Automatic information extraction from historical collections
The case of the 1808 venetian cadaster

Sofia Ares Oliveira
Digital Humanities Laboratory – Ecole Polytechnique Fédérale de Lausanne, Switzerland

CREATE Seminar, 5 April 2018
Documents
X Savi alle Decime : Filze
X Savi alle Decime: Quaderni dei Trasporti
X Savi alle Decime: Indici dei Quaderni dei Trasporti
X Savi alle Decime : Giornali dei Traslati
X Savi alle Decime: Catastici delle parrocchie
Provveditori alla Sanità, Necrologi
Segretario alle Voci

30 Agosto 1726
30 Agosto 1731
30 Agosto 1736
22 Decembre 1726
11 Agosto 1729
11 Novembre 1729
11 Agosto 1731
10 Agosto 1732
10 LUGLIO 1733
16 Novembre 1733
16 Agosto 1734
16 Agosto 1735
10 Luglio 1731
10 Luglio 1732
10 Luglio 1733
10 Luglio 1734
10 Luglio 1735
Giustizia Vecchia : Accordi dei Garzoni
Fondazione Cini
Photo Library
Sommarioni : censo stabile (napoleonici)
Digitization
Layout Analysis
dhSegment

Generic deep learning pixel-wise segmentation
dhSegment

Generic deep learning pixel-wise segmentation

Assign to each pixel a label
Generic deep learning pixel-wise segmentation

Assign to each pixel a label

Train a neuronal network by showing several examples
Overview of the segmentation system
Architecture network

- Encoder is a pre-trained convolutional neural network (ResNet-v1)
- Each encoding layer has a corresponding decoding layer
  → **U-shaped architecture**

Transcription
Towards automatic transcription

First subset of 18th century fiscal documents from the Venetian State Archives (Condizione di Decima, Quaderni dei Trasporti, Catastici)

23,000 units (image segments) used

55,000 Venetian names of persons and places
A Convolutional Recurrent Neural Network combines the best of CNN and RNN

Machine vs human performance experiment

Training set : 20 712 units
Evaluation set : 2 317 units

36 Italian-speaking transcribers
8 674 valid transcriptions

Character Error Rate (CER)
Transcribers’ average : 13% CER
CRNN : 8% CER
The **automatic transcriber** delivers better transcriptions performances than human average, making **accessible** document content through **textual search**

This opens up new prospects for **massive indexing**, analysis and study of historical documents
The case of 1808 cadaster
Venice
1808
<table>
<thead>
<tr>
<th>Numero della Mappa</th>
<th>Possessori</th>
<th>Denominazione dei Pezzi di terra</th>
</tr>
</thead>
<tbody>
<tr>
<td>9001</td>
<td>Primani</td>
<td>Posta di Spalto</td>
</tr>
<tr>
<td>9002</td>
<td>Marcello Marina e Suedo</td>
<td>25</td>
</tr>
<tr>
<td>9003</td>
<td>Venier del Fortane e R. Ballesta</td>
<td>28</td>
</tr>
<tr>
<td>9004</td>
<td>Suddetto</td>
<td>27</td>
</tr>
<tr>
<td>9005</td>
<td>Vacariello Giuseppe e Ant.</td>
<td>25</td>
</tr>
<tr>
<td>9006</td>
<td>Ruoti Giuseppe e Sta. Maria</td>
<td>22</td>
</tr>
<tr>
<td>9007</td>
<td>Nuuti Andrea e S. Morbizzi Antonell</td>
<td>27</td>
</tr>
<tr>
<td>9008</td>
<td>Delfini Luigi e Delfini</td>
<td>25</td>
</tr>
<tr>
<td>9010</td>
<td>Ruoti Antonio e Delfini</td>
<td>22</td>
</tr>
<tr>
<td>9009</td>
<td>Ruoti Antonio e Delfini</td>
<td>25</td>
</tr>
<tr>
<td>9011</td>
<td>Ruoti Antonio e Delfini</td>
<td>22</td>
</tr>
<tr>
<td>9012</td>
<td>Ruoti Antonio e Delfini</td>
<td>25</td>
</tr>
<tr>
<td>9013</td>
<td>Ruoti Antonio e Delfini</td>
<td>22</td>
</tr>
<tr>
<td>9014</td>
<td>Ruoti Antonio e Delfini</td>
<td>25</td>
</tr>
<tr>
<td>9015</td>
<td>Ruoti Antonio e Delfini</td>
<td>22</td>
</tr>
</tbody>
</table>
Overview of the system

input image → Segmentation (dhSegment) → text → Transcription (CRNN)

georeferenced parcels → contours → text

georeferenced parcels

5599
3457
6681
number transcription
Data annotation
Pixel labelling
Training

Input {image, label} → data augmentation → neural network → dhSegment → 2-class prediction
Parcel extraction and georeferencing

If the image is georeferenced, the geographical coordinates are directly inferred and the parcels can be exported into a GIS system.
Pipeline

input image

Segmentation (dhSegment)

contours

text

Transcription (CRNN)

georeferenced parcels

5599
3457
6681

number transcription
Text extraction

For each parcel, find its corresponding text region

- parcel localization
- text probabilities
- orientation finding
- rotation and cropping
Transcription

Convert image segments into features sequences that will be mapped into text.
Training data

Synthetic data generated from MNIST dataset (100K)

Handwritten numbers extracted from venetian archives (~ 30K)

+ data augmentation
Results
## Parcel extraction results

<table>
<thead>
<tr>
<th>IoU threshold</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parcel recall</td>
<td>0.90 (1062)</td>
<td>0.79 (941)</td>
<td>0.51 (605)</td>
</tr>
<tr>
<td>Parcel precision</td>
<td>0.50</td>
<td>0.44</td>
<td>0.28</td>
</tr>
<tr>
<td>Extracted parcels</td>
<td></td>
<td>2121</td>
<td></td>
</tr>
<tr>
<td>Ground truth</td>
<td></td>
<td>1185</td>
<td></td>
</tr>
</tbody>
</table>

IoU: Intersection over Union

\[
\text{recall} = \frac{\text{true positives}}{\text{total ground truth}} \in [0, 1] \quad \text{precision} = \frac{\text{true positives}}{\text{total retrieved}} \in [0, 1]
\]
## Label extraction and transcription results

<table>
<thead>
<tr>
<th><strong>Label locating</strong></th>
<th>Inter : 0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.86 (633)</td>
</tr>
<tr>
<td>Precision</td>
<td>0.37</td>
</tr>
<tr>
<td>Extracted labels</td>
<td>1693</td>
</tr>
<tr>
<td>Ground truth (labels)</td>
<td>736</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Label transcription</strong></th>
<th>Inter : 0.8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall (correctly transcribed)</td>
<td>0.83 (608)</td>
</tr>
<tr>
<td>Precision</td>
<td>0.36</td>
</tr>
<tr>
<td>Character Error Rate (CER)</td>
<td>0.14</td>
</tr>
<tr>
<td>Ground truth (labels)</td>
<td>736</td>
</tr>
</tbody>
</table>
Label transcription: what are the errors?

Type of errors
- Deletions: 78%
- Insertions: 3%
- Substitution: 19%

Edit distance
- 1 edit: 46%
- 2 edits: 14%
- 3 edits: 19%
- 4 edits: 21%
Next steps

Automatically process the Sommarioni documents to link the parcels to their owners:

- Extract the table layout
- Transcribe the content of the cells
- Link each entry with the corresponding parcel
Resources and contacts

Digital Humanities Laboratory
dhlab.epfl.ch

github.com/

Ʉ dhlab-epfl/dhSegment
Ʉ solivr/tf-crnn
Ʉ dhlab-epfl/cadasters

Sofia Ares Oliveira
sofia.oliveiraarees@epfl.ch

Isabella di Lenardo
Benoit Seguin
Frederic Kaplan