



Italian National Agency for New Technologies,
Energy and Sustainable Economic Development

Artificial Intelligence Methods for Historical Documents Transcription

Claudio Ronchetti, Serena D'Onofrio and Nicola Quercioli

ENEA TERIN ICT HPC



ECCV Tel Aviv - October 26th, 2022



1101 0110 1100
1101 0110 1101
1101 0110 1100
1101 0110 1101
1101 0110 1100



Outline

- 1. Introduction**
- 2. Project pipeline:**
 - 2.1 Text Detection**
 - 2.2 HTR**
 - 2.3 Table Recognition**
- 3. Work in progress**

Introduction

AI project using an archive of eighteenth-century ledgers, a combination of letters and numbers in a complex format.



Italian National Agency for New Technologies,
Energy and Sustainable Economic Development



Essential impact on the community of the historian: it will allow a fast and deeper analysis of the historical ledgers.

Societal impact: this AI work will allow disclosure and spreading of historical knowledge.



Introduction

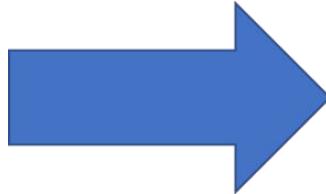
Our dataset consists of thousands of scanned historical documents (XVII and XVIII century). Each document contains handwritten information structured like table.

London		Imperialation of Foreign Goods and Merchandise from Christmas y ^r to Christmas y ^r & Co with an Estimate of the first ^{or} Value		
When Import and for whence	To London Eng Ships Frigates	Summ ^r of the Goods imported	Amount at the value	
Sondon Upper Wm ^r	Bough Great 101 ^l	W ^r 100 ^l	512 ^l	
Africa	Cape Horn 101 ^l 2 ^s	W ^r 100 ^l 2 ^s	120 2 ^s	
From	Mauritius 101 ^l 2 ^s	W ^r 100 ^l 2 ^s	120 2 ^s	
Comm ^r Comodoro 101 ^l 2 ^s	W ^r 100 ^l 2 ^s	120 2 ^s	120 2 ^s	
India	Lahore 101 ^l 2 ^s	W ^r 100 ^l 2 ^s	120 2 ^s	
China	Singap ^r 101 ^l 2 ^s	W ^r 100 ^l 2 ^s	120 2 ^s	
Indostan	Gol ^r 101 ^l 2 ^s	W ^r 100 ^l 2 ^s	120 2 ^s	
Egyptia	Tell 101 ^l 2 ^s	W ^r 100 ^l 2 ^s	120 2 ^s	
Portu ^r Gal ^r 101 ^l 2 ^s	W ^r 100 ^l 2 ^s	120 2 ^s	120 2 ^s	
Gr ^r America ^r 101 ^l 2 ^s	W ^r 100 ^l 2 ^s	120 2 ^s	120 2 ^s	

Introduction

The main goal is to transcribe the scanned ledgers into a digital format while keeping the original table structure.

London		Importation of Foreign Goods and Merchandise from Christmas y ^t to Christmas y ^t & with an Estimate of the first ^{first} or ^{last} Value		
Where Imp'd and for whence	To Merchant Eng. Ships, for Ships	Amount of the first or last value	Amount of the value	
S. America	Augt Great	104 £	101 8 6 £	5 12 £
London (opp. Newgate)	Augt 24		101 8 6 £	1002 2 0
Almonds	Augt 1-1-3-4		101 8 6 £	101 8 6 £
Common Salt	Sept 1-2		101 8 6 £	101 8 6 £
India	Sept 3-8		101 8 6 £	101 8 6 £
Linen	Sept 10-20		101 8 6 £	101 8 6 £
Timber	Sept 24		101 8 6 £	101 8 6 £
Cd. Potas	Sept 22		101 8 6 £	29 8 10
Egyptians	Sept 1-10 0-18		101 8 6 £	101 8 6 £
Southern Africa	Sept 28		101 8 6 £	101 8 6 £
Gn. Almonds	Sept 30		101 8 6 £	101 8 6 £



London Importation of Foreign Goods and Merchandise from Christmas 1767 to Christmas 1768 with an Estimate of the first cost or Value					
Where Imp. id and fro whence	For Merchan. ez	Eng Ships	For Ships	Estimate of the first Cost or Value	Amount of the value
To London From Africa	Bugle Great	154		at 8 to 10	5 15 6
	Copper	1187 2 87		at 3 15 to 40	4602 2 11
	Almond Butter	1454 3 4		at 2 5 to 2 15	3636 19 3
	Cumin Seeds	31 3 10		at 1 8 to 1 17 4	52 3
	Drugs	Arabic	594 3 8	at 38 to 47	1264
		Seneca	3049 0 20	at 38 to 47	6479 10
		Sandrake	667 2 21	at 27 to 29	934 15 3
	Oil Palm	29 1 22		at 18 to 22	29 8 10
	Elephants Theeth	143 0 18		at 4 to 7 10	858 19 2
	Feathers Ostrich	218		at 10 to 14	130 16
	Gro Almonds Seed	70 0 0		at 2 to 2 10	157 10

Project Pipeline

TEXT DETECTION

HTR

TSR

FINAL STEP

Prediction of bounded boxes around the text with Deep Learning techniques trained on digital documents.
From .png to .json file.

Automatic Handwritten Text Recognition via DL method.
From the crops of the words in the image to their digital transcription.

Table Structure Recognition.
Detect the structure of the table starting from the image.
The goal is to understand the positions of the table blocks.

Construction of the digital archive,
merging together the output of the previous steps.

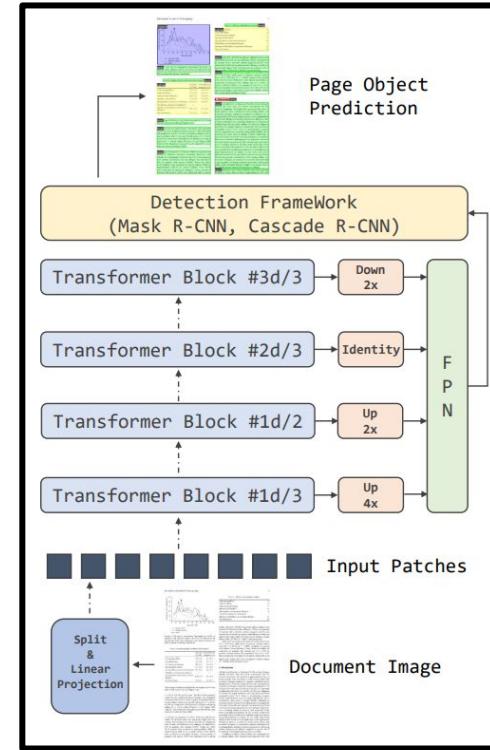
Text Detection

DiT: Self-Supervised Pre-Training for Document Image Transformer

Transformer-based architecture pretrained on FUNSD dataset for Text Detection task.



The online learning has been developed using an external annotation web tool:
Label Studio.



Text Detection

Comparison of the predictions

Pre-trained model



Retrained model

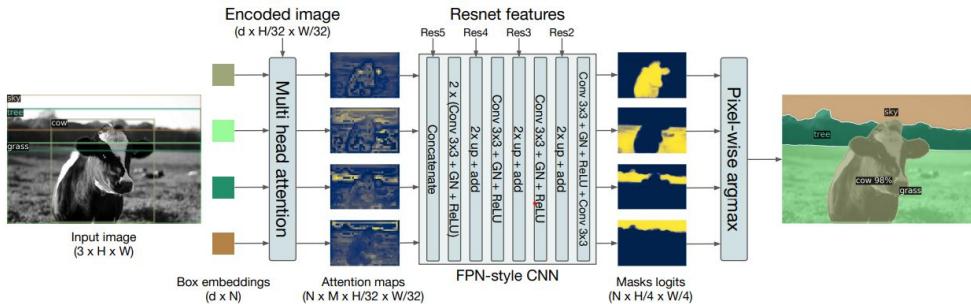


Handwritten Text Recognition

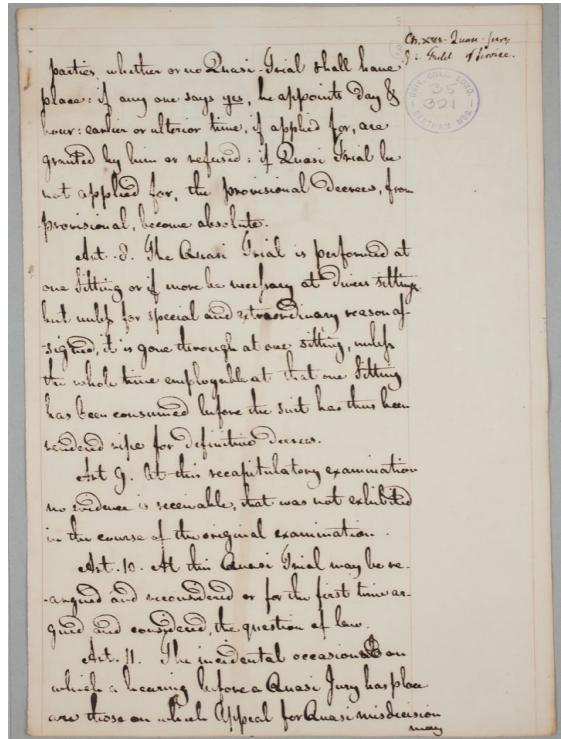
A transformer based Neural Network for handwritten recognition.

It is inspired by a Transformer OCR repository¹ based on End-to-End Object Detection with Transformers.

Network pretrained on Bentham dataset.



1. <https://github.com/him4318/Transformer-ocr>



Handwritten Text Recognition

First experiments with numbers and words. The labelling of the dataset is under construction.

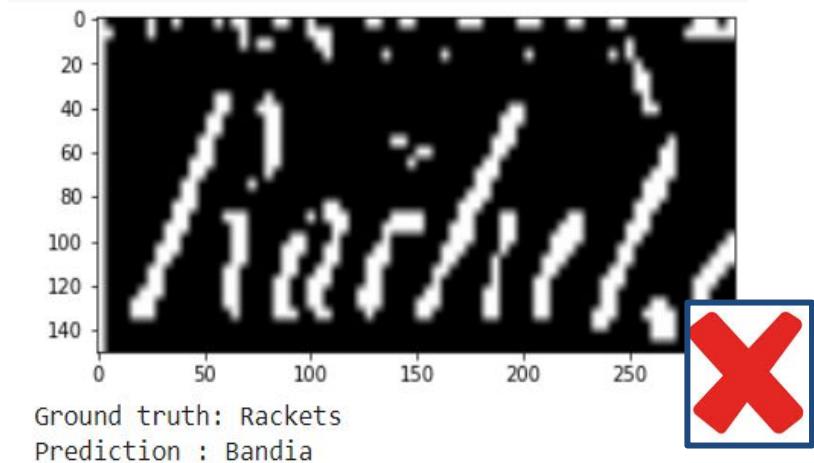
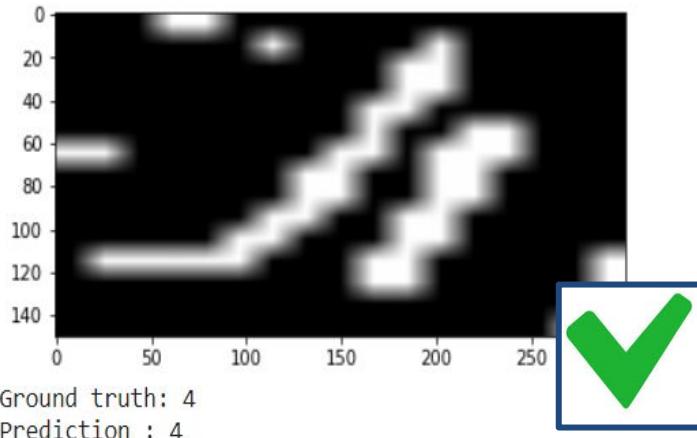


Table Recognition

Variables	Adjusted Prevalence Ratio (APR)	Std. Err.	P-value	95 % Conf. Interval
Oral hygiene status				
Good oral hygiene status	1.00			
Fair oral hygiene status	0.02	0.02	0.18	-0.007 - 0.05
Poor oral hygiene status	0.07	0.03	0.002	0.03 - 0.12
Caries status				
Absence of caries	1.00			
Presence of caries	0.005	0.02	0.77	-0.03 - 0.04
Gender				
Male	1.00			
Female	-0.006	0.01	0.64	-0.03 - 0.02
Socioeconomic status				
High socioeconomic class	1.00			
Middle socioeconomic class	-0.001	0.02	0.95	-0.03 - 0.03
Low socioeconomic class	-0.007	0.02	0.68	-0.04 - 0.03

Column

Row

Variables	Adjusted Prevalence Ratio (APR)	Std. Err.	P-value	95 % Conf. Interval
Oral hygiene status				
Good oral hygiene status	1.00			
Fair oral hygiene status	0.02	0.02	0.14	-0.007 - 0.05
Poor oral hygiene status	0.07	0.03	0.002	0.03 - 0.12
Caries status				
Absence of caries	1.00			
Presence of caries	0.005	0.02	0.77	-0.03 - 0.04
Gender				
Male	1.00			
Female	-0.006	0.01	0.64	-0.03 - 0.02
Socioeconomic status				
High socioeconomic class	1.00			
Middle socioeconomic class	-0.001	0.02	0.95	-0.03 - 0.03
Low socioeconomic class	-0.007	0.02	0.68	-0.04 - 0.03

Spanning Cell

Grid Cell

PubTables-1M and GriTS

Transformer based structures which allows the recognition of the table structure.

Work in progress!

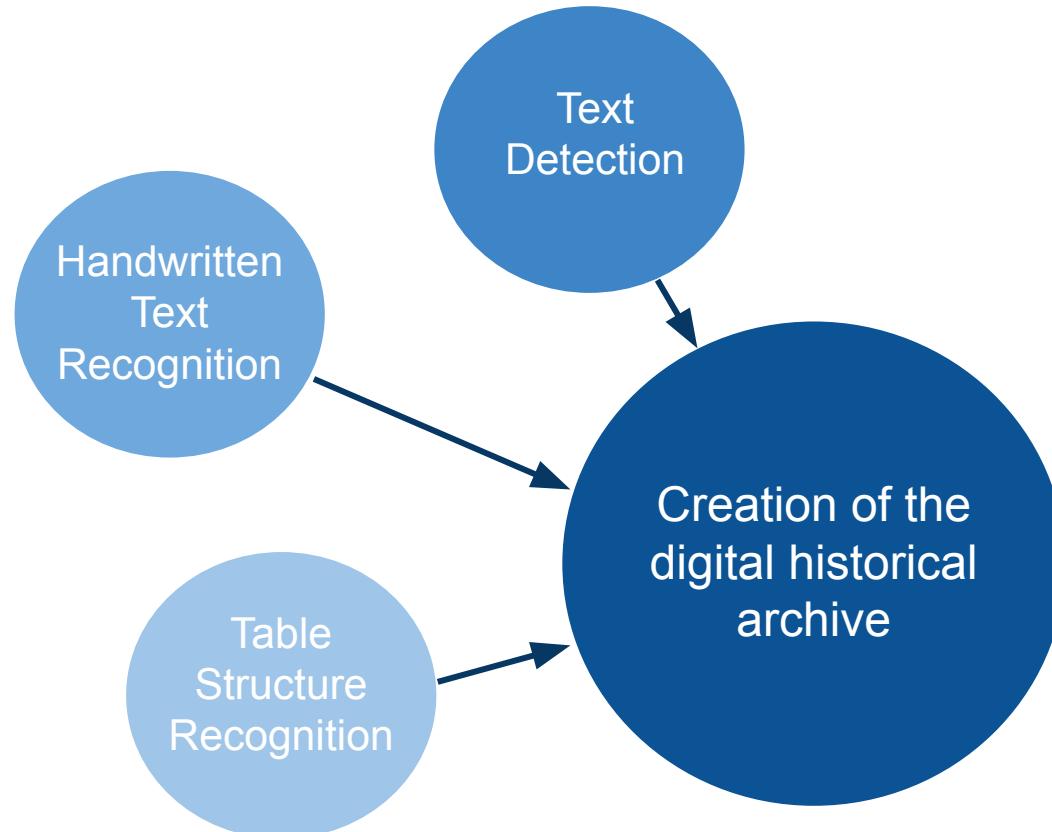
ΔSDM		better	equal	Worse	Sum
ΔSCA					
better	better	19457 (28.9)	12 (0.02)	14654 (21.8)	34,123 (50.8)
better	equal	1158 (1.7)	21989 (32.7)	1024 (1.5)	24,171 (36.0)
better	worse	3755 (5.6)	2 (0.003)	5183 (7.7)	8,940 (13.2)
better	Sum	24370 (36.2)	22003 (32.7)	20861 (31.0)	67,234 (100.0)

Final Step

Our guess is to obtain the digital historical archive with a deterministic way.

Question:

Is it convenient to do it with a ML approach?



Digital platform

LABEL STUDIO



Label Studio

ML BACKEND

PYTHON



STORAGE



CRESO INFRASTRUCTURE

GPU



Work in progress

- ❑ **Text Detection:**
 - ❑ Increase annotated dataset using Label Studio web tool
 - ❑ Improve accuracy of DiT model
- ❑ **Handwritten Text Recognition:**
 - ❑ Transcript words cropped in Label Studio
 - ❑ Retrain DETR-based model with large size dataset
- ❑ **Table Recognition:**
 - ❑ Develop the solution using either AI or no-AI method
- ❑ Integrate all solutions together

References

1. Ledgers: <https://discovery.nationalarchives.gov.uk/details/r/C5580>
2. Li, Junlong, et al. "Dit: Self-supervised pre-training for document image transformer." *arXiv preprint arXiv:2203.02378* (2022).
3. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020, August). End-to-end object detection with transformers. In *European conference on computer vision* (pp. 213-229). Springer, Cham
4. Transformer OCR repository: <https://github.com/him4318/Transformer-ocr>
5. Smock, Brandon, Rohith Pesala, and Robin Abraham. "PubTables-1M: Towards comprehensive table extraction from unstructured documents." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
6. Smock, Brandon, Rohith Pesala, and Robin Abraham. "GriTS: Grid table similarity metric for table structure recognition." *arXiv preprint arXiv:2203.12555* (2022).
7. F. Iannone et al., "CRESCO ENEA HPC clusters: a working example of a multifabric GPFS Spectrum Scale layout," 2019 International Conference on High Performance Computing & Simulation (HPCS), Dublin, Ireland, 2019, pp. 1051-1052, doi:10.1109/HPCS48598.2019.9188135.

Acknowledgements and contacts

This presentation is a part of the collaborative work of our research team. We would like to thank every member of our group. Daniela Alderuccio, Angelo Mariano, Silvio Migliori, Maria Luisa Mongelli, Marco Puccini, Jeremy Land (University of Helsinki), Rodrigo Dominguez (University of Minho), Chris Nierstrasz (Erasmus University Rotterdam).

Furthermore, we deeply thank ENEA and Giovanni Ponti for the support.

Contacts:

- Serena D'Onofrio: serena.donofrio@enea.it
- Nicola Quercioli: nicola.quercioli@enea.it
- Claudio Ronchetti: claudio.ronchetti@enea.it

The End

