Methodology

The methodology followed to achieve the task of detecting lines of text for historical documents is given below.

In literature, we have found deep learning model called ARU-Net [1] that works on historical documents by detecting lines of text. The ARU-Net model performs pixel-labelling. For achieving this task, ARU-Net model is used that takes historical document image as input shown in figure 1 (a) and predicts lines of text and gives output in binarized image as shown in figure 1 (b).

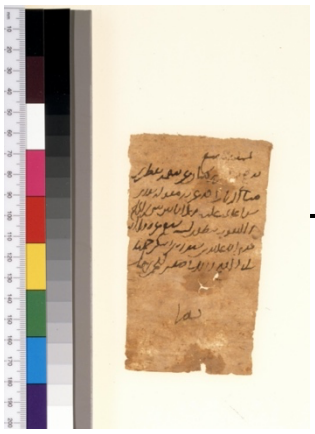

Figure 1 (a)

ARU-Net Model

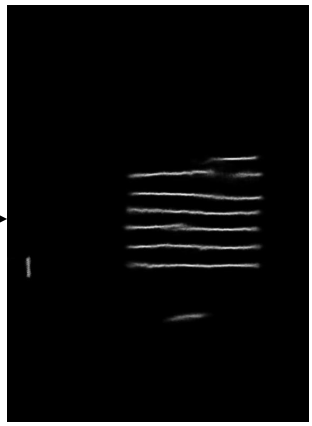

Figure 1 (b)

Once the output of ARU-Net is produced which is binarized image containing lines of text on that location in form of pixels, both original historical document image and binarized image are combined with image processing techniques to highlight lines below the text in original image. Two main post-processing tasks were performed to combine and blend image using image resize and weighted function from opencv library. Image resizing is performed to make sure both images are of same size so that both images can easily be combined. After that, two experiments are performed. Experiment one involves using addWeighted function in which weight for original image was 0.8 and for binarized image, it was 0.5 to make more visible line below the text. Experiment one involves selecting pixels with white color from predicted image are highlighted with red color on original image.

Results for these two experiments are mentioned below, however experiment two looks more suitable among these two.

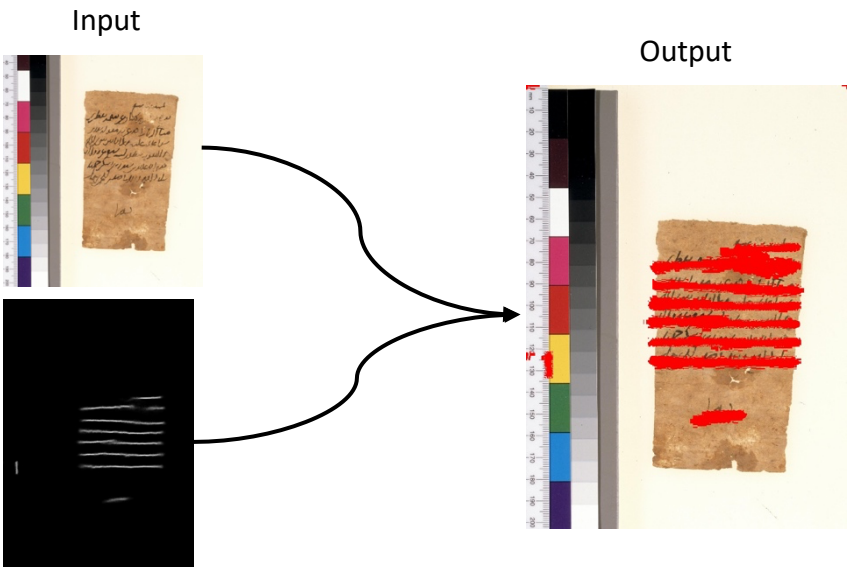

Figure 2: Experiment two, performed using pixel to pixel mapping for highlighting red color to each pixel in original image.

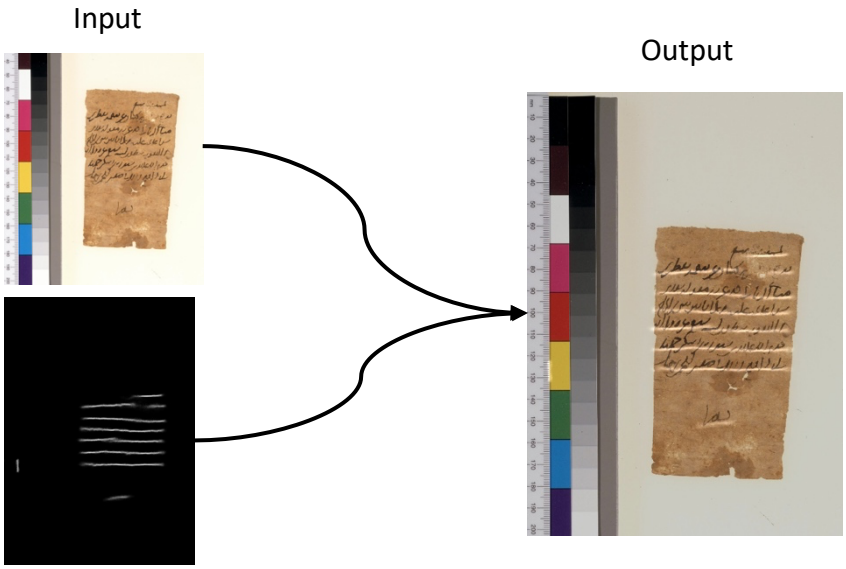

Figure 3: Experiment one, performed using addWeighted function to combine both images to highlight text in original image.

Challenges

Challenge 01: Most of these historical documents are dark and contains text whose pixel values are matching with paper pixels values so to separate both, it is much difficult.

- Solution: Need to perform background subtraction based on text and different operations will be involved, depending on different types of text

Challenge 02: Some images were different, so ARU-Net model didn't detect even single line as shown in following example.

- Solution: Need to train model on these types of images, for training, again we need dataset. To the best of my knowledge, this type of dataset is not available publicly.

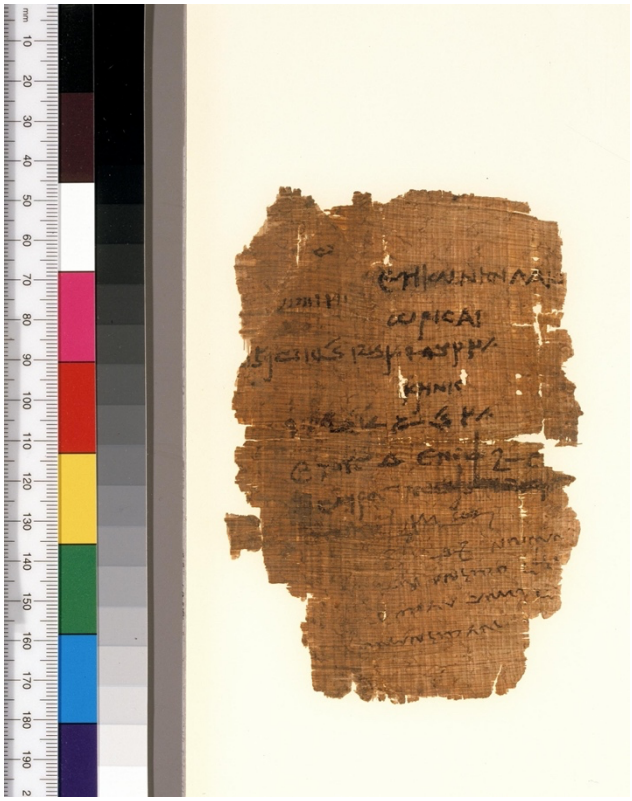

Figure 4: Example of historical document image, on which ARU-Net didn't detect even single line.

References

1. Grüning, Tobias, et al. "A two-stage method for text line detection in historical documents." International Journal on Document Analysis and Recognition (IJDAR) 22.3 (2019): 285-302. APA