UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

XDOCS: an Application to Index Historical Documents

F. Bolelli, G. Borghi, C. Grana

Università degli Studi di Modena e Reggio Emilia

Outline of the Presentation

- Introduction
- The XDOCS application:
  - Words extraction pipeline
  - Annotation tools
- A new challenging dataset
- Conclusions

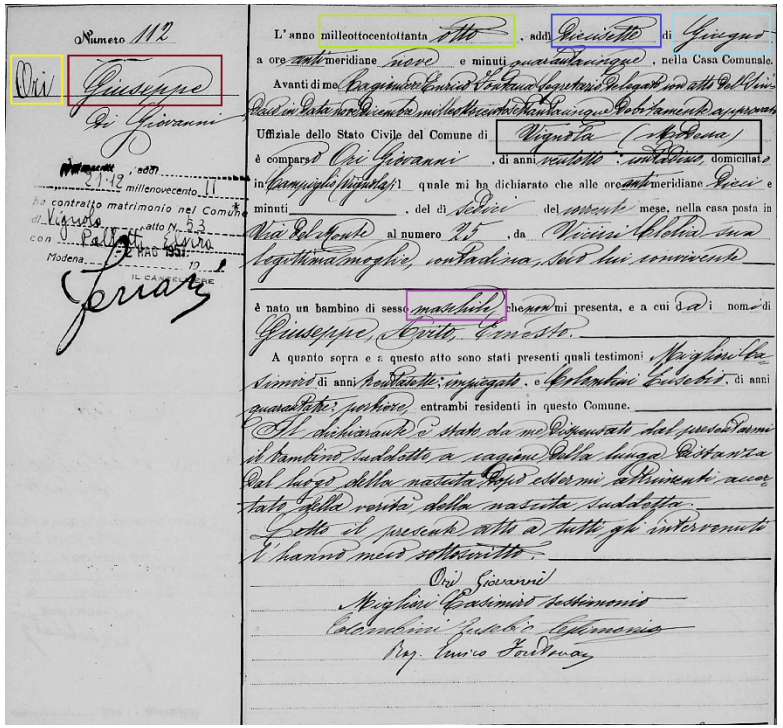
Introduction

- The availability of large collection of handwritten historical manuscripts is often required but their diffusion is limited by:
  - Physical condition
  - Handwriting style
  - Graphic artifacts
- Dematerialization and digitalization represent a possible solution but:
  - Costly and time-consuming
  - *Optical Character Recognizers* (OCRs) often fail

The XDOCS Application

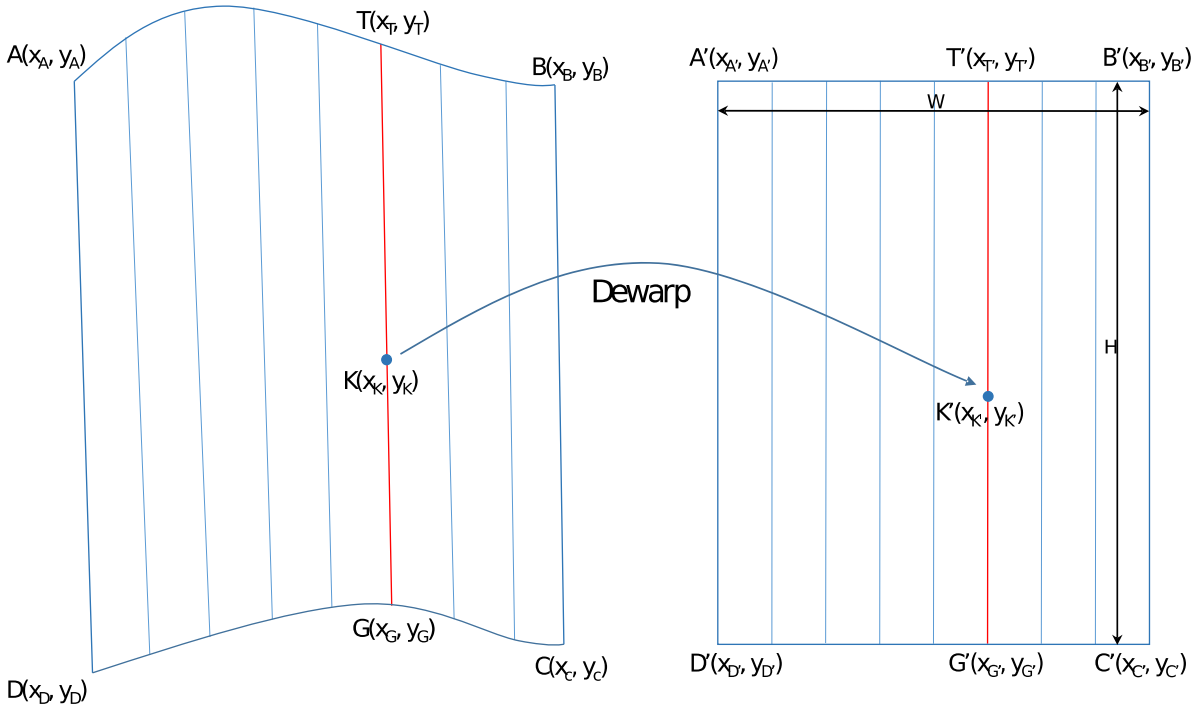
- Goal: develop an innovative data capturing technique able to extract document indexes semi-automatically.

- The application can be splitted into three main blocks:
  - Page Dewarping
  - Word Spotting
  - Annotation tools



AImage Lab

Page Dewarping - Method



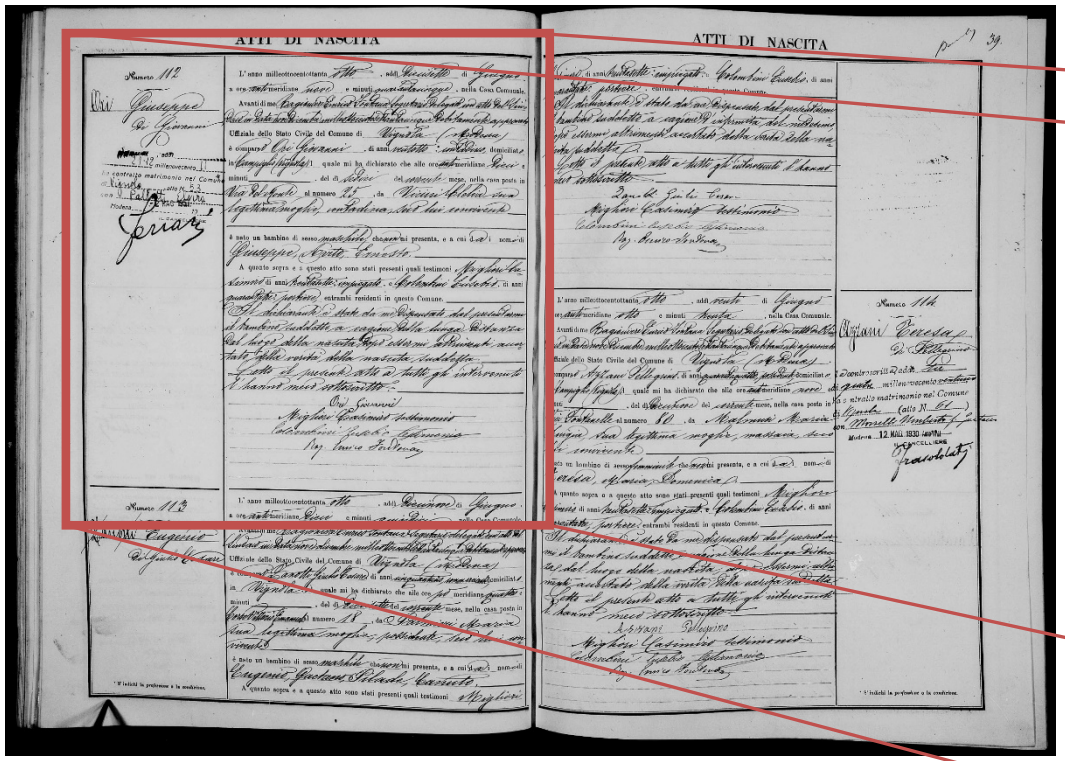
Aims at mapping the projection of the curved surface, represented by four polynomial lines, to a 2D rectangular area with fixed size.

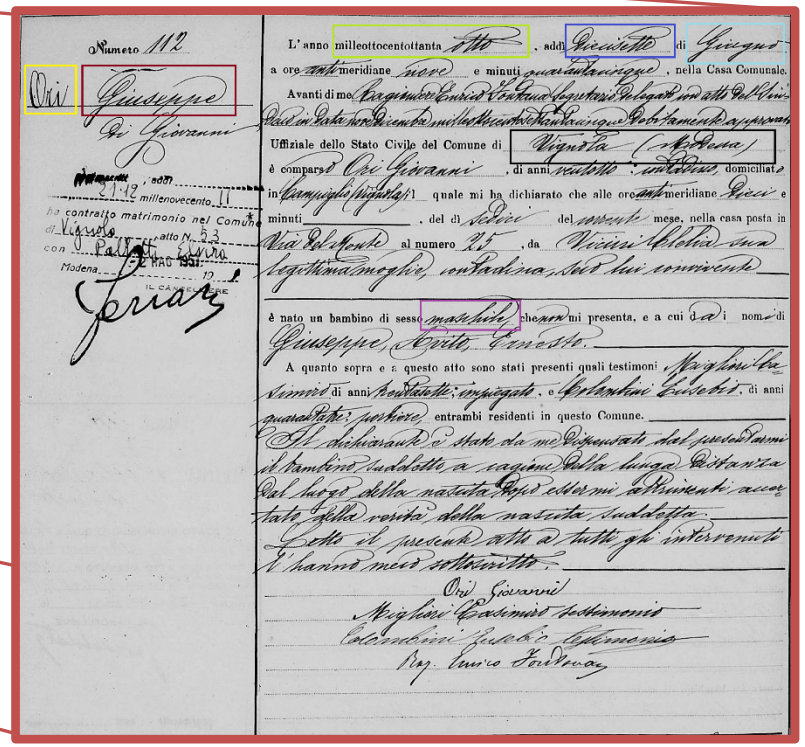
$$\begin{cases} x'_k = x'_A + W * \dfrac{|\widehat{AT}|}{|\widehat{AB}|} \\ y'_k = y'_A + H * \dfrac{|TK|}{|TG|} \end{cases}$$

[1] N. Stamatopoulos, B. Nikolaos: "A two-step dewarping of camera document images." Document Analysis Systems, 2008

AImage Lab

Page Dewarping - Results

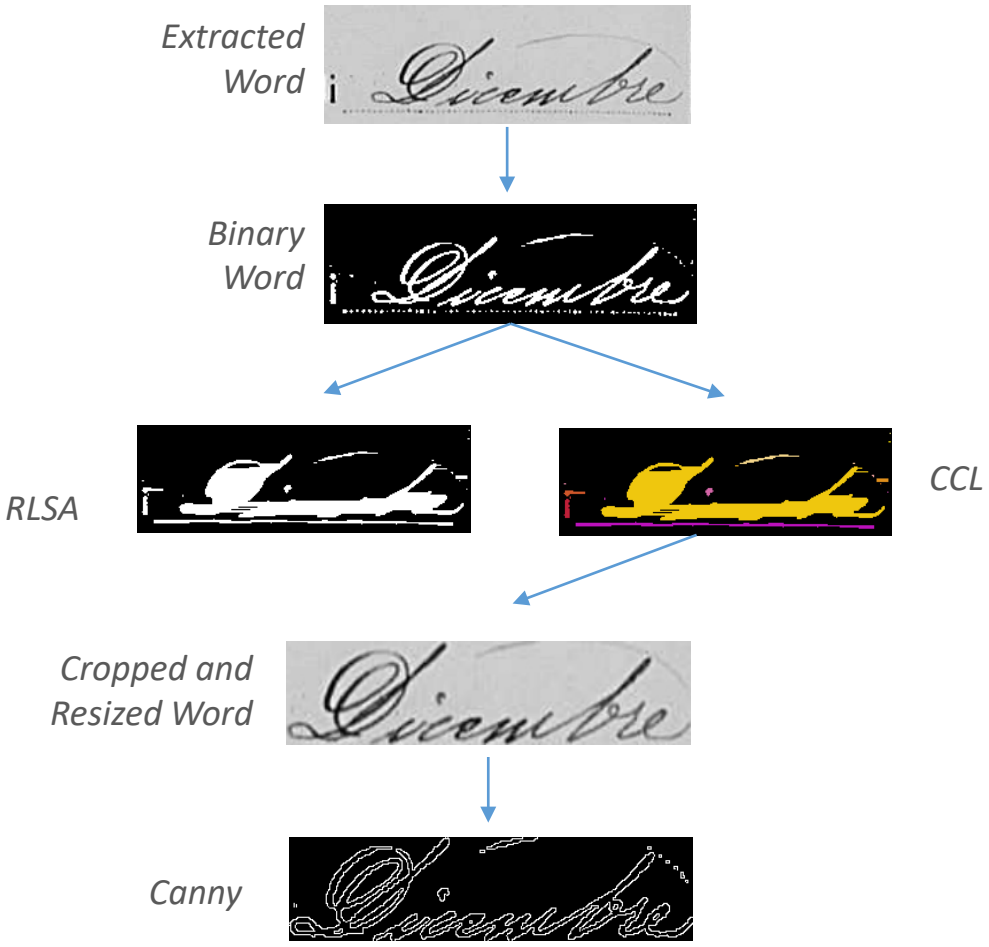




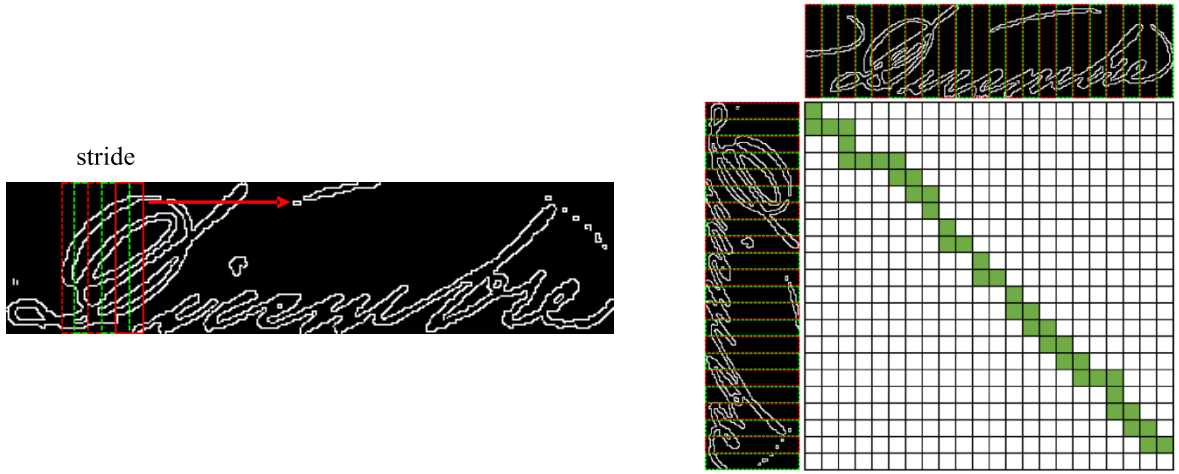
AImage Lab

Word Spotting - Method

① Word Extraction and Preprocessing



② Word Matching Based on HOG Descriptor and Dynamic Time Warping



$$DTW(i,j) = min \begin{cases} DTW(i-1,j) \\ DTW(i,j) \\ DTW(i,j-1) \end{cases} + \sum_{k=1}^{N} |x_{ik} - y_{jk}|$$

AImage Lab

Word Spotting - Results

***Intra* and *Inter* dataset evaluation of word spotting algorithm with:**

- *Mean Averages Precision* (MAP) with *cut-off* at C = {5, 10, 15}:

$$MAP@n = \frac{\sum_{i=1}^{Q} ap@n_i}{N} \qquad ap@n = \frac{\sum_{k=1}^{n} P(k)}{\min(m,n)}$$

- Correct Match First (CMF): percentage of queries with P(1) = 1.

		Vignola	*Carpi*	*Formig.*
Vignola	MAP@05	0.528	0.1	0.181
	MAP@10	0.38	0.086	0.144
	MAP@15	0.306	0.093	0.132
	CMF	75.25%	17.82%	26.73%
Carpi	MAP@05	0.135	0.466	0.095
	MAP@10	0.101	0.434	0.078
	MAP@15	0.079	0.414	0.072
	CMF	14.53%	63.25%	15.38%
Formig.	MAP@05	0.192	0.127	0.644
	MAP@10	0.156	0.114	0.541
	MAP@15	0.135	0.121	0.476
	CMF	24.69%	19.25%	77.82%

Results with 16 pixels stride

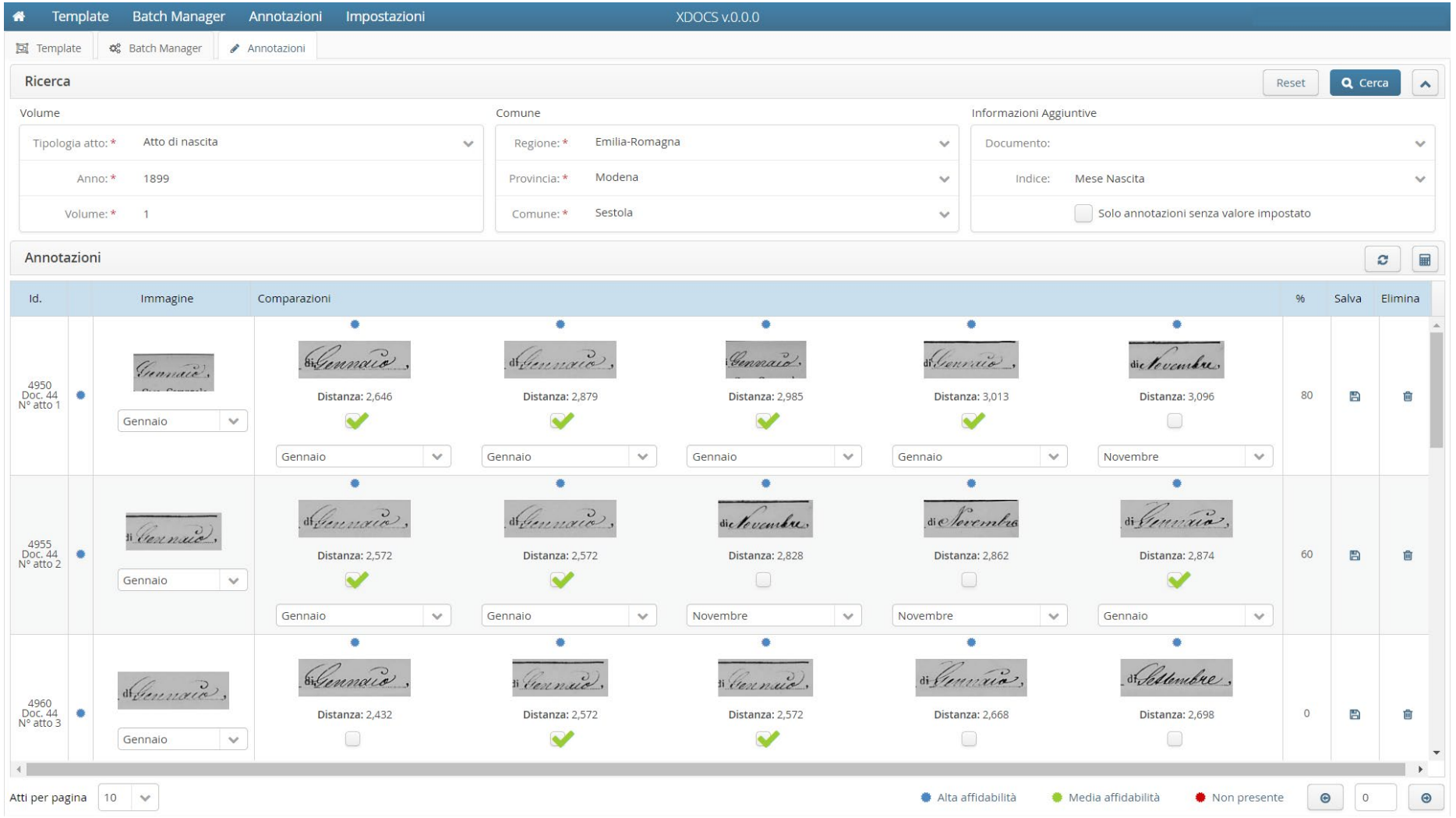
		Vignola	*Carpi*	*Formig.*
Vignola	MAP@05	0.665	0.102	0.222
	MAP@10	0.493	0.093	0.189
	MAP@15	0.4	0.098	0.17
	CMF	87.13%	14.85%	27.22%
Carpi	MAP@05	0.159	0.578	0.125
	MAP@10	0.117	0.536	0.101
	MAP@15	0.091	0.527	0.096
	CMF	19.66%	73.50%	17.95%
Formig.	MAP@05	0.309	0.177	0.823
	MAP@10	0.235	0.152	0.708
	MAP@15	0.194	0.153	0.621
	CMF	40.59%	26.77%	94.14%

Results with 2 pixels stride

		Vignola	*Carpi*	*Formig.*
Vignola	MAP@05	0.468	0.042	0.077
	MAP@10	0.347	0.034	0.057
	MAP@15	0.276	0.028	0.05
	CMF	68.32%	9.90%	13.37%
Carpi	MAP@05	0.086	0.445	0.087
	MAP@10	0.06	0.411	0.067
	MAP@15	0.05	0.382	0.058
	CMF	13.78%	51.70%	15.34%
Formig.	MAP@05	0.097	0.053	0.557
	MAP@10	0.071	0.045	0.413
	MAP@15	0.06	0.042	0.342
	CMF	19.25%	9.21%	80.33%

[3] T. M. Rath and R. Manmatha: *Word Image Matching Using Dynamic Time Warping*. In: Proc. of the Conf. on Computer Vision and Pattern Recognition (CVPR), 2007

Annotation Tool

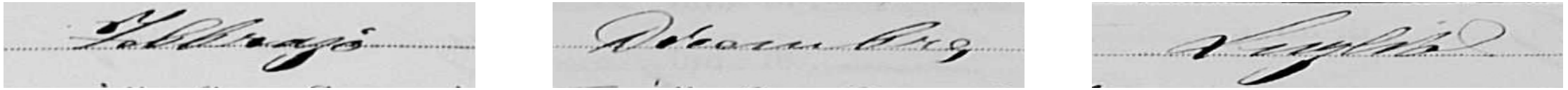


AImage Lab

A New Challenging Dataset

- To test and evaluate the XDOCS indexing pipeline, a huge amount of single word images has been collected and annotated.
- The dataset consists of more than 3000 annotated word images of handwritten names, surnames, birthdays, municipalities, and months and ..
- .. it is publicly available at: aimagelab.ing.unimore.it/XDOCS

Inter dataset variations



Intra dataset variations

AImage Lab

Conclusions

- XDOCS as a tool to encourage the diffusion of handwritten historical documents
- Technical details on which the system is based:
  - Page Dewarping and Word Spotting
- Description of the *Annotation* tools
- Publication of a new dataset

Acknowledgement

- The project is co-funded by the Emilia-Romagna regional administration, SATA s.r.l. and the University of Modena and Reggio Emilia

AImage Lab

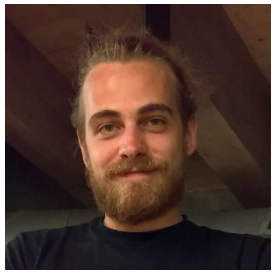
UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

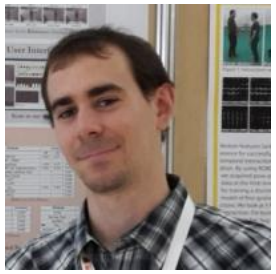
XDOCS: an Application to Index Historical Documents

Thank you!

federico.bolelli@unimore.it



Federico Bolelli



Guido Borghi

Costantino Grana