and pad to 28x28px for each connected component image from the previous step.
11. Feed this image into the predictor and add the result to a list with an index based on the current line.
12. When finished with the current line, repeat the process starting at step 9. Keep repeating until no more labels are contained in the set of labels
13. Return the list of strings that represent each line of the scanned text.

Figures 1-4 below show images from the utes.jpg file and how the image is denoised, thresholded, divided into letters, and rescaled for the letter classifier.
This methodology works well for breaking up the image into individual characters. However, if the text becomes too small, thin parts of the text can be lost, resulting in multiple images for a single character. I found that my results were better for images with no serifs, which were all capitalized. The serif font seemed to confuse the machine learning model and it would return “R” as “K”.

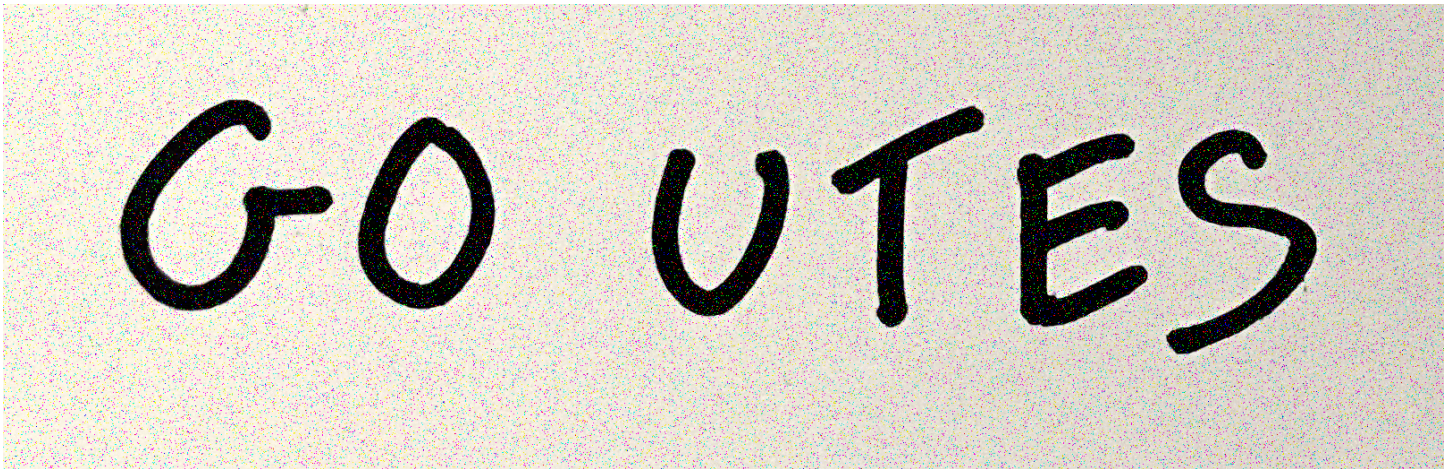

Figure 1: Unmodified utes.jpg

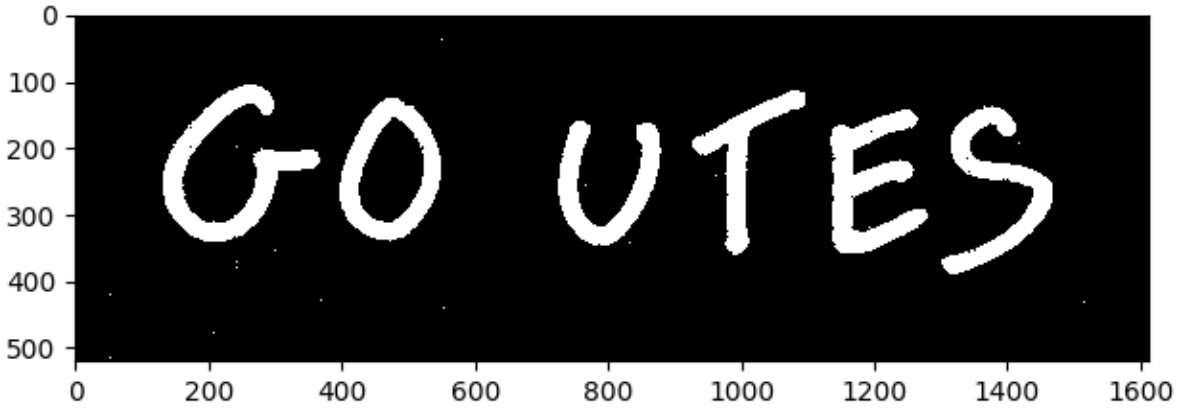

Figure 2: Denoised and thresholded utes.jpg

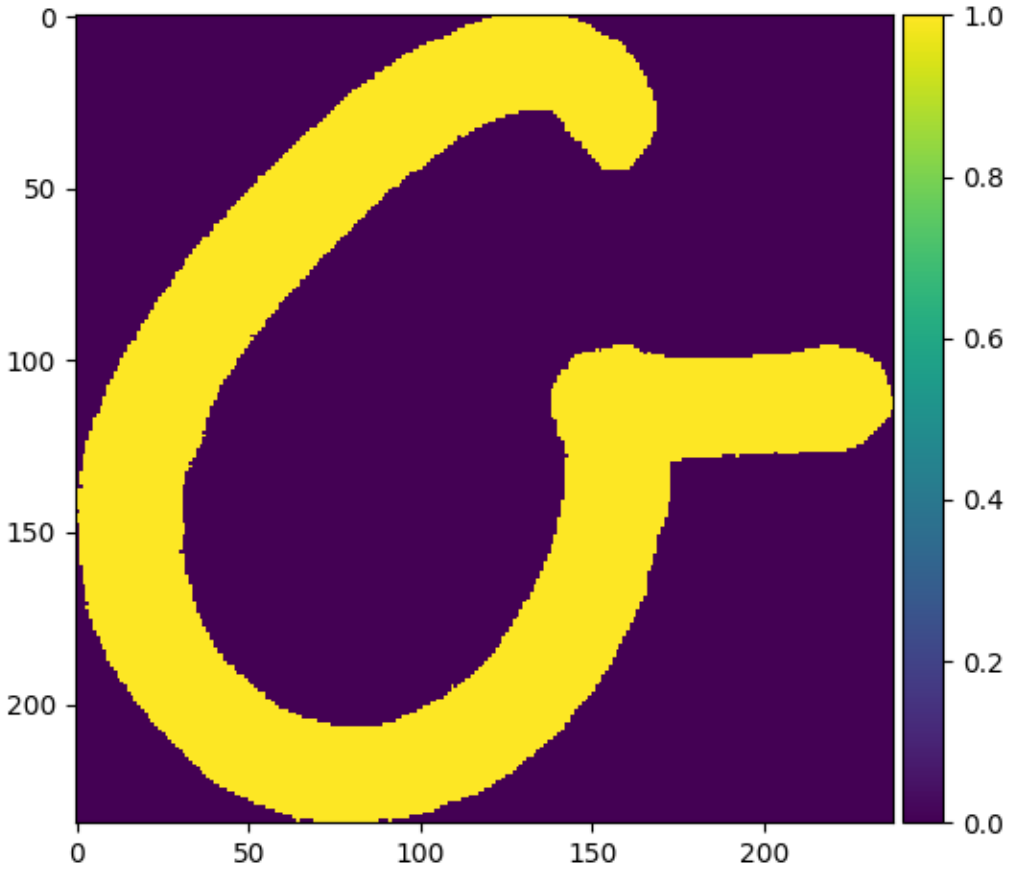

Figure 3: Connected component 'G' from utes.jpg

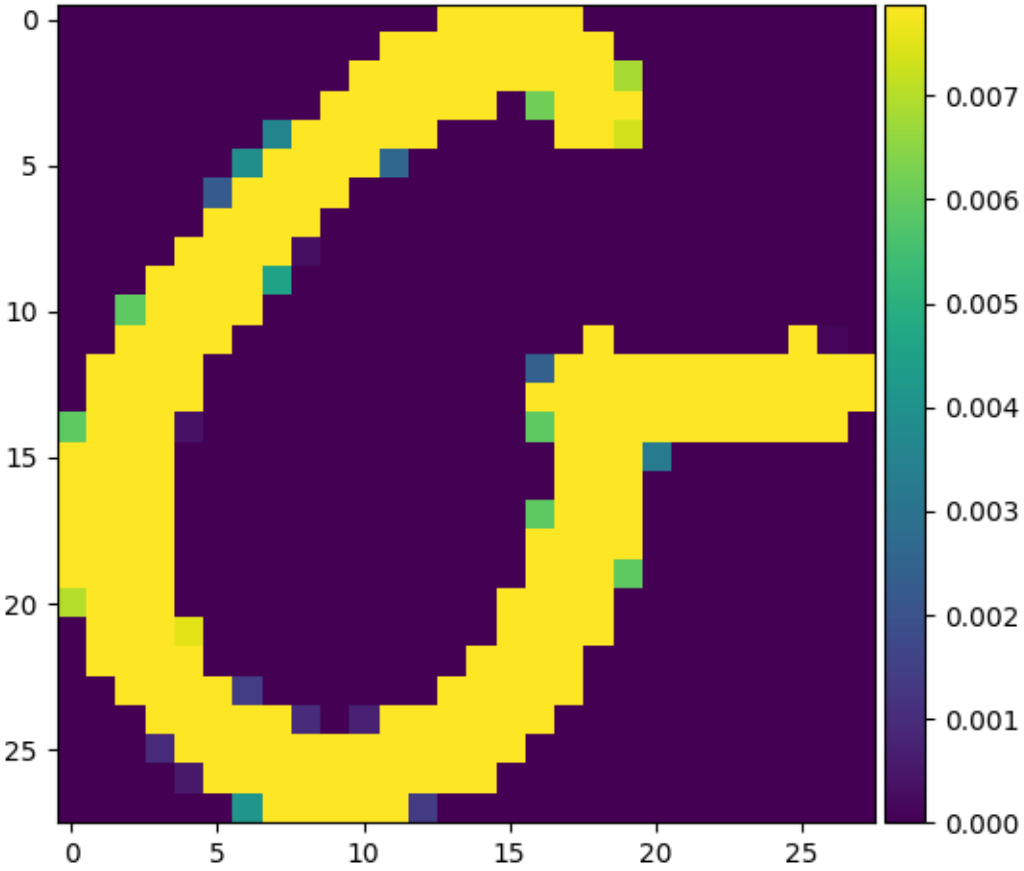

Figure 4: Rescaled input to classifier 'G' from utes.jpg

noisy_one_paragraph.jpg
Error = 141/765 characters = 18.4%

Issues with O-D, K-R, L-I, T-I, M-W, L-M, W-T, N-M, W-V, Q-O, etc.
Prediction/ truth

```
LDWCAKBDHYDRATEDLEISHAVEBECDWELNCREASLNGLYPDPULARSUPPDRTERSCLALWIHEYARE - 14 wrong
LOWCARBOHYDRATEDIETSHAVEBECOMEINCREASINGLYPOPULARSUPPORTERSCLAIMTHEYARE

NDIABLYNDREEFFECTLWETHANQTHERDLEISFDRWEIGHTLDSSANDPRDWLDEDTHERHEALIH - 16 wrong
NOTABLYMOREEFFECTIVETHANOTHERDIETSFORWEIGHTLOSSANDPROVIDEOTHERHEALTH

BENEFLTSSUCHASLDWERBLDDDPRESSUREANDIWPKDVEDCHDLESTERDLLEVELSHDWEVERSQNE - 11 wrong
BENEFITSSUCHASLOWERBLOODPRESSUREANDIMPROVEDCHOLESTEROLLEVELSHOWEVERSOME

DDCIDRSBELIEVETHESEDIEJSCARRYPDTENIJALLDNGIERNHEALTHRISKSAKEVIEWDFTHE - 12 wrong
DOCTORSBELIEVETHESEDIETSCARRYPOTENTIALLONGTERMHEALTHRISKSAREVIEWOFTHE

AVAILABLERESEARCHLJTERATURELNDJCATESTHAILDWCARBDHYDRAIEDLEISAREHLGHLYEFFECTJVE - 11 wrong
AVAILABLERESEARCHLITERATUREINDICATESTHATLOWCARBOHYDRATEDIETSAREHIGHLYEFFECTIVE

FARSHDRTTERWWELGHILDSSBUTTHATTHEIRLDNGFERMEFFECTJVENESSJSNDFSLGNLFLCANTLY - 14 wrong
FORSHORTTERMWEIGHTLOSSBUTTHATTHEIRLONGTERMEFFECTIVENESSISNOTSIGNIFICANTLY

GREATERTHANDTHERCDNNDNDLETPLANSTHEJRLANGTERWEFFELTSDNCHDLESTERDLLEVELS - 13 wrong
GREATERTHANOTHERCOMMONDIETPLANSTHEIRLONGTERMEFFECTSONCHOLESTEROLLEVELS

ANDBLDDDPRESSUREAREUNKNDWNRESEARCHLLFERAIURESUGGESTSSDWEPDTENTIALFDR - 8 wrong
ANDBLOODPRESSUREAREUNKNOWNRESEARCHLITERATURESUGGESTSSOMEPOTENTIALFOR

NEGATLVEHEALTHDUTCDNESASSDCLAJEDWLTHLNCREASEDCDNSUWPILDNDFSATURAFEDFAITHIS - 17 wrong
NEGATIVEHEALTHOUTCOMESASSOCIATEDWITHINCREASEDCONSUMPTIONOFSATURATEDFATTHIS

CDNCLUSIDNPDLNTSTDTHELNPDRIANCEDFFQLLQWLNGABALANCEDNDDEKATEDFEF - 17 wrong
CONCLUSIONPOINTSTOTHEIMPORTANCEOFFOLLOWINGABALANCEDMODERATEDIET

APPRDPRIATEFDRIHELNDLVLDUALASWELLASIHENEEDFDRFURTHERRESEARCH - 8 wrong
APPROPRIATEFORTHEINDIVIDUALASWELLASTHENEEDFORFURTHERRESEARCH
```

Noisy_one_sentance.jpg
Error = 5/57 = 8.8%

Issue with O-D, N-M, L-I (which all look similar)
Prediction/ truth

```
YDUWILLFACENANYDEFEATSLNLLFEBUTNEVERLETYDURSELFBEDEFEATED
YOUWILLFACEMANYDEFEATSINLIFEBUTNEVERLETYOURSELFBEDEFEATED
```

We can note that the errors with the typed font seem to be with letters that look like one another, such as D and O. The classifier could probably predict these better, but further training on this font would be required.

msg_from_annie.jpg
Error = 1/25 = 4%

Issue with O-Q (which look very similar)
Prediction/ truth

```
BESURETO
BESURETO

DRINKYQUR
DRINKYOUR

OVALTINE
OVALTINE
```

This image was much closer in accuracy than the previous one. The classifier we were provided seems to have been trained on smoother letters without serifs.

---

utes.png
Error = 0/6 = 0%

No issues
Prediction/ truth

```
GOUTES
GOUTES
```