The Venice Time Machine

Presented by: Yotam Gani
How would you define a “Time Machine”? 
Google Maps
Venice nowadays
Google Maps
Street view
Google Maps
What if we had a rewind button?
What if we could go back even further?
If we could create the “Google Maps of the past” and “Facebook of the past” could we say that we can design a “Time Machine”? 
THE “VENICE TIME MACHINE PROJECT”

+ A large international scientific program in the field of Digital-Humanities.
+ Launched by EPFL (Lausanne, Switzerland) and the University (Ca'Foscari) of Venice in 2012.
+ Aims to build a collaborative 4D model of Venice, covering 1000 years of evolution for research and education.
+ Creating an open digital archive of cultural heritage.
+ Tracing circulation of Venetian documents to create the “Big-Data of the past”.
+ Stands out for scale and new technologies.
The “Information Mushroom”
The “Information Mushroom”
The “Information Mushroom”
Why Venice?
Why Venice?

80 KM OF SHELVES
2,5 BN PAGES
1000 YEARS OF HISTORY

“…Venice was is a permanent state of risk management, and recorded everything.”
A UNIQUE (digital) HUMANITIES LAB

+ A historian usually studies one or a few documents.
+ A future historian may be able to study 5000 of those at the same time.
+ From the study of one document to the study of “System of documents”.

“If all this material becomes available and searchable, the way we do research may change.”

John Cogon, Professor of the humanities, Harvard University
How to make the data available?
THE PROBLEM
THE WORKFLOW

Scanning/Digitization

Transcription

Connecting Data
THE WORKFLOW

Scanning/Digitization

Transcription

Connecting Data
SCANNING OR DIGITIZATION

Main challenge:
Large, fragile and heterogenic corpus.
25 MACHINES
50 WORKERS
~ 100 YEARS
~ 240 MILLION$
**Scanning or Digitization**

A Better Solution:

**Tomography (X-Ray)**

Fundamental issue is the chemistry of the inks used for administrative documents.
The Workflow

Scanning/Digitization

Transcription

Connecting Data
Transcription using OCR

Optical Character Recognition
Optical Character Recognition

- The conversion of images of types, handwritten or printed text into machine-encoded text.
- Various types of OCR algorithms (on/off line, handwritten/printed, etc.)
- Our focus: off-line handwritten OCR.
Main stages of OCR system:

- Pre-Processing
- Segmentation
- Feature Extraction
- Classification
- Post-Processing
Main stages of OCR system

Pre-Processing → Segmentation → Feature Extraction → Classification → Post-Processing
Pre-processing

Raw data (image) is passing through preliminary processing steps to make it usable for character analysis.

- Binarization
- Noise reduction
- Normalization
- Skew correction
- Slant removal
Binarization (a.k.a. thresholding) is the conversion of gray-scale image into a binary one.

Two categories:

- Global – pick one threshold for the entire document image.
- Adaptive (local) – use different values for each pixel.
The image resulting from scanning process may contain certain amount of noise.

Main challenge – distinguishing noise from text (i.e. dots).

Two main approaches:

- Filtering (masks)
- Morphological operations
NORMALIZATION

+ Aim to remove the variations of the writing and obtain standardized data.
+ Common methods:
  ✶ Skew normalization and baseline extraction
  ✶ Slant normalization
  ✶ Size normalization
Inaccuracies in the scanning process and/or writing style can make the writing tilted or curved.

Some characters are distinguished according to the relative position with respect to the baseline (9 & g).
One of the measurable factors of different handwriting styles is the slant angle between longest stroke in a word and the vertical direction.

Aims to normalize all characters to a standard form.
Size Normalization

- Used to adjust the character size to a certain standard.
- May be applied for both horizontal and vertical size normalizations.
Main stages of OCR system

- Pre-processing
- Segmentation
- Feature Extraction
- Classification
- Post-processing
MAIN STAGES OF OCR SYSTEM

- Pre-Processing
- Segmentation
- Feature Extraction
- Classification
- Post-Processing
SEGMENTATION

- Image is segmented into its subcomponents.
- The ability to separate various lines in the characters directly affects recognition rate.
- Segmentation of cursive writing as a major challenge.
- Strategies are divided into categories:
  - Explicit segmentation
  - Implicit segmentation
Main stages of OCR system

- Pre-Processing
- Segmentation
- Feature Extraction
- Classification
- Post-Processing
Main stages of OCR system

- Pre-processing
- Segmentation
- Feature Extraction
- Classification
- Post-processing
FEATURE EXTRACTION

+ Capture essential characteristics of symbols.
+ Each character is represented as a feature vector, which becomes its identity.
+ Goal: maximize recognition rate and minimize amount of elements.
+ Extraction methods based on 3 types:
  ※ Statistical
  ※ Geometrical and topological
  ※ Global transformations (FT) and moments
Statistical Features

Representation of a character image by statistical distribution of points takes care of style variations to some extent.

The major statistical features used for character representation are:

- Zoning
- Projections and profiles
- Crossings and distances
ZONING

Goal: obtain the local characteristics.
Character image is divided into NxM zones.
Zones are computed as average gray level.
“Darker” squares indicate higher density of zone.
From each zone features are extracted to form the feature vector.
**Direction Features**

+ Based on the contour (קווי מתאר) of the character.
+ For each zone the contour is followed and a directional histogram is obtained by analyzing adjacent pixels.
PROJECTION

- Characters can be represented by projecting the image onto lines in various directions.
- This representation creates a 1-D signal from a 2-D character image that can be used to represent it.
Profiles

- Measures the distance between the edges of a character and the bounding box.
- The profile describes the external shapes of characters.
Crossings is the number of transitions from background to foreground pixel along vertical and horizontal lines.

Distances is the number of pixels from the first pixel detected to the boundaries.
GEOMETRICAL AND TOPOLOGICAL FEATURES

- Encode knowledge about the structure and components of an object.
- High tolerance to distortions and style variations.
- Predefined structures are searched in a character.
- Extract topological features such as the extreme points, maxima and minima, symmetry, curves etc.
- Extract geometrical features such as ratio between width and height, distance between points etc.
Main stages of OCR system

- Pre-Processing
- Segmentation
- Feature Extraction
- Classification
- Post-Processing
Main Stages of OCR System

- Pre-Processing
- Segmentation
- Feature Extraction
- Classification
- Post-Processing
CLASSIFICATION

- Assign an unknown sample into a predefined class.
- Techniques vary widely according to the features set selected.

Four general approaches:

- Template matching
- Statistical techniques
- Structural techniques
- ANN’s – Artificial Neural Networks
DIRECT MATCHING

Direct matching – the binary input is directly compared to a standard set of store prototypes. Can be done using K-nearest neighbor algorithm.

Deformable templates and elastic matching – the characters are matched by deforming the contour of one to fit the edges of the other.
Statistical matching

- Goal: maximize the probability of the observed pattern given the model of certain class.
- Mostly assume that the distribution of the feature set is Gaussian or in worst case uniform.
- Common approach: HMM – Hidden Markov Model.
Structural matching

- Describe a complex pattern in terms of simpler patterns based on the shape of the object.
- These (simpler) patterns are used for classification.
- Characters are represented as the union of the structural primitives.
Main stages of OCR system

- Pre-processing
- Segmentation
- Feature Extraction
- Classification
- Post-processing
Main stages of OCR system

- Pre-Processing
- Segmentation
- Feature Extraction
- Classification
- Post-Processing
Post Processing

- Grouping of symbols into strings.
- Error detection and correction by the use of context.
- Done by using rules defining syntax of word.
- Usually, the probability of two or more characters appearing together in sequence can be computed and is utilized to detect errors.
- Can also use dictionaries.
Main stages of OCR system:

- Pre-Processing
- Segmentation
- Feature Extraction
- Classification
- Post-Processing
THE FUTURE - HTR

- Handwritten Text Recognition.
- From a single character (classical OCR) to sub-images.
- Recurrent Neural Networks based method.
The Workflow

Scanning/Digitization

Transcription

Connecting Data
The connected data exposes the “big picture”
Main goal: make data searchable
Organize information into giant graphs of interconnected data

A graph is constructed based on the extracted information.
**FUTURE WORK**

- Digitization (Tomography)
- Image processing (Pre-processing Suite)
- Content extraction (Automatic transcription, READ)
- Information modelling (Garzoni Project)
- Building an information system (Document Viewer)
- Content enrichment and network effects (Linked Books Project)
- Valorization and use (GIS, digital experiences)
TO CONCLUDE

- Change is scale
- Change in methodology
- Development in technology
- The future: A Global Time Machine
Future Time Machines

Jerusalem Time Machine

The project aims to digitize and analyse the Big Data of the Past of Jerusalem. This project is part of the European Time Machine Project. Research team includes: Prof. Avraham Faust, Prof. Joshua Shwartz, Prof. Yvonne Freidman and Prof. Moshe Koppel.
SOURCES

- Optical Character Recognition, Line Eikvil December 1993
- Optical Character Recognition Systems for Different Languages with Soft Computing, Springer International Publishing AG 2017
- ROGER LABAHN: HANDWRITTEN TEXT RECOGNITION. KEY CONCEPTS
- VENICE GETS A TIME MACHINE, By Alison Abbott, June 2017
- Venice Time Machine homepage, EFPL https://vtm.epfl.ch/
- Big data in Venice: creating the toolkit for a time machine, Seminar by Michele Petochi
- Frédéric Kaplan: The Venice Time Machine (2017 WORLD.MINDS Annual Symposium)
THANKS FOR LISTENING!

Any questions?