Character Recognition- Document Analysis

Willem Hunkeler & Alexander Stewart

ISCS-380- Digital Image Processing- Keene State College

What is Document Analysis?

- Document Analysis is a vital component in day to day operations at any business or firm. It is used by many businesses to extract specific data that is relevant to existing documentation. Document analysis may also be necessary when stakeholders are not available to offer insight into existing business processes or systems.

- Image acquisition, in order to process a document on a computer one needs to convert the document into a numerical representation. Pixel-level processing include binarization, noise reduction, signal enhancement and segmentation.

- There are two categories of processing the data, textual processing (optical character recognition, page layout analysis) and graphical processing (line processing, region and symbol processing).

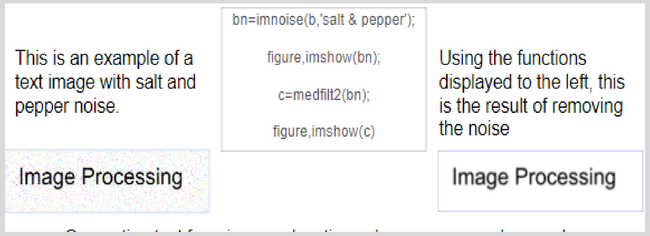
Preprocessing Techniques

- Binarization is to choose a threshold that separates the foreground and background information, i.g bank checks (handwriting and background image).

- Binarization- Involves converting the image into two values usually black and white. This is done to show the separation the background from the foreground.

Preprocessing Techniques

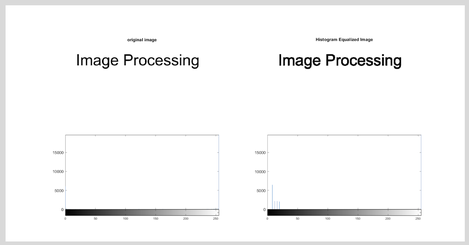
- Removal of salt and pepper noise using certain filters. In the example below we used median filtering to clear the text image.



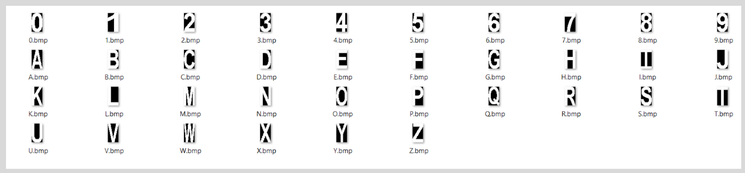
- Salt-and-pepper noise (also called impulse noise, speckle noise, or just dirt) is a common form of noise on a binary image.

Methods

- Separating text from images, locating columns, paragraphs, words, titles and captions. Retrieving intensity values of the text to capture the information. Comparing each pixel to one another to accurately retrieve what characters are recognized. The best way to find a global threshold is by the use of histograms. Hough transform is useful for line detection.



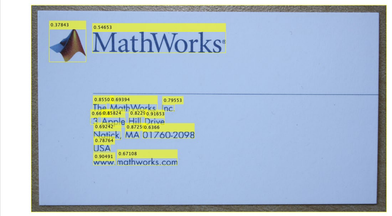
- Multiple templates can be used for detection. This allows several fonts that have different shapes, sizes, lower case, and uppercase characters to be detected.



- After the image is captured the characters are separated and compared to the available templates.

Conclusion

- Optical character recognition first finds the words on the image. The located words are then highlighted by each line as a located region.
- It is then separated by characters. The beginning and end of the characters are compared to each the available binary characters. An algorithm is used to determine the character it had the statically the most in common with. The measurement from the actual character (white pixels 0) to the white space is used (black pixels 1) is run through the available trained fonts.



- Only a good threshold will give the proper results, a low or high threshold will result in inaccurate character recognition. Low resolutions, blurry images, and handwriting can have an negative effect on the final character detection. It depends on how well the image can be processed.

- With OCR recognition rates now in the mid to high 90% range, and other document processing methods achieving similar improvements, these advances in research have also driven document image analysis forward.

References

- https://www.mathworks.com/help/vision/examples/recognize-text-using-optical-character-recognition-ocr.html
- http://www.cse.usf.edu/~r1k/DocumentImageAnalysis/DIA.pdf
- https://www.mathworks.com/help/vision/ref/ocr.html
- https://www.mathworks.com/help/vision/examples/automatically-detect-and-recognize-text-in-natural-images.html