Text Extraction from Image with Eigenvalues

Bedir Halcı[1] and Alper Şişman[2]

[1]Marmara University, Istanbul, Turkey, ***bedirhalci@marun.edu.tr***
[2]Marmara University, Istanbul, Turkey, ***alper.sisman@marmara.edu.tr***

Introduction & Motivation

- Text extraction from image is openly used with open libraries with Machine Learning based database. But, it is slightly non-usable in memoryless self driven apparatus.
- It is doable and plausible to use Advanced Linear Algebra to extract such textual information directly without any database requirement. It can also be well-based in a sense that Fuzzy Logic can be abandoned.
- It is essential to have concrete Math behind any perceptual usage to create confidence against Statistics. Because Math is strictly well-prepared against extremes while Statistics just cause error against slight differences in inputs about their media.
- PCA is also another way for text extraction but it also heavily depends on strict “Test Group” as Machine Learning does. So it is not a well-based alternative to it.
- Our approach marks a brand-new epoch, because it differs from the others by directly using eigenvalues. It’s the real-extractor and the others are just recognisers.

Methodology

- Our methodology can be written like this in sequential order:
  (1) Taking RGB image and transforming it into grayscale.
  (2) Taking grayscale image and transforming it into binary.
  (3) Divide binary image into rows.
  (4) Divide rows into symbols by column-wise division to get matrix form.
  (5) Multiply the matrix form with its transpose to have a square matrix.
  (6) Extract eigenvalues from the square matrix.
  (7) Calculate correlations with the definitive eigenvalues of the structural archetypes of symbols.
  (8) Take maximum correlation to have extracted text.
- The eigenvalue extraction can be understood better with the schemes below.

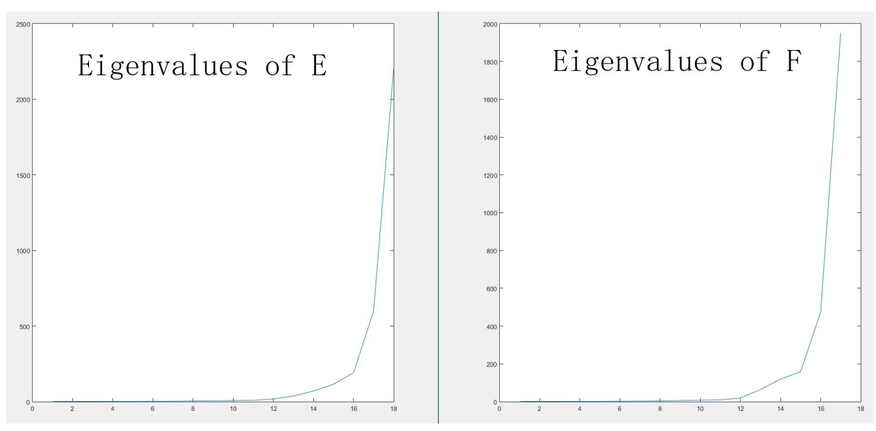
$$A_i(m_i \times n_i) \rightarrow B_i = A_i \times A_i^T$$

$$e_i = [z_i ; eig(B_i)]$$

$$z_i = [0 ; \dots ; 0]\ ((u - m_i) \times 1)$$

Experiments

- Eigenvalues of the structural archetypes of symbols “E” & “F” can be seen below.



- Text extraction example can be seen below.

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Loremipsumdolorsitamet,consecteturadipiscingelit,sed doeiusmodtemporincididuntutlaboreetdoloremagnaaliq ua.

Results

- There are positive & negative sides of this project: In positive side, text easily can be extracted in a certain punto range (between program minima & maxima) and also, it is consistent in itself; but in negative side, it starts to lose information as image text punto decreases, exponentially in round letters, and also, it begins to have singularity, as punto becomes smaller and smaller.

WILLIAMSHAKESPEARE
WILLIAMSHAKESPEARE
WILLIAMSHAKESPEARE
WILLIAMSHAKESPEARE
WILLIAMSHAKESPEARE
WILLIAMSHAKESPEARE
WILLIAMSHAKESPEARE

- At the end of the day, it is doing what it needs to do.

Future Works

- We have three main plan for future to develop and improve it:
  (1) To have size-free algorithm to avoid punto-based information loss. (Up-Down Interpolations & Wavelet Transform are in.)
  (2) To define even better eigenvalues of the structural archetypes of symbols to avoid similarity-based singularities and their consequences. (Font Media *status quo*)
  (3) To implement it into a FPGA-based apparatus to use in Antique Book Classification. (Realisation)
- In fact, we have started these three, even, we applied to Digital Design Contest 2019 with it; and we are enrolled.