BIRZEIT UNIVERSITY
Electrical and Computer Engineering Department
Computer Vision, Second Semester, 2020-2021
Assignment# 1

Text-Line Extraction for Arabic Handwritten Documents

Background:

Optical character recognition (OCR) is a system specially designed to convert text-images into an editable form. This conversion goes through six main sequencing stages: image acquisition, preprocessing, segmentation, feature extraction, recognition, and post-processing4. According to the segmentation stage, OCR systems can be classified into holistic (segmentation-free) and analytical (segmentation-based) approaches5,6, in which both approaches require segmentation stage at the text-line level. Therefore, it is important the correct extraction of the text-lines.

Text-line extraction is a basic and critical stage in developing OCR systems. The main objective of text-line segmentation is to identify all pixels that belong to a text-line. Although text-line segmentation for machine-printed text is often seen as a solved problem, text-line segmentation for handwritten Arabic text still presents a significant challenge. In addition, to the general (language independent) text-line segmentation challenges such as the quality of the input image (e.g. existence of noise, blurring, skewness, slant, and image degradation) and the layout complexity of the scanned image, Arabic text has a set of unique features that make the text-line segmentation challenging including: (i) Arabic text is written cursively from right to left, (ii) Arabic text has a large diversity of font types, which makes the shape (height and width) and the contour of the characters irregular and diverse, (iii) words are often divided into sub-words and letters which makes the spaces between them variable, (iv) characters in adjacent text-lines can be touching or overlapped, which treated as the major challenge in text-lines extraction, and (iv) the presence of diacritical components called "Harakat" increase the overlapping between characters in the adjacent text-lines. In fact, diacritics has an important role in the meaning of word in which same word with different diacritics leads to different meanings. They are widely used in religious documents (e.g. Quran), literature texts, and historical documents. Finally, Arabic calligraphy specifications contribute in making the empty spaces between every two-consecutive text-lines too narrow. This increases the complexity of determining the right segmentation points.

Task:

Write a code to extract the text lines from the scan documents. The input of your code is a scanned page and the output is a contour surrounding the segmented text lines. Note that you will not be allowed to used deep learning methods.

Related works:

https://www.researchgate.net/publication/341982308_Survey_on_Segmentation_and_Recognition_of_Handwritten_Arabic_Script

Data set:

VML-AHTE dataset: Arabic Handwritten Text Line Extraction dataset: is a natural handwritten benchmark dataset for text lines with crowdy diacritics, touching and overlapping characters. It is fully labeled at line level by native Arabic speakers. The dataset contains 20 training pages and 10 test pages. Every document image has a corresponding ground truth in the form of pixel labels and PAGE xml.

Evaluation measure:

To measure the performance of your proposed method, use the ICDAR2017 line segmentation evaluation metrics.
https://www.cs.bgu.ac.il/~berat/papers/icpr2020_unsupervised_deep_learning_for_text_line_segmentation.pdf

Submission:

1. Deadline: **1/5/2021**
2. Source code or link to GitHub
3. Comprehensive report that include the following:
    a. Introduction
    b. Related work
    c. Your approach step by step with sufficient examples on each step
    d. Results and discussion
    e. Comparing to other related methods
    f. Limitation and contribution