Seminar on Digital Humanities

JAKOV ZINGERMAN

Lecture Structure

The Problem

Understanding solution’s toolbox

The solution

Results

A Simple and Fast Word Spotting Method

WRITTEN BY ALON KOVALCHUK, LIOR WOLF AND NACHUM DERSHOWITZ

Other use cases?

Book scanning

Converting handwriting in real time to control a computer

Data entry for business documents

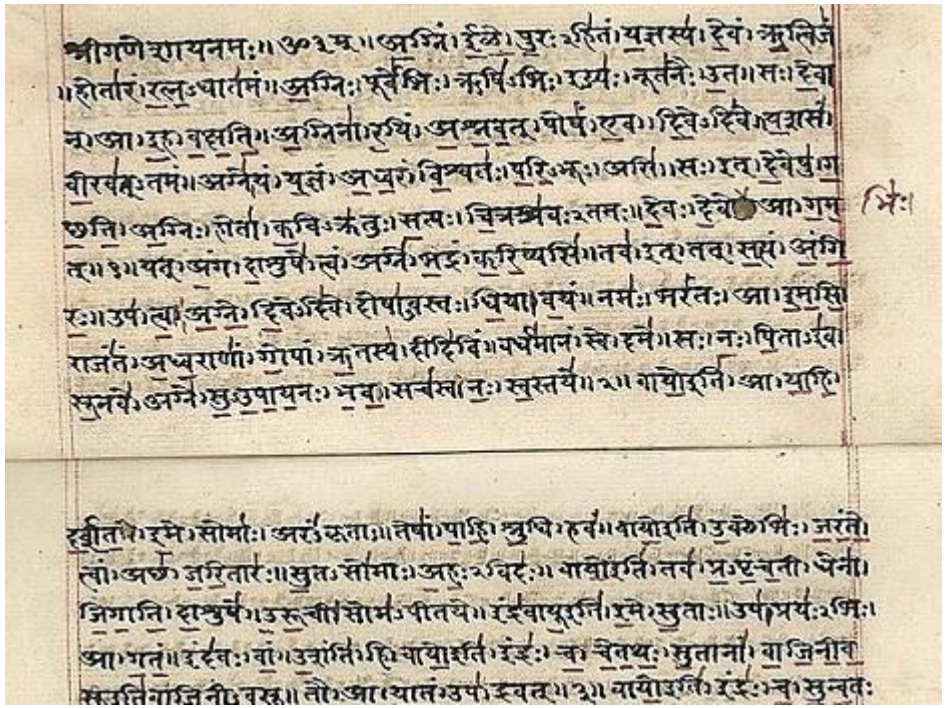
Archeological

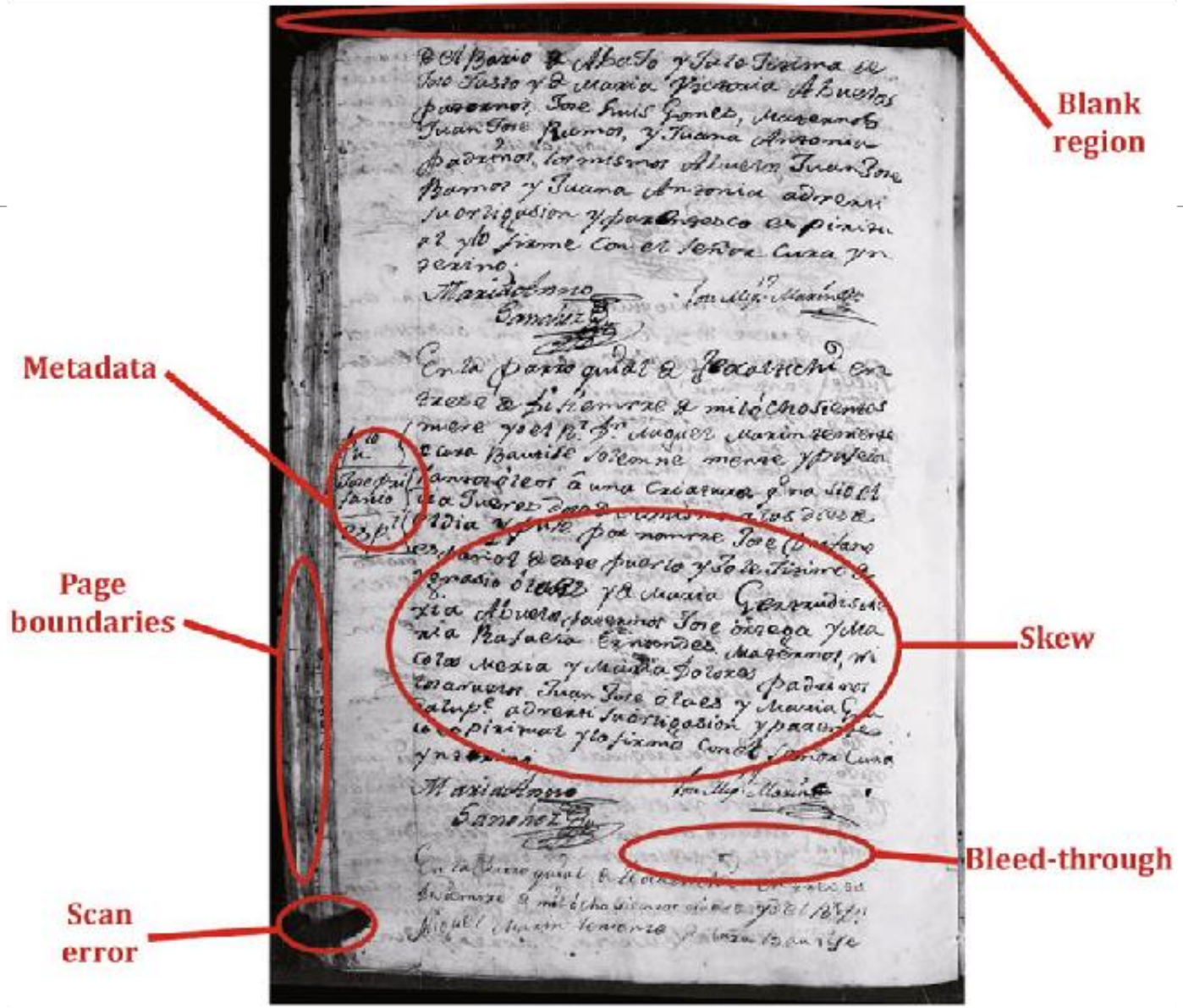
Why is it hard?

“Off-line handwriting recognition is comparatively difficult, as different people have different handwriting styles. And, as of today, OCR engines are primarily focused on machine printed text and ICR for hand "printed" (written in capital letters) text.”

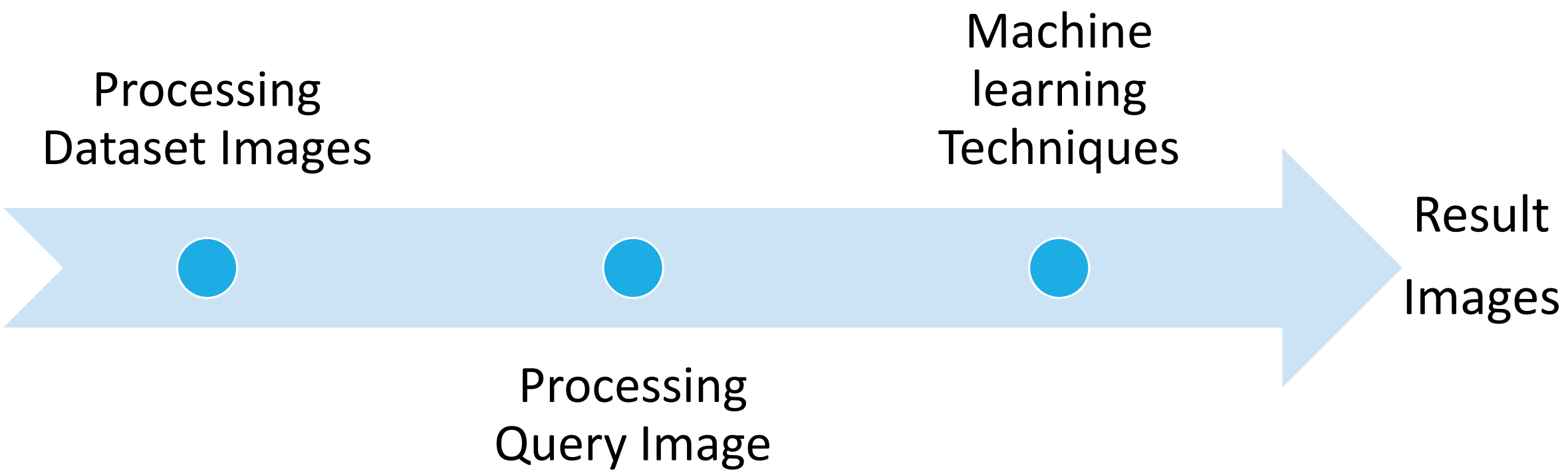
https://en.wikipedia.org/wiki/Handwriting_recognition

Why is it hard?



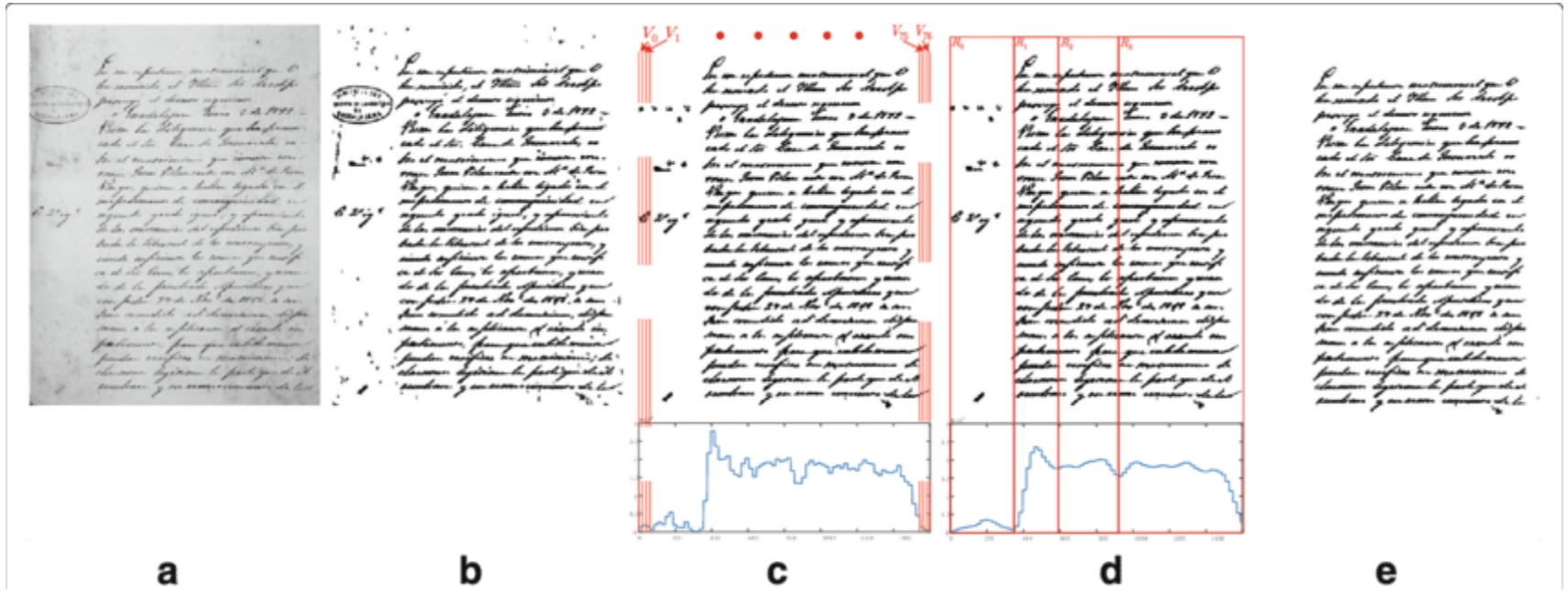

Blank region
Metadata
Page boundaries
Skew
Bleed-through
Scan error

Solution overview

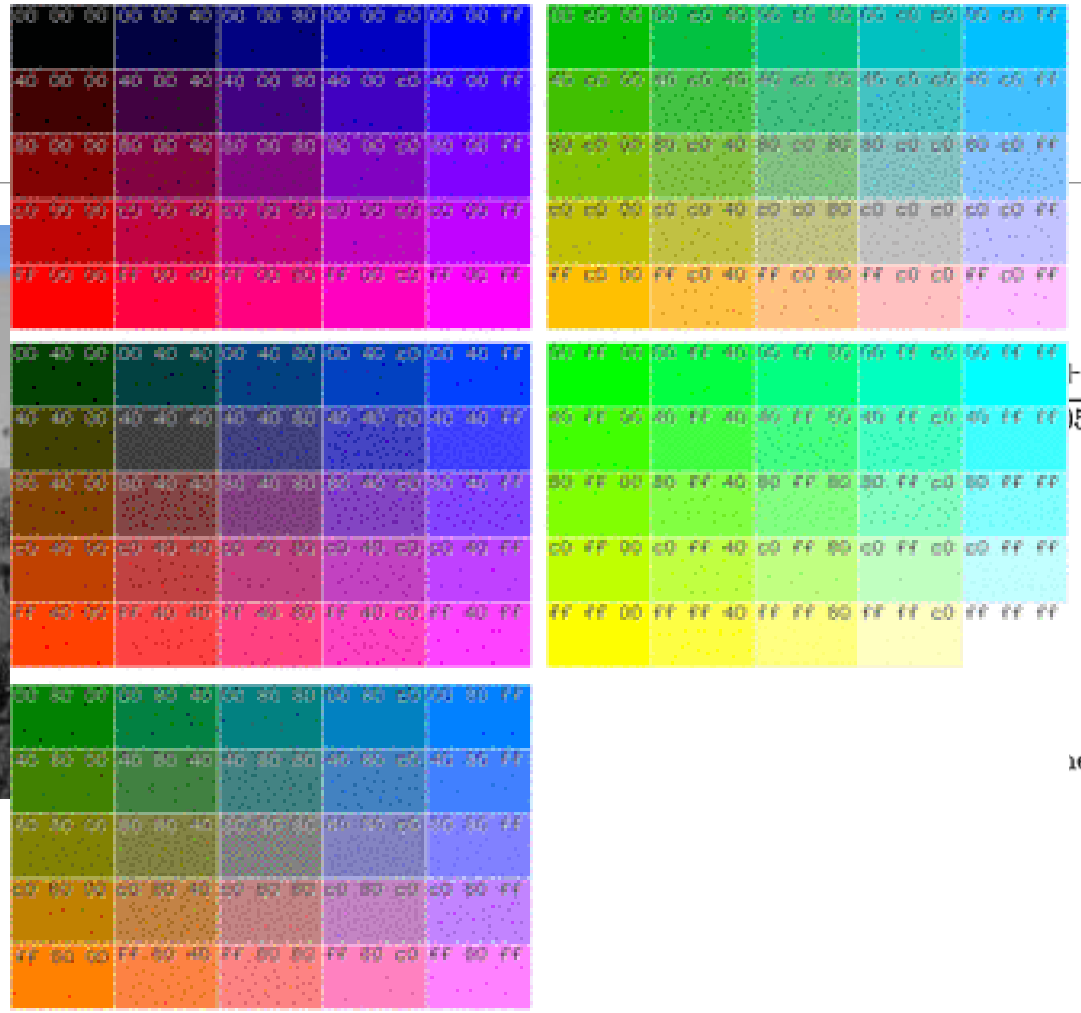


Let’s dive in

Binarization



Color to



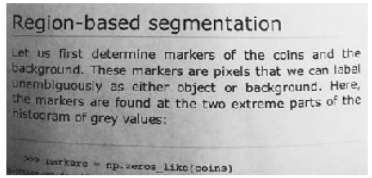
$$C_{\text{srgb}} \leq 0.04045$$

$$\left(\frac{\text{-0.055}}{\text{055}}\right)^{2.4}, \quad C_{\text{srgb}} > 0.04045$$

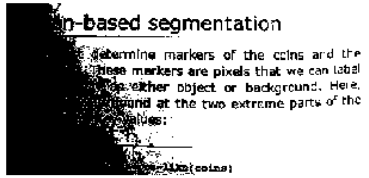
$$\text{ıear} + 0.7152G_{\text{linear}} + 0.0722B_{\text{linear}}$$

Thresholding

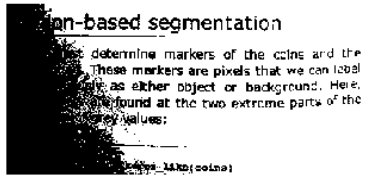
Original



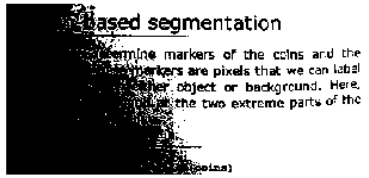
Isodata



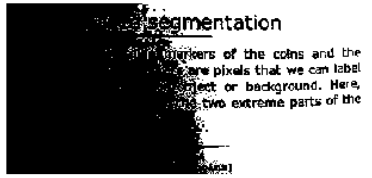
Li



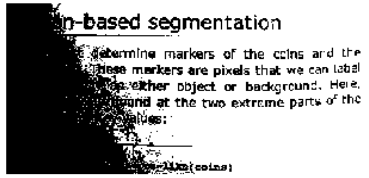
Mean



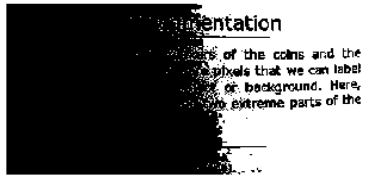
Minimum



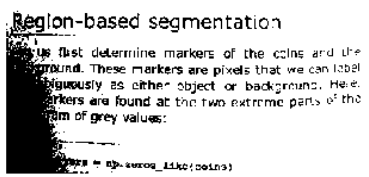
Otsu



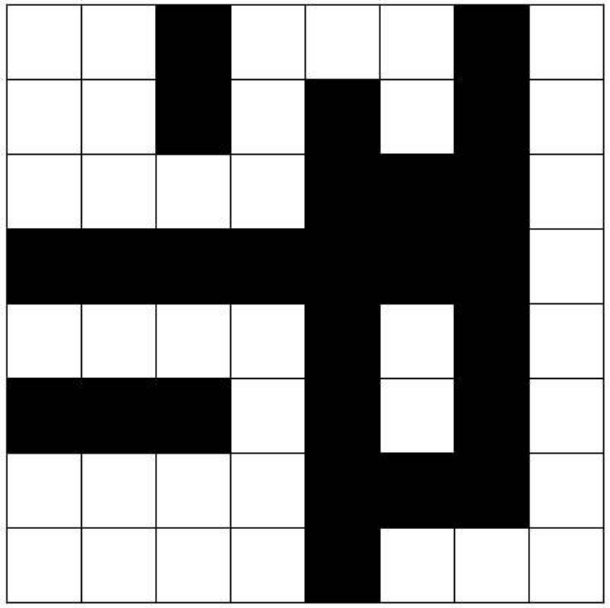
Triangle



Yen



Binary image

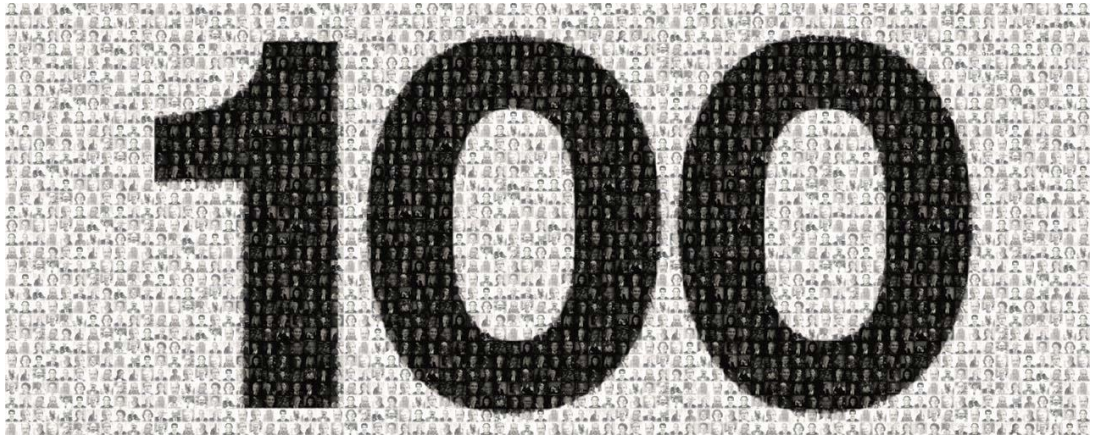


1	1	0	1	1	1	0	1
1	1	0	1	0	1	0	1
1	1	1	1	0	0	0	1
0	0	0	0	0	0	0	1
1	1	1	1	0	1	0	1
0	0	0	1	0	1	0	1
1	1	1	1	0	0	0	1
1	1	1	1	0	1	1	1

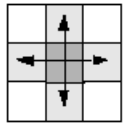
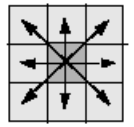
Algorithms Quiz

How can you find a connected component in an undirected graph?

How can we do that for a pixels matrix?

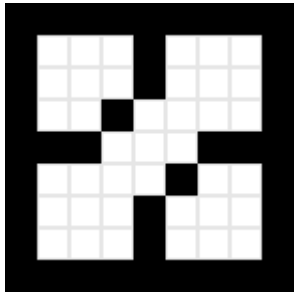


Finding Connected Components

Value	Meaning	
Two-Dimensional Connectivities		
4-connected	Pixels are connected if their edges touch. Two adjoining pixels are part of the same object if they are both on and are connected along the horizontal or vertical direction.	
8-connected	Pixels are connected if their edges or corners touch. Two adjoining pixels are part of the same object if they are both on and are connected along the horizontal, vertical, or diagonal direction.	

Finding connected components

4 connected:



8 Connected:

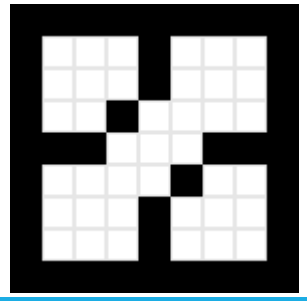


Image descriptors

Histogram of oriented gradients (HOG)

- Object detection in images

Local binary patterns (LBP)

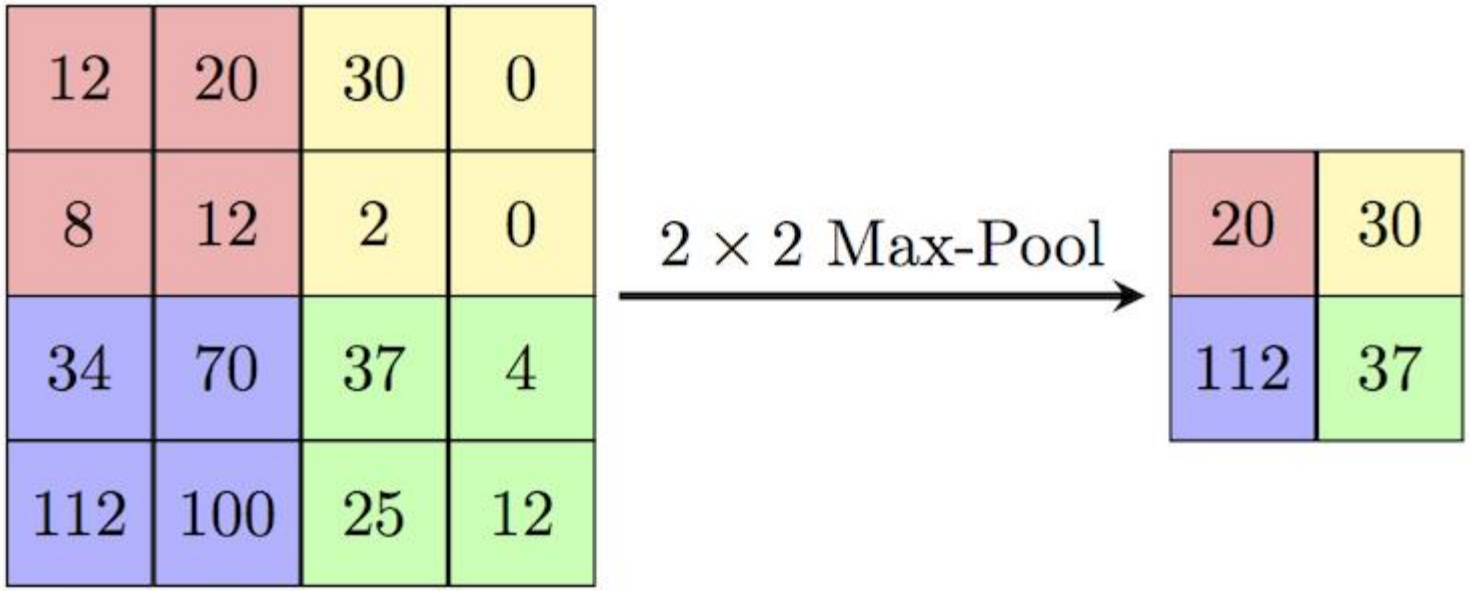
- Powerful for texture classification

Often used together

- improves the detection performance considerably on specific datasets

Max pooling

The process of down-sampling and input representation



Basic Machine learning

The past provides information about the future

Supervised

Unsupervised

K-nn Algorithm

The solution

Approaches

Simplicity vs. Efficiency and Effectiveness

Modularity

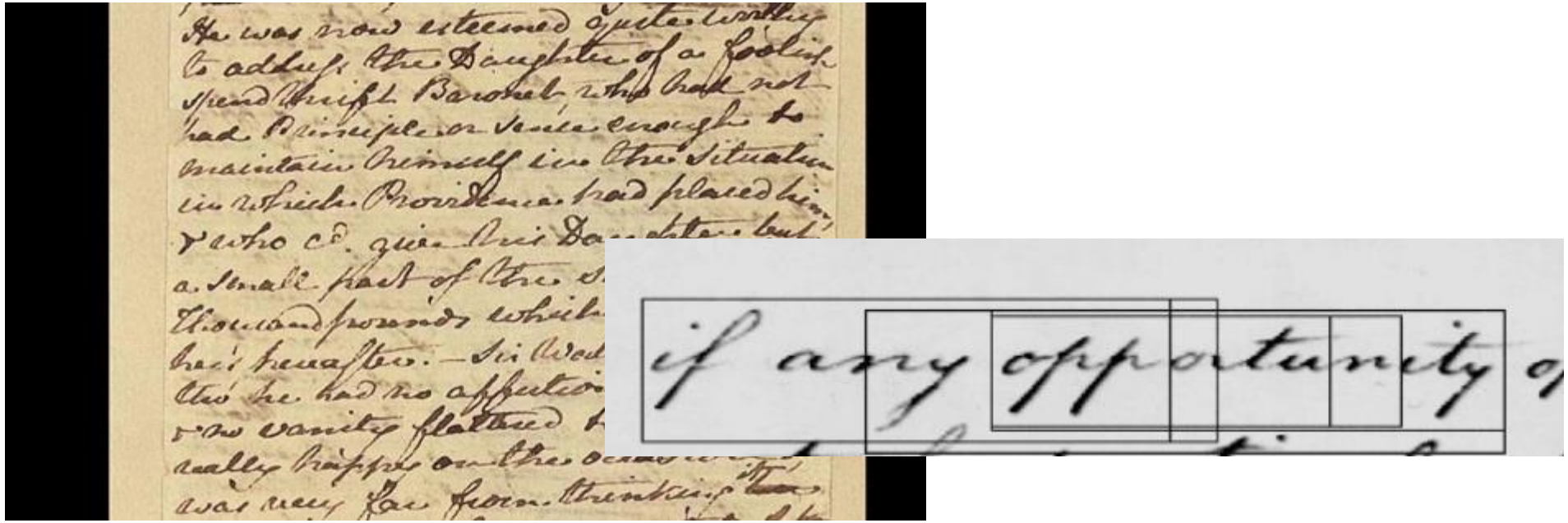
Scalability

The solution

Preprocessing Dataset Images

- Binarization
- Finding Word-like targets

Overlapping Candidates

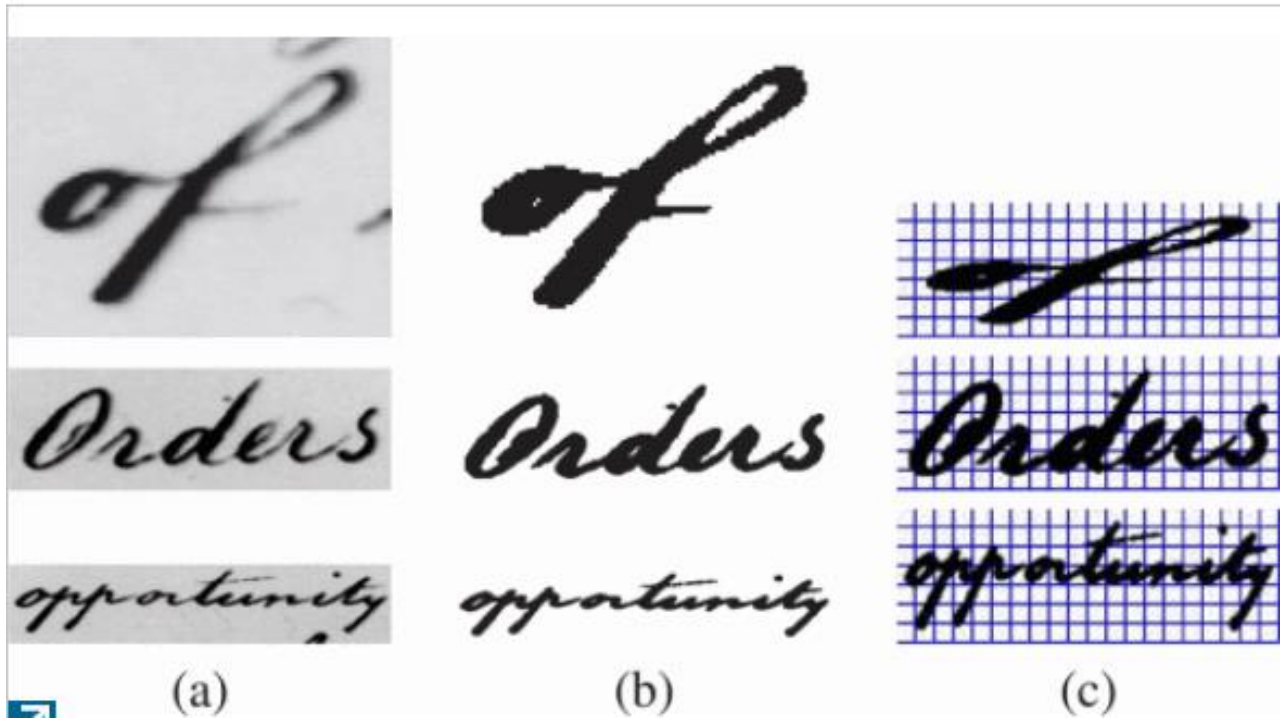


The solution

Preprocessing Dataset Images

- Binarization
- Finding Word-like targets
- Resizing potential targets
- https://mrl.nyu.edu/~dzorin/ig04/lecture08/lecture08.pdf

What we've got so far



The solution

Preprocessing Dataset Images

- Binarization
- Finding Word-like targets
- Resizing potential targets
- Calculating image descriptors
- Normalization
- Max-pooling

The solution

Processing query image

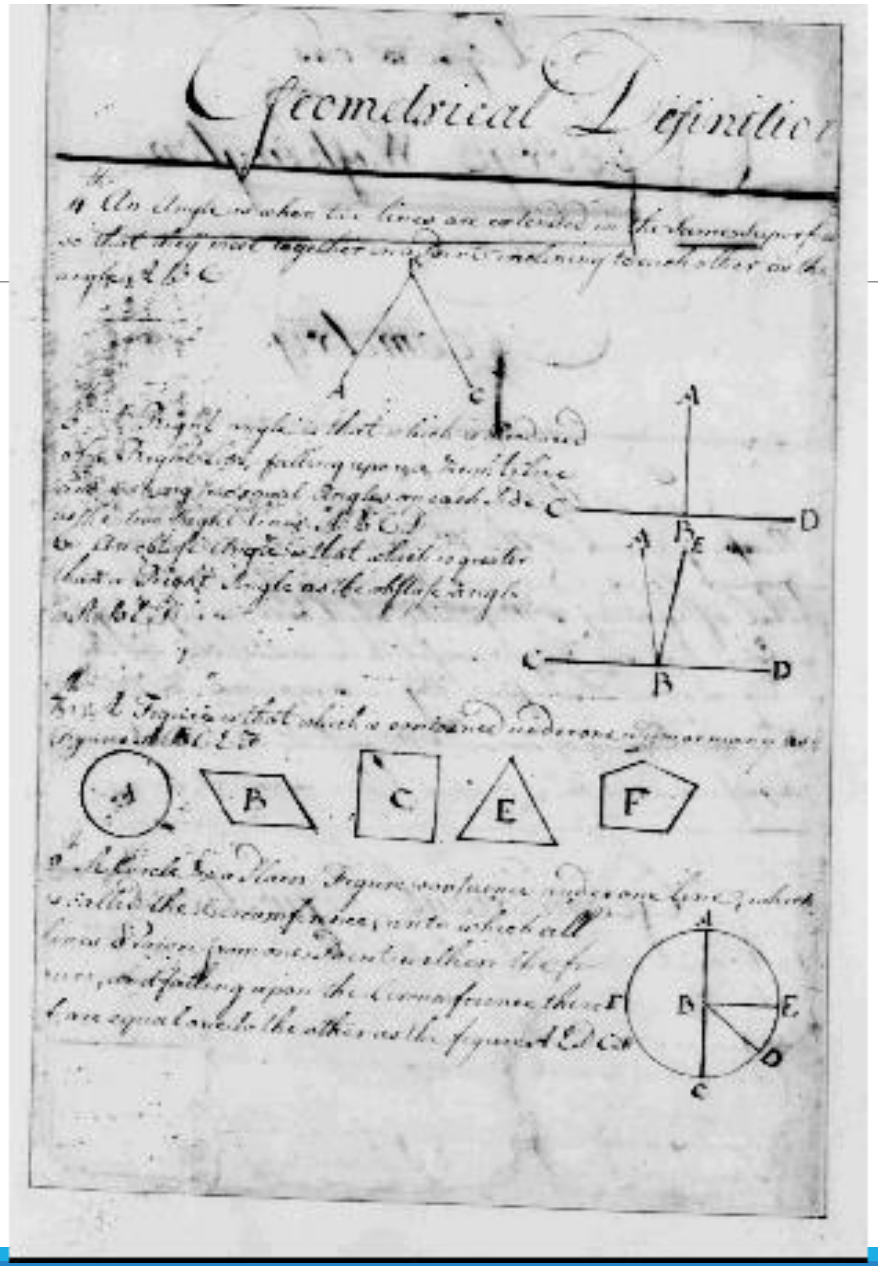
- Binarization
- Resizing
- Calculating Image Descriptors
- Max-pooling

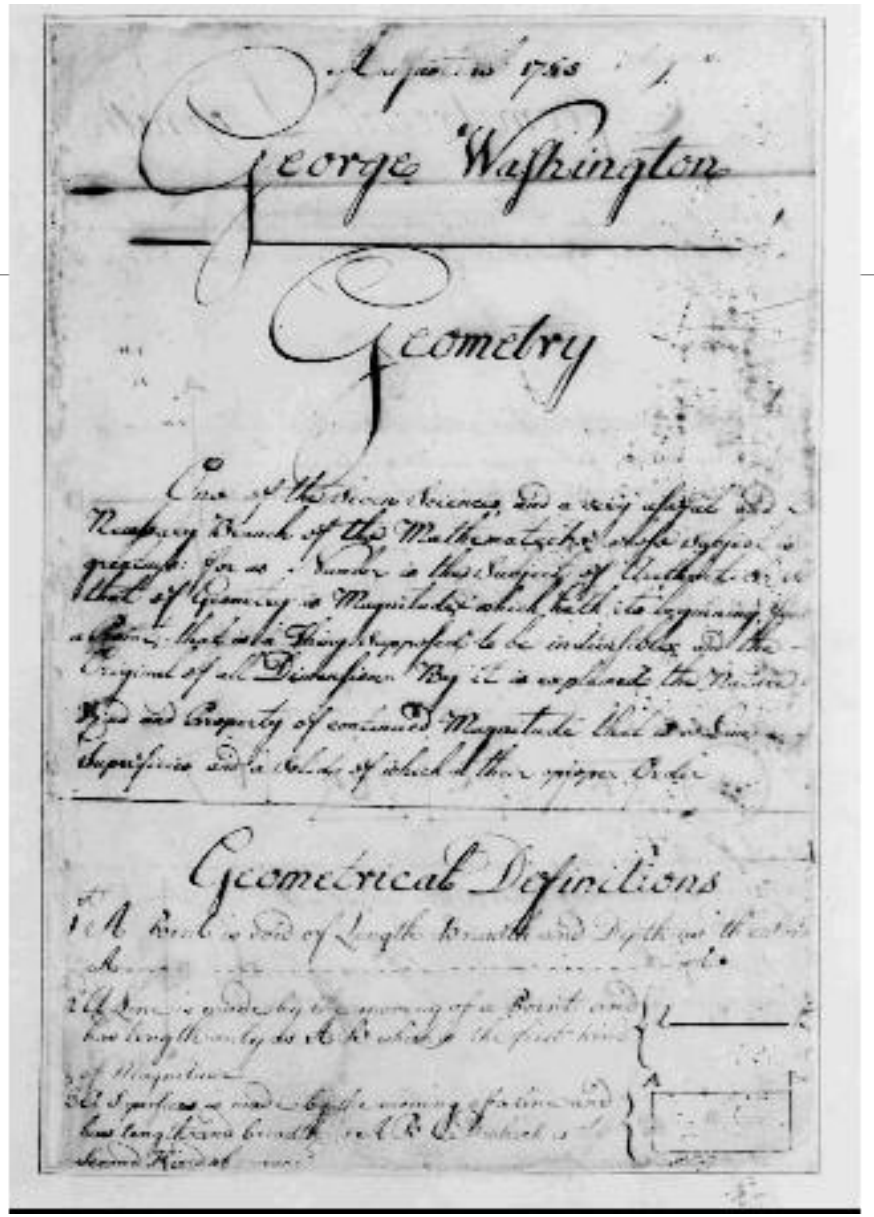
The solution

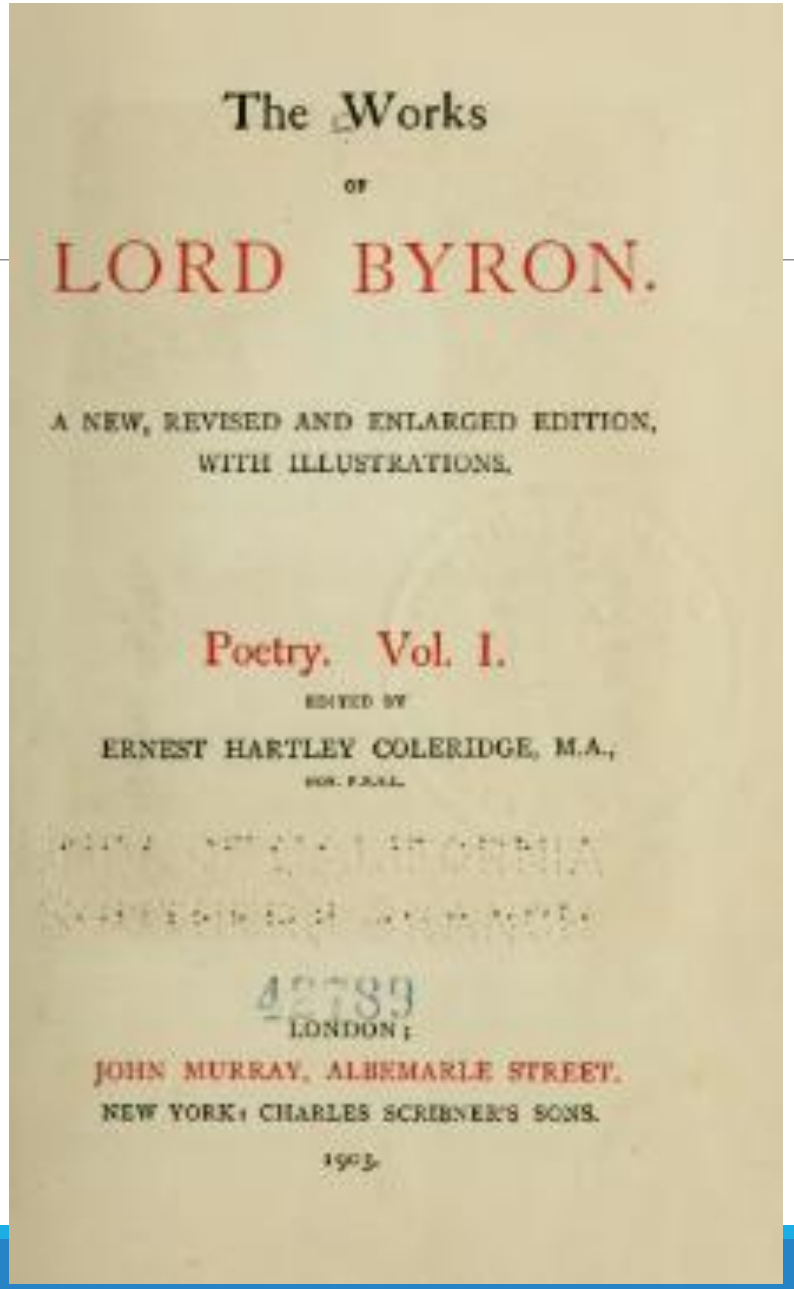
Calculating L2 distance between query and dataset images

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2}.$$

Picking the closest k images (KNN algorithm)


Geometrical Definitio
A
B
C
D
E
F


George Washington
Geometry
Geometrical Definitions



The Works

OF

LORD BYRON.

A NEW, REVISED AND ENLARGED EDITION,
WITH ILLUSTRATIONS.

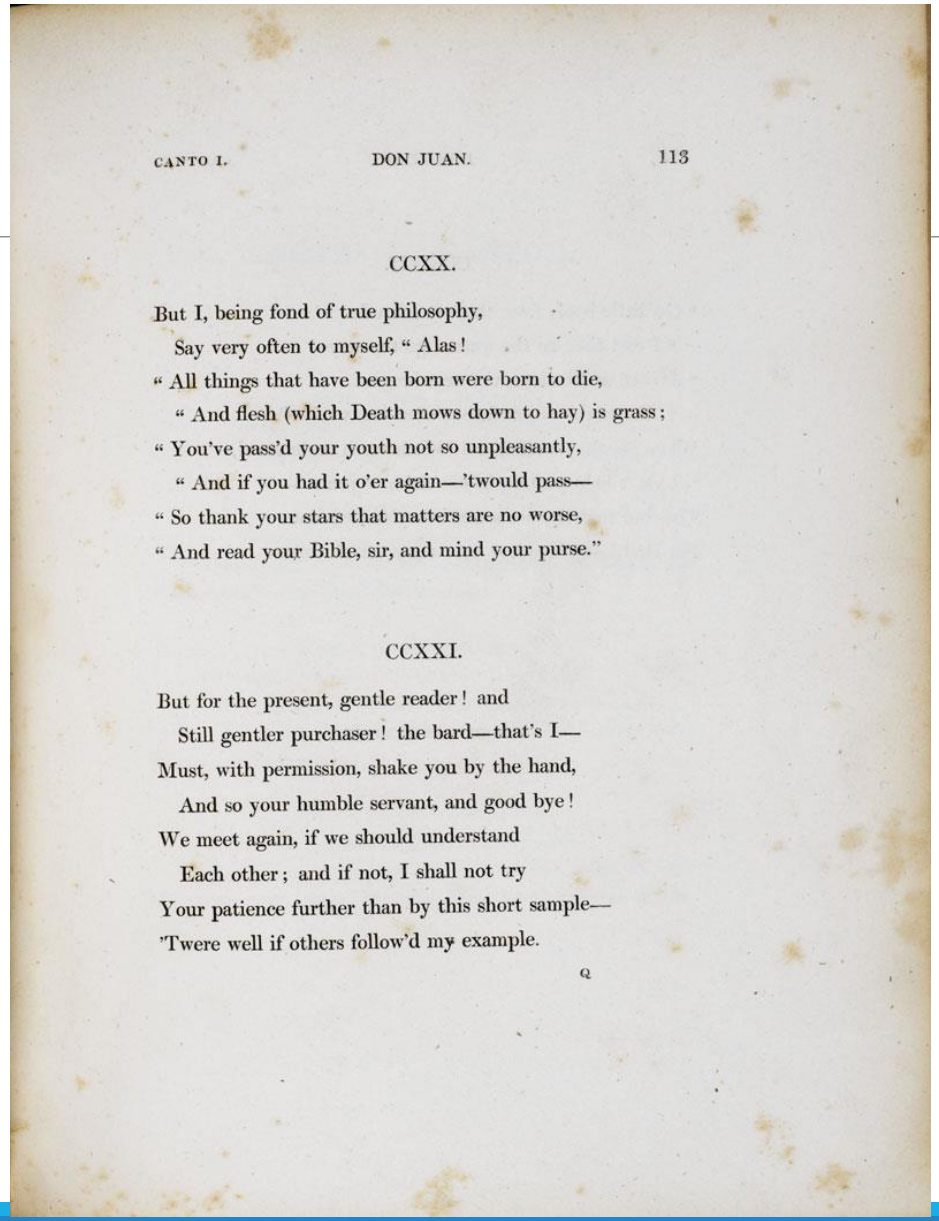
Poetry. Vol. I.

EDITED BY

ERNEST HARTLEY COLERIDGE, M.A.,

HON. FELLOW.

LONDON:
JOHN MURRAY, ALBEMARLE STREET.
NEW YORK: CHARLES SCRIBNER'S SONS.
1903.



CCXX.

But I, being fond of true philosophy,
Say very often to myself, "Alas!
"All things that have been born were born to die,
"And flesh (which Death mows down to hay) is grass;
"You've pass'd your youth not so unpleasantly,
"And if you had it o'er again—'twould pass—
"So thank your stars that matters are no worse,
"And read your Bible, sir, and mind your purse."

CCXXI.

But for the present, gentle reader! and
Still gentler purchaser! the bard—that's I—
Must, with permission, shake you by the hand,
And so your humble servant, and good bye!
We meet again, if we should understand
Each other; and if not, I shall not try
Your patience further than by this short sample—
'Twere well if others follow'd my example.

Method Results

2 Datasets

- The George Washington dataset
- The Lord Byron dataset
- ~20 pages, ~5,000 Words each

Table II
RUN TIME STATISTICS

Method/component	GW	LB
Number of queries	4,860	4,988
[2] all queries	5,058sec	4,159sec
[2] average per query	1.04sec	0.83sec
Our, all queries	158sec	46sec
Our, average per query	0.033sec	0.009sec
Our, single query	0.08sec	0.03sec
Preprocessing one page (ours)	46sec	3sec
Average memory per page (ours)	1,875KB	136KB

Table I
MEAN AVERAGE PRECISION FOR VARIOUS METHODS.

Method	**GW**	**LB**
Efficient exemplar word spotting [2]	54.5%	85.5%
Segmentation-free word spotting [5]	30.5%	42.8%
Complete pipeline	50.1%	90.7%
Same applied to segmented words	66.3%	92.9%
Without max-pooling ($v \in \mathbb{R}^{3750}$)	48.8%	90.8%
Without max-pooling ($v \in \mathbb{R}^{250}$)	47.6%	90.7%

Table II
RUN TIME STATISTICS

Conclusion

Word spotting is a useful substitute to OCR

Simplest method that can provide “state-of-the-art” results

Results can be improved in each of the steps

Potential of the method

Thank you for listening