

Seminar on Digital Humanities

JAKOV ZINGERMAN

A solid blue horizontal bar at the bottom of the slide.

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



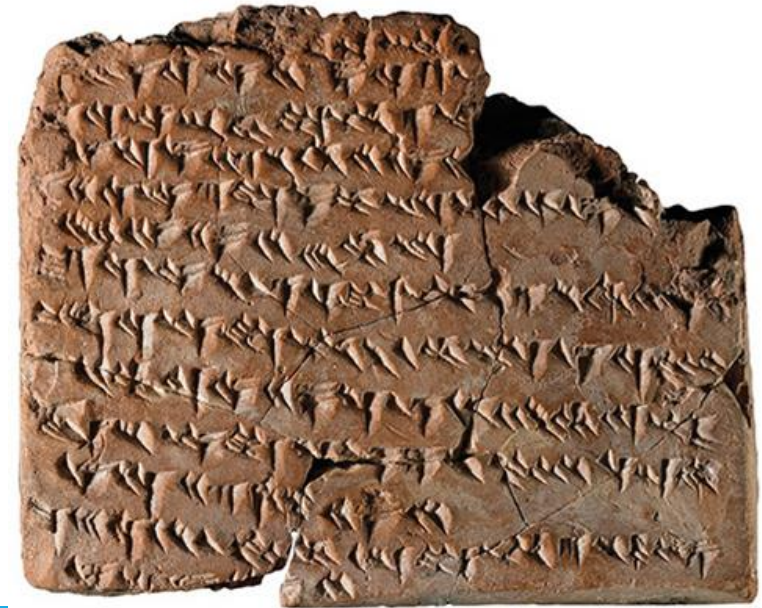
Other use cases?

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?

电话: 88310532

姓名: [redacted] 性别: [redacted] 年龄: 24岁

诊断:

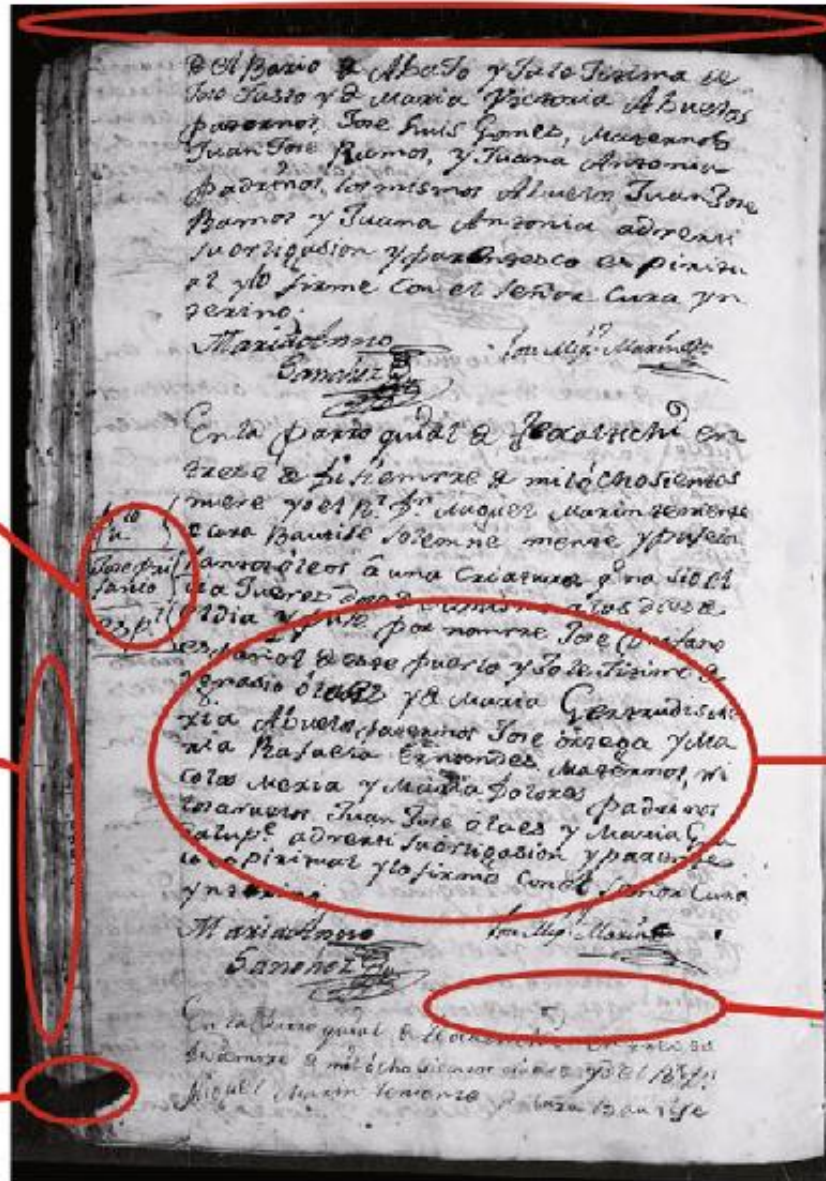
处方:

医生: [redacted]

weibo.com/gdzkb

श्रीगणेशायनमः॥ ॐ नमः ॥ अग्निं देवेभ्यः पुरः स्तितं यज्ञस्य देवं कृत्विति
 ॥ होतारं रत्नं पातमं ॥ अग्निः पूर्वभिः कृषिभिः इयं नृत्तैः उत ॥ स देवा
 न् ॥ आ ॥ इत् ॥ वसति ॥ अग्निना ॥ इयं ॥ अश्वत् ॥ पोषं ॥ एव ॥ द्विवे ॥ द्विवे ॥ लवसं
 वीरवत् ॥ तमं ॥ अग्नेयं ॥ यत् ॥ अश्वत् ॥ विश्वतः ॥ परिभूः ॥ अति ॥ स ॥ इत् ॥ देवेषु ॥ ग
 छति ॥ अग्निः ॥ होता ॥ कृषिः ॥ ऋतुः ॥ सत्यः ॥ चित्रश्रवः ॥ रतमः ॥ देवः ॥ देवो ॥ आ ॥ गम
 त ॥ इयं ॥ अग्ने ॥ दासुषे ॥ तं ॥ अग्ने ॥ प्रदं ॥ कुरिष्यसि ॥ तव ॥ इत् ॥ तव ॥ सुसं ॥ अग्नि
 रः ॥ उप ॥ त्वा ॥ अग्ने ॥ देव ॥ देवे ॥ होता ॥ इत् ॥ वयं ॥ नमः ॥ भरतः ॥ आ ॥ इत् ॥ सि
 राजते ॥ अश्वराणां ॥ गोपां ॥ कृतस्य ॥ होतारं ॥ यर्षमानं ॥ स्वे ॥ हम ॥ सः ॥ नः ॥ पिता ॥ इत् ॥
 सुनवे ॥ अग्ने ॥ सु ॥ उपाधनः ॥ भव ॥ सर्वसानः ॥ सुस्तय ॥ २ ॥ वायो इति ॥ आ ॥ युधिः

द्वीतरे ॥ इमे ॥ सोमाः ॥ अरं ॥ रुता ॥ तेषां ॥ पाहि ॥ शुचिः ॥ हव ॥ वायो इति ॥ उक्तेभिः ॥ जर्ते
 त्वां ॥ अर्कं ॥ जग्ति ॥ सुतः ॥ सोमाः ॥ अहः ॥ अर्कः ॥ वायो इति ॥ तव ॥ प्र ॥ इत् ॥ नृत्तैः ॥ येना
 जिगति ॥ दासुषे ॥ उरु ॥ वा ॥ सोमं ॥ पीतये ॥ इत् ॥ वा ॥ इति ॥ इमे ॥ सुताः ॥ उप ॥ प्रयः ॥ अग्निः
 आ ॥ गत ॥ इत् ॥ वः ॥ उजाति ॥ हि ॥ वायो इति ॥ इत् ॥ का ॥ चेतय ॥ सुतानां ॥ वा ॥ जिनीव
 स ॥ इति ॥ जिनीव ॥ वत् ॥ तौ ॥ आ ॥ यातं ॥ उप ॥ इत् ॥ ३ ॥ वायो इति ॥ इत् ॥ नः ॥ सुनवः



Blank region

Metadata

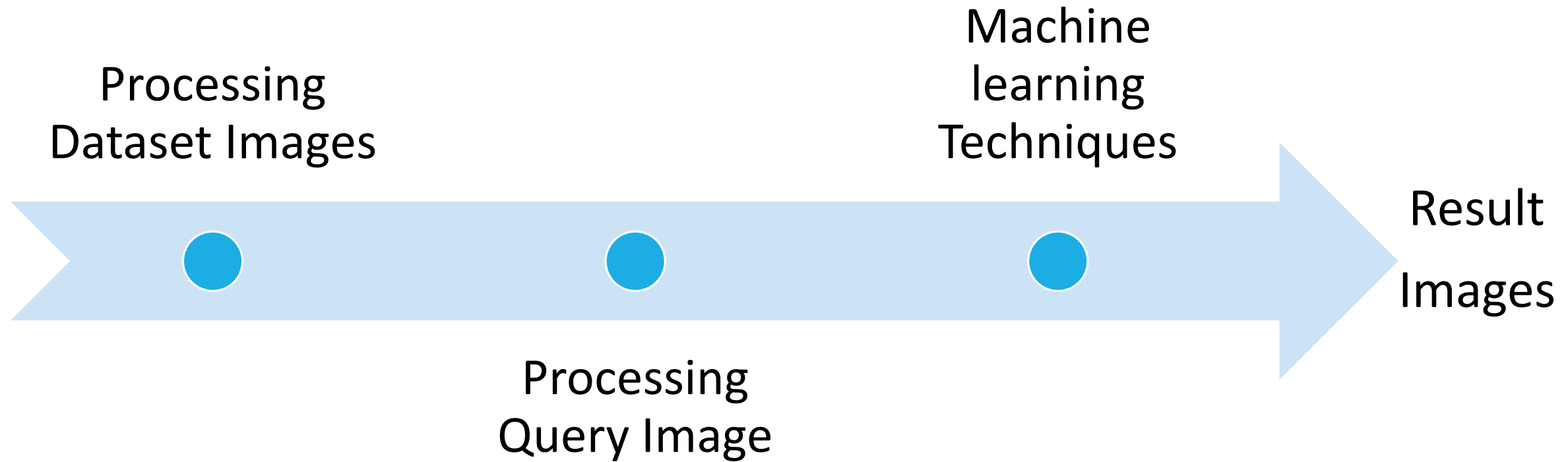
Page boundaries

Scan error

Skew

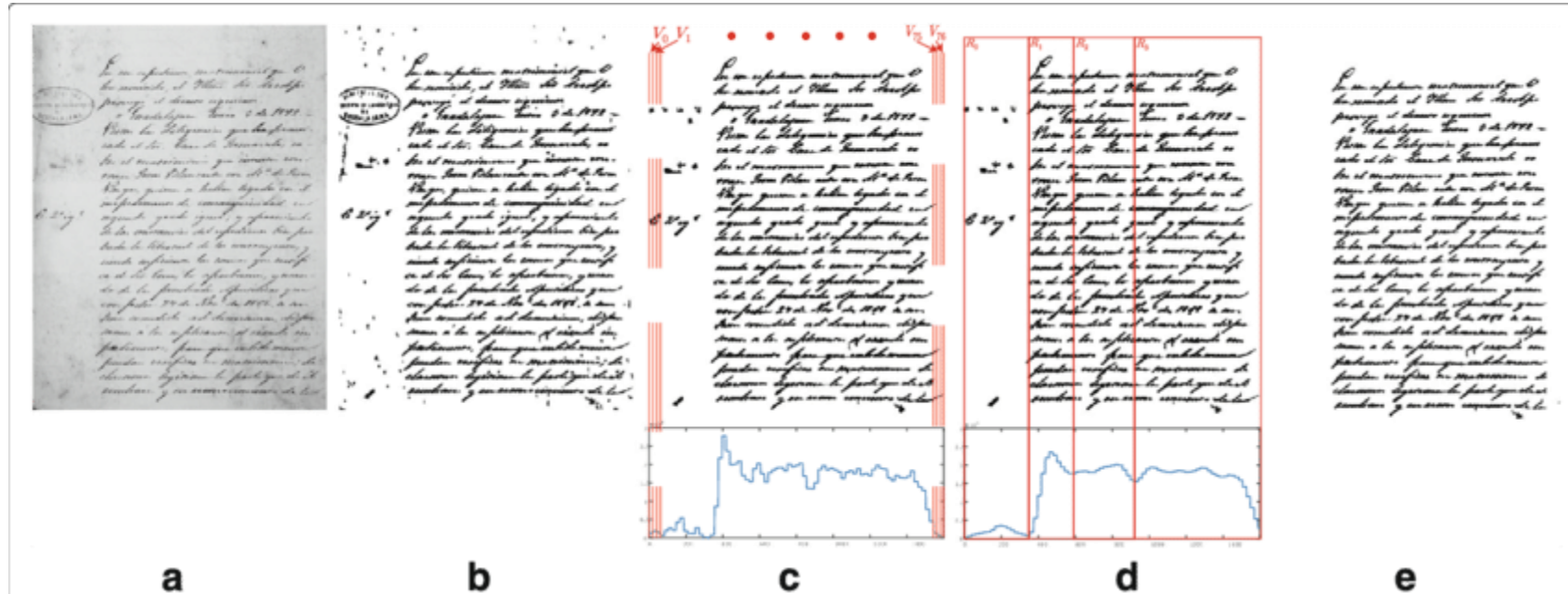
Bleed-through

Solution overview

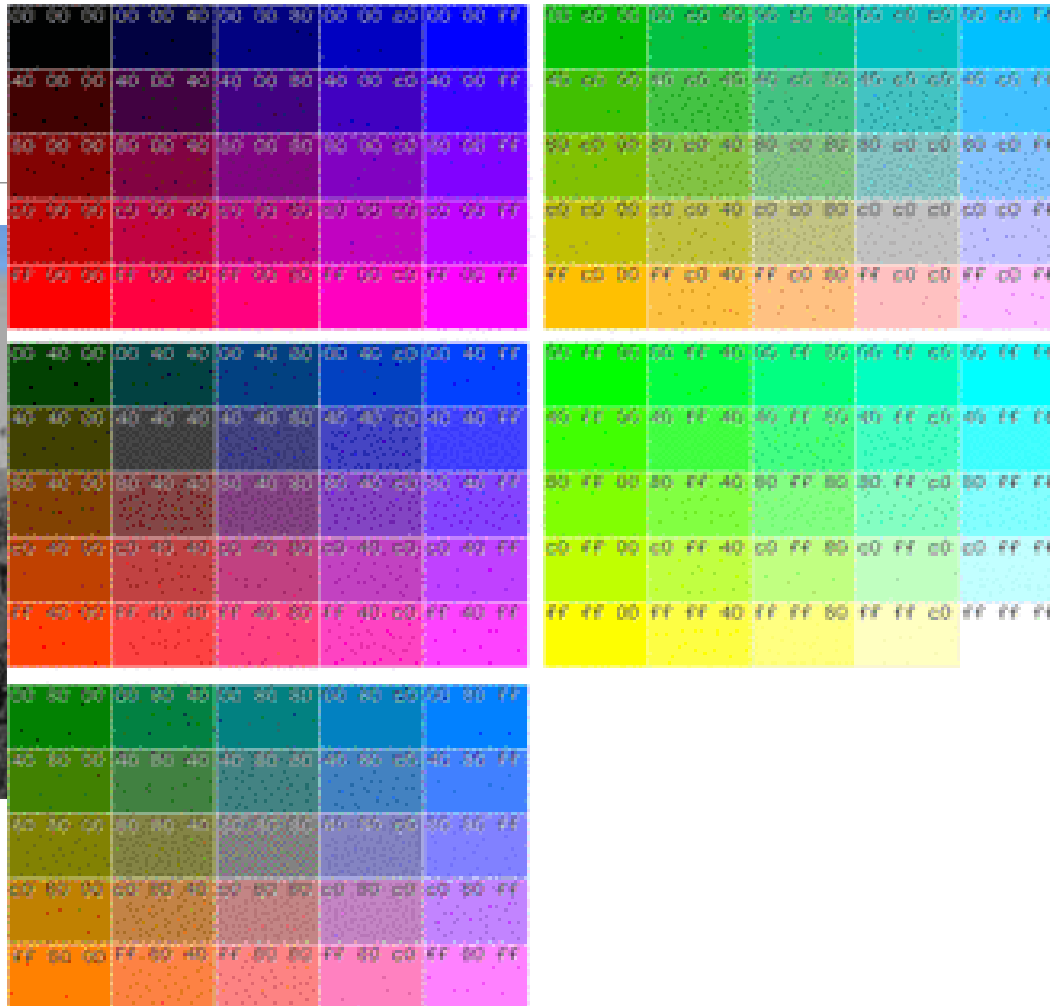


Let's dive in

Binarization



Color to

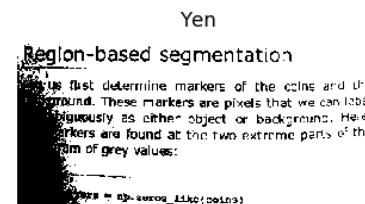
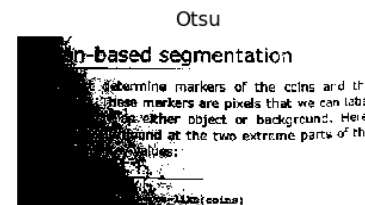
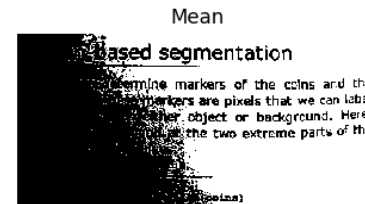
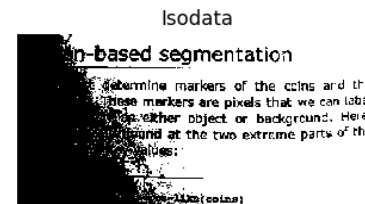
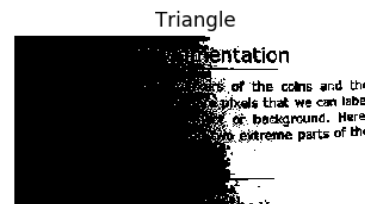
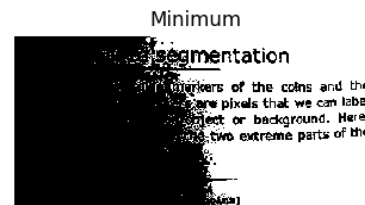
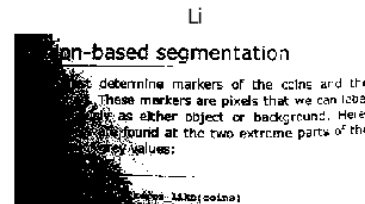
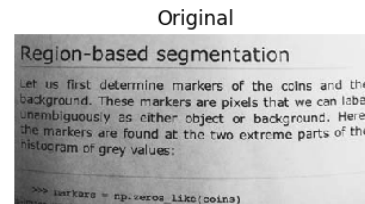


$$C_{\text{srgb}} \leq 0.04045$$

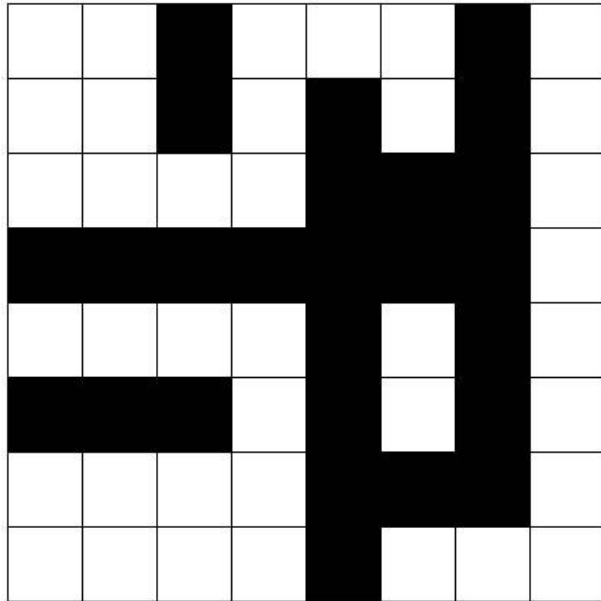
$$\left(\frac{C_{\text{srgb}} + 0.055}{1.055} \right)^{2.4}, \quad C_{\text{srgb}} > 0.04045$$

$$R_{\text{near}} + 0.7152G_{\text{linear}} + 0.0722B_{\text{linear}}$$

Thresholding



Binary image

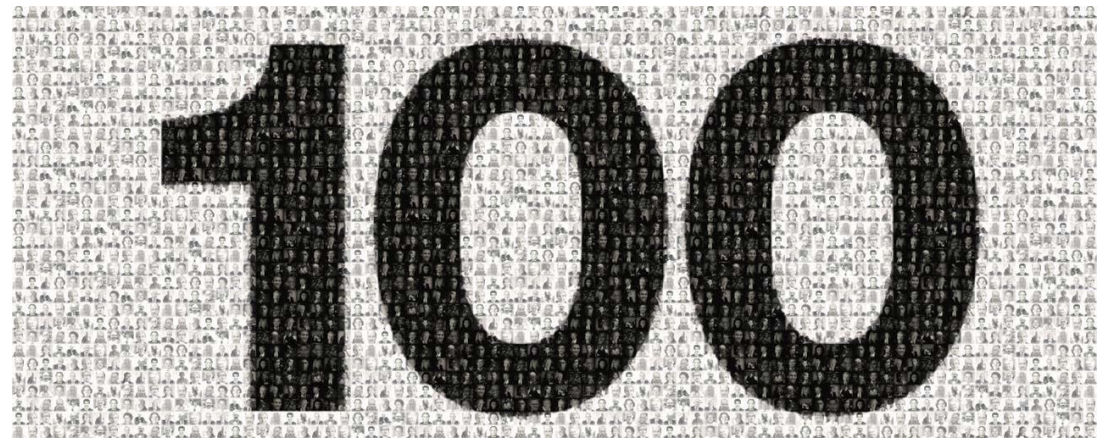


1	1	0	1	1	1	0	1
1	1	0	1	0	1	0	1
1	1	1	1	0	0	0	1
0	0	0	0	0	0	0	1
1	1	1	1	0	1	0	1
0	0	0	1	0	1	0	1
1	1	1	1	0	0	0	1
1	1	1	1	0	1	1	1

Algorithms Quiz

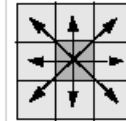
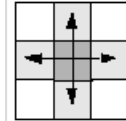
How can you find a connected component in an undirected graph?

How can we do that for a pixels matrix?



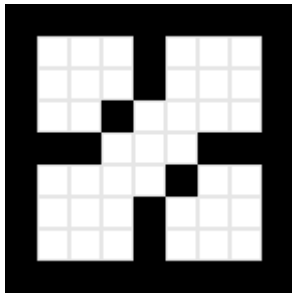
Finding Connected Components

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:

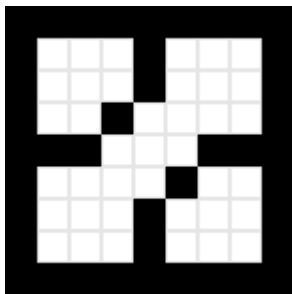


Image descriptors

Histogram of oriented gradients (HOG)

- Object detection in images

Local binary patterns (LBP)

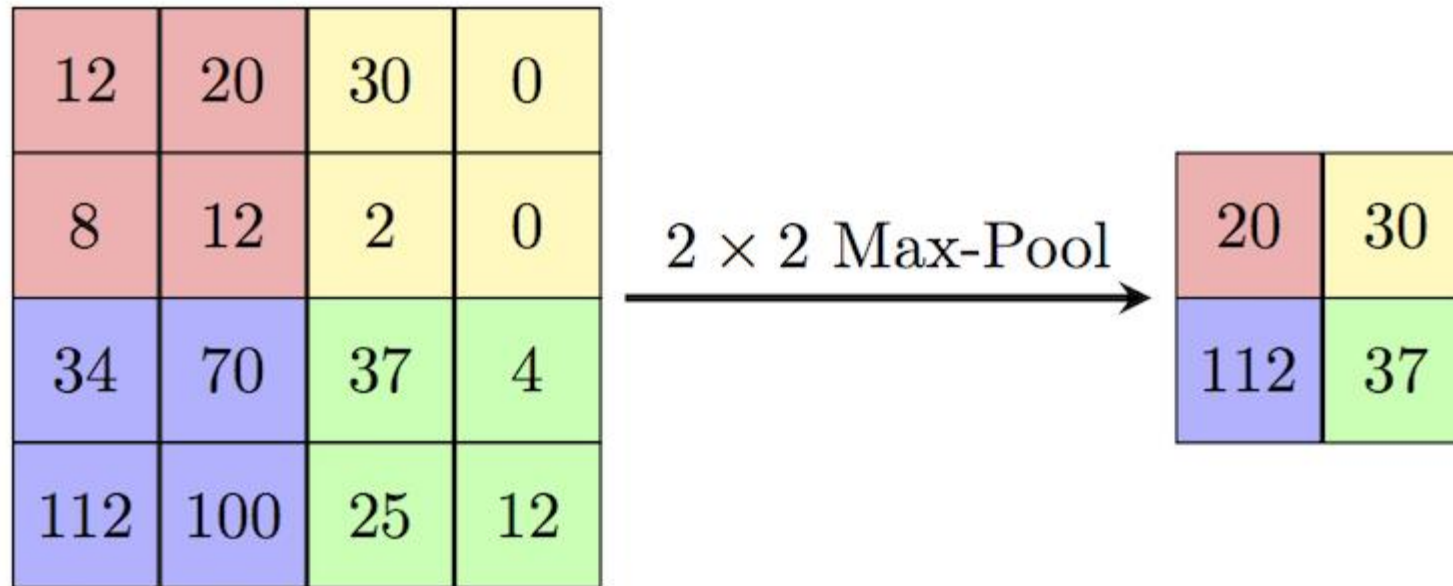
- Powerful for texture classification

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

He was now esteemed quite worthy
to address the Daughter of a foolish
spendthrift Baronet, who had not
had Principle or Sense enough to
maintain himself in the Situation
in which Providence had placed him,
& who c^d. give his Daughter but
a small part of the
Thousand pounds which
he had hereafter. — Sir Wood
tho' he had no affection
& no vanity flattered to
really happy on the occasion
was very far from thinking

if any opportunity of

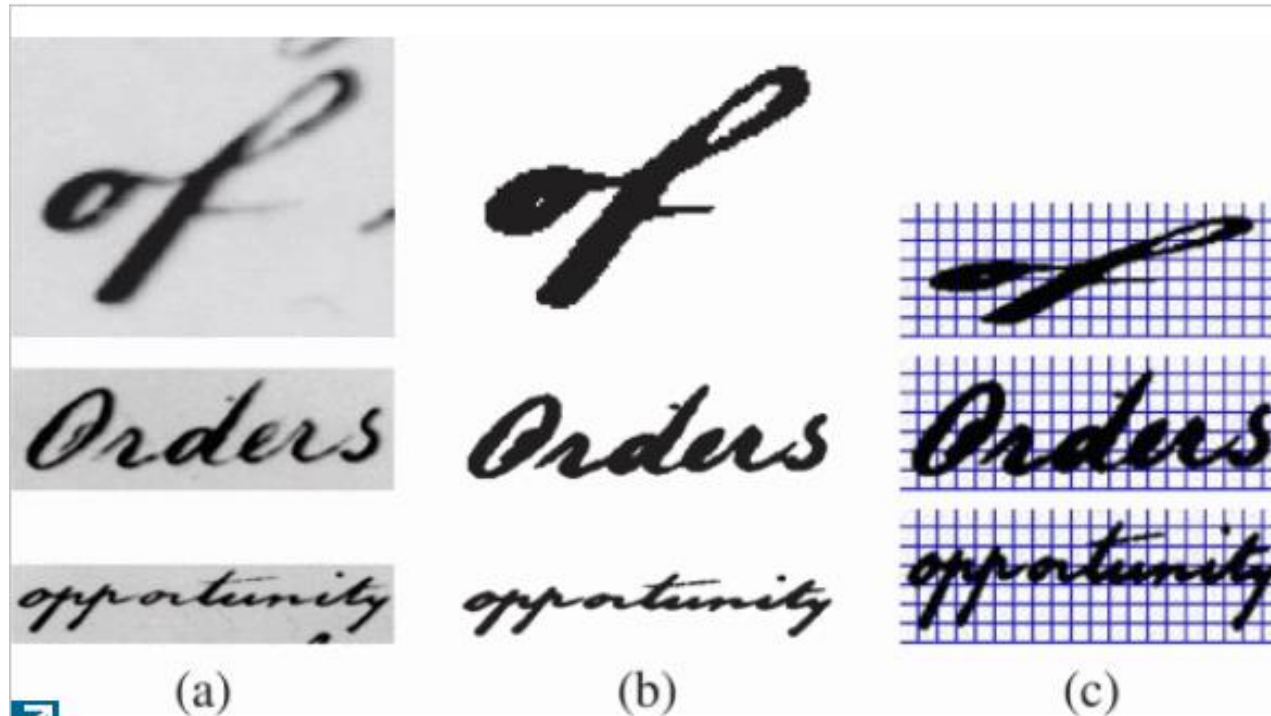
The solution

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

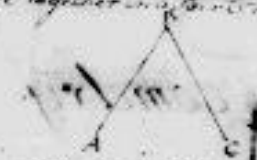
Calculating L2 distance between query and dataset images

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

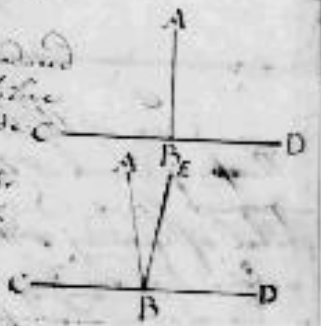
Picking the closest k images (KNN algorithm)

Geometrical Definitions

1. An Angle is when two lines are extended in the direction of a point so that they meet together and do not incline to touch either as the angles A & B & C



2. A Right angle is that which is contained of a straight line falling upon a straight line and making two equal angles or each. See C with the two right lines A & B & C



3. An obtuse angle is that which is greater than a right angle as the obtuse angle A & B & C

4. A Figure is that which is contained under one or more lines as the figures A & B & C & E & F



5. A Circle is a plain figure contained under one line which is called the circumference, unto which all lines drawn from a certain within the figure, and falling upon the circumference, there are equal to the other as the figure A & B & C



August 1788

George Washington

Geometry

One of the most ancient Sciences and a very useful and necessary branch of the Mathematics whose origin is ancient. For as Number is the Subject of Arithmetick, so that of Geometry is Magnitude which hath its originating from a point that is a thing supposed to be indivisible and the Original of all Dimensions. By it is explained the Nature, Uses and Properties of continued Magnitude that is a line, a superficies and a solid of which in their proper Order

Geometrical Definitions

1. A Point is that of Length breadth and Depth or that which has no dimensions.
2. A Line is made by the moving of a Point and has length only as A & B which is the first kind of Magnitude.
3. A Surface is made by the moving of a line and has length and breadth as A & B & C which is the second kind of Magnitude.



The Works
OF
LORD BYRON.

A NEW, REVISED AND ENLARGED EDITION,
WITH ILLUSTRATIONS.

Poetry. Vol. I.

EDITED BY
ERNEST HARTLEY COLERIDGE, M.A.,
FOR. F.R.S.E.

42789
LONDON:
JOHN MURRAY, ALBEMARLE STREET.
NEW YORK: CHARLES SCRIBNER'S SONS.

1903.

CANTO I.

DON JUAN.

113

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.

q

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
