Topics

• What?
• Why?
• Crash course in Latin paleography
• Digital Paleography
  • Vectorization - SIFT, Bag of (visual) words
  • Classification - SVM
• The CLaMMM Competition
• Conclusion
What is Paleography?

The study of ancient and medieval handwriting.
Paleography

Palaios (Old)  Graphein (To Write)
וִירוּשָׁלִָם, يְהוּדָה - עַל, אָמוֹץ - בֶּן יְשַׁעְיָהוּ, חָזָה אֲשֶׁר הַדָּבָר א

אֵהֶבֶם אָשֶר חָזָה, יְשַׁעְיָהוּ בֶן-יאָמוֹץ, עַל-יהוּדָה, ירוּשָׁלָם.
Modern Hebrew

Late Semitic

Middle Semitic

Early Semitic

20th-12th century BC

12th-4th century BC

4th century BC
Motivation

• Dating and localization of Manuscripts

• Pre-OCR/HTR

• Paleography Research/ Descriptive

Paleography
Crash Course in Latin Paleography
Paleography Terminology

• Aspect
  • general appearance of a hand (round/square/left-inclined etc.)

• Scale of script
  • relative size (big/small/middle-sized)
Ductus

- succession of strokes forming a letter
- speed and care of execution of a letter (cursive/non-cursive)
Ligature

\[ AE \rightarrow \AE \]
\[ ae \rightarrow \ae \]
\[ OE \rightarrow \OE \]
\[ oe \rightarrow \oe \]
\[ ff \rightarrow \ff \]
\[ fi \rightarrow \fi \]
\[ ij \rightarrow \iij \]
\[ st \rightarrow \st \]
\[ ft \rightarrow \ft \]
\[ et \rightarrow \et \]
\[ fs \rightarrow \fs \]
\[ ffi \rightarrow \fffi \]
Paleography Terminology
Which of the following letters does NOT have a descender?
Latin Scripts

- Trajan
- Rustic
- Greek Uncial
- Uncial
- Half-Uncial
- Visigothic
- Luxeuil
- Beneventan
- Carolingian
- Insular
- Protogothic
- Textualis quadrata
- Fraktur
- Humanist
- Times
Uncial script

- Majuscule script
- Uncia (1/12)
- “Inch high” letters
- 4th to 8th century.
A (λ)  E (ε)  L (l)

M, N and U (m, n, u)
PRECAITIONEM MEAEM
TRIBUS PERCIP ET ORATIONEM MEAM NON
SOLONISIUS DEUILTUTAO JUDICIO MEE
EXAMINASTI ET NOSTRINVENTAINME
NONQUXUR OSMEAEM PERXHOM
Carolingian Minuscule

• Minuscule
• 8th to 12th century.
• Emperor Charlemagne
• Uniform script for his empire


Lower case letters: a b c d e f g h i j k l m n o p q r s t u v w x y z
Hæc hannibal post cannesim pugnam capit. Ac diem quod se estam exaulitam samnum mouerat. Accus in hynpis et sa stagnolentes secoms traditurum compassus est ab trebis nobiliis inter se depromebaturum compassorium factione samnicae per zнятия romano rum potens post samnum cannesis pugnas volgatique trebi sermoni bı aduentu hannibalis cum compass urbem excessit sent sine costamine tradita urbis poeno praedidiumque acceptae ibi placemnit argi impedimenta relicta exercitu proptauto magno regioni seius urbes aut defeci.
Gothic

• Blackletter/ Old English
• Minuscule
• 12\textsuperscript{th} - 17\textsuperscript{th} century.
• Moveable type.
• slender, angular, and compressed
• “Barbaric”
Typical Gothic capitals. 14th century.

 Capitals from English and German area designs. 15th century.

 Capital letters used by Gutenberg in his 42-line Bible. Mid-15th century.

 Capitals of Schwabacher type, very popular script by the late 15th century and early 16th.


 Fraktur type: capitals used in the "Gebetbuch" (prayer book) of the Emperor Maximilian I. Early 16th century.
Gothic or Caroline?

Caroline

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz

Gothic

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz

Typical Gothic capitals, 14th century.

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz

Capital letters used by Gutenberg in his 42-line Bible, Mid-15th century.

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz

Capitals of Schwabacher type, very popular script by the late 15th century and early 16th.

ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz


ABCDEFGHIJKLMNOPQRSTUVWXYZ
abcdefghijklmnopqrstuvwxyz

Fraktur type: capitals used in the "Gebethbuch" (prayer book) of the Emperor Maximilian I. Early 16th cent.
Digital Paleography
SPI - System for Paleographic Inspections

Dr. Arianna Ciula
Reminder – OCR Process

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

• Image into a vector
  • Global Descriptors - whole image
  • Local Descriptors - patch within an image.
SIFT Descriptor

- Scale-invariant feature transform
- Key points - scaled
SIFT descriptors
- Vector size - number of cells in grid * number of angles

(8 angels) x (4x4 grid) = 128 dimension descriptor
Feature Extraction

• Features encoding
  • letter form
  • spacing
  • curve style
  • Quill
  • slant
  • proportions
  • consistency
  • More..
### Bag of (Visual) Words

<table>
<thead>
<tr>
<th>ID</th>
<th>MONEY</th>
<th>FREE</th>
<th>FOR</th>
<th>GAMBLING</th>
<th>FUN</th>
<th>MACHINE</th>
<th>LEARNING</th>
<th>SPAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>true</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>true</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>true</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>false</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>false</td>
</tr>
</tbody>
</table>
Interest point extraction
Global Descriptors

- **Histogram-type vectorization**
  - Vector dimension as amount of clusters.
  - Normalize the vector - sum to 1.

- **Distance-Based vectorization**
  - Vector dimension as amount of clusters
  - For each cluster - calculate minimum distance to all descriptors from cluster center (prototype).

- **VLAD - vectors of locally aggregated descriptors**
  - Vector dimension as amount of clusters

\[ v_k = \sum_{x_l : \text{NN}(x_l) = \mu_k} (x_l - \mu_k) \]
Classification
Classification

- What is a classifier?
- kNN - k nearest neighbors
- SVM - support machine vectors
- DNN - Deep neural network
SVM

• Support vector machines

• Supervised learning

• Invented by Vladimir N. Vapnik and Alexey Ya. Chervonenkis

• 1963
Which one is the best?
What can we do?
Mapping to Higher Dimension

\[ \varphi((a, b)) = (a, b, a^2 + b^2) \]
Maximum Margin
The SVM Problem

Maximum margin

All samples classified correctly
Pro and Cons of SVM

• **Pros**
  • Accuracy
  • Works well on smaller cleaner datasets

• **Cons**
  • Isn’t suited to larger datasets - long time
  • Less effective on noisier datasets with overlapping classes
The CLaMM Competition

• CLaMM: Classification of Latin Medieval Manuscripts.
• 2016
• Grayscale images.
• 6th - 17th AD
• 12 Labels
Tasks

- “Crisp” classification of 1,000 single-script type manuscripts
- “Fuzzy” classification of 2,000 dual-script type manuscripts
hinc si bi non dat ut videmus. et abh
ab monasteri fide maiestatem et
quis punctum deesse det.

Symhachi...
Classes/ Labels

- Uncial
- Semi-Uncial
- Caroline
- humanistic
- humanistic cursive
- Gothic
  - 7 sub scripts
## “Crisp” Classification

<table>
<thead>
<tr>
<th>System</th>
<th>Accuracy (%)</th>
<th>Ranking according to accuracy</th>
<th>Average Intraclass Distance</th>
<th>Ranking according to AID</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepScript</td>
<td>76.49</td>
<td>4</td>
<td>0.039</td>
<td>3</td>
</tr>
<tr>
<td>FAU</td>
<td>83.90</td>
<td>1</td>
<td>0.068</td>
<td>4</td>
</tr>
<tr>
<td>FRDC-OCR</td>
<td>79.80</td>
<td>3</td>
<td>0.018</td>
<td>1</td>
</tr>
<tr>
<td>NNML</td>
<td>83.80</td>
<td>2</td>
<td>0.026</td>
<td>2</td>
</tr>
<tr>
<td>TAU-1</td>
<td>49.90</td>
<td>7</td>
<td>0.421</td>
<td>7</td>
</tr>
<tr>
<td>TAU-2</td>
<td>50.10</td>
<td>6</td>
<td>0.417</td>
<td>6</td>
</tr>
<tr>
<td>TAU-3</td>
<td>52.80</td>
<td>5</td>
<td>0.393</td>
<td>5</td>
</tr>
</tbody>
</table>
## "Fuzzy" Classification

<table>
<thead>
<tr>
<th>System</th>
<th>Average score (%)</th>
<th>Ranking according to score</th>
<th>Average Intraclass Distance</th>
<th>Ranking according to AID</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>DeepScript</em></td>
<td>2.967</td>
<td>1</td>
<td>0.146</td>
<td>3</td>
</tr>
<tr>
<td><em>FAU</em></td>
<td>2.784</td>
<td>2</td>
<td>0.174</td>
<td>4</td>
</tr>
<tr>
<td><em>FRDC-OCR</em></td>
<td>2.631</td>
<td>4</td>
<td>0.120</td>
<td>1</td>
</tr>
<tr>
<td><em>NNML</em></td>
<td>2.771</td>
<td>3</td>
<td>0.134</td>
<td>2</td>
</tr>
<tr>
<td><em>TAU-1</em></td>
<td>0.615</td>
<td>6</td>
<td>0.260</td>
<td>6</td>
</tr>
<tr>
<td><em>TAU-2</em></td>
<td>0.590</td>
<td>7</td>
<td>0.259</td>
<td>5</td>
</tr>
<tr>
<td><em>TAU-3</em></td>
<td>1.226</td>
<td>5</td>
<td>0.356</td>
<td>7</td>
</tr>
<tr>
<td>True label</td>
<td>True</td>
<td>Uncial</td>
<td>Half-uncial</td>
<td>Caroline</td>
</tr>
<tr>
<td>------------</td>
<td>------</td>
<td>--------</td>
<td>-------------</td>
<td>----------</td>
</tr>
<tr>
<td>Uncial</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Half-uncial</td>
<td>1</td>
<td>86</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Caroline</td>
<td>0</td>
<td>4</td>
<td>82</td>
<td>0</td>
</tr>
<tr>
<td>Humanistic</td>
<td>0</td>
<td>0</td>
<td>79</td>
<td>3</td>
</tr>
<tr>
<td>Hum-Cursive</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>75</td>
</tr>
<tr>
<td>Praegothica</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>South-Textual</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Semitextualis</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Textualis</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Hybrida</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Semihybrida</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Cursiva</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
### Confusion matrix Task 1 (DeepScript)

<table>
<thead>
<tr>
<th>True label</th>
<th>Uncial</th>
<th>Half-uncial</th>
<th>Caroline</th>
<th>Humanistic</th>
<th>Hum-Cursive</th>
<th>Praegothica</th>
<th>South-Textual</th>
<th>Semitextualis</th>
<th>Textualis</th>
<th>Hybrida</th>
<th>Semi-hybrida</th>
<th>Cursiva</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncial</td>
<td>87</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Half-uncial</td>
<td>19</td>
<td>36</td>
<td>32</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caroline</td>
<td>1</td>
<td>0</td>
<td>84</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Humanistic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hum-Cursive</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>72</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Praegothica</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>79</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>South-Textual</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>61</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Semitextualis</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>60</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Textualis</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>10</td>
<td>68</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Hybrida</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>12</td>
<td>0</td>
<td>38</td>
<td>30</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Semi-hybrida</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>12</td>
<td>25</td>
<td>45</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cursiva</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>8</td>
<td>74</td>
<td>0</td>
</tr>
</tbody>
</table>
Research in TAU

- Dead Sea Scrolls
- The Cairo Genizah
- Tibetan Buddhist corpus
Conclusion

• New features extractions

• Writer Identification
Kahoot.it
Questions?
Sources

- Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts By Kestemont, Christlein, and Stutzmann (2017)
- A case study on the System for Paleographic Inspections (SPI): challenges and new developments By Aiolli and Ciula (2005)
- Robust and applicable handwriting biometrics, Axel Brink
- Unsupervised Feature Learning for Writer Identification and Writer Retrieval, ICDAR (2017)
- User Manual for Latin Fonts by Juan-José Marcos