ImageCLEF 2016 Handwritten Scanned Document Retrieval Task

Mauricio Villegas, Joan Puigcerver, Alejandro H. Toselli,
Joan-Andreu Sánchez and Enrique Vidal

mauvilsa@prhlt.upv.es

Outline

Introduction

- Currently there is a boom in digitization of documents:
  - To provide digital access, thus a larger audience.
  - Access to fragile historical documents.
  - Ease extraction of information from records.
  - ...
- A large percentage of these documents is handwritten.
- Users expect information access tools to ease searching and processing of these handwritten collections.
- Manually transcribing is generally too expensive for most applications.

Introduction

- Scalable indexing and retrieval techniques for handwritten documents can be based on automatic recognition.
- Not as mature or accurate as printed text recognition.
- To learn models, a similar document is used, or a small part is transcribed.
- The goal is not just recognizing and then using standard text retrieval. Better retrieval performance is attainable by considering the uncertainty of the recognition.

Outline

About the challenge

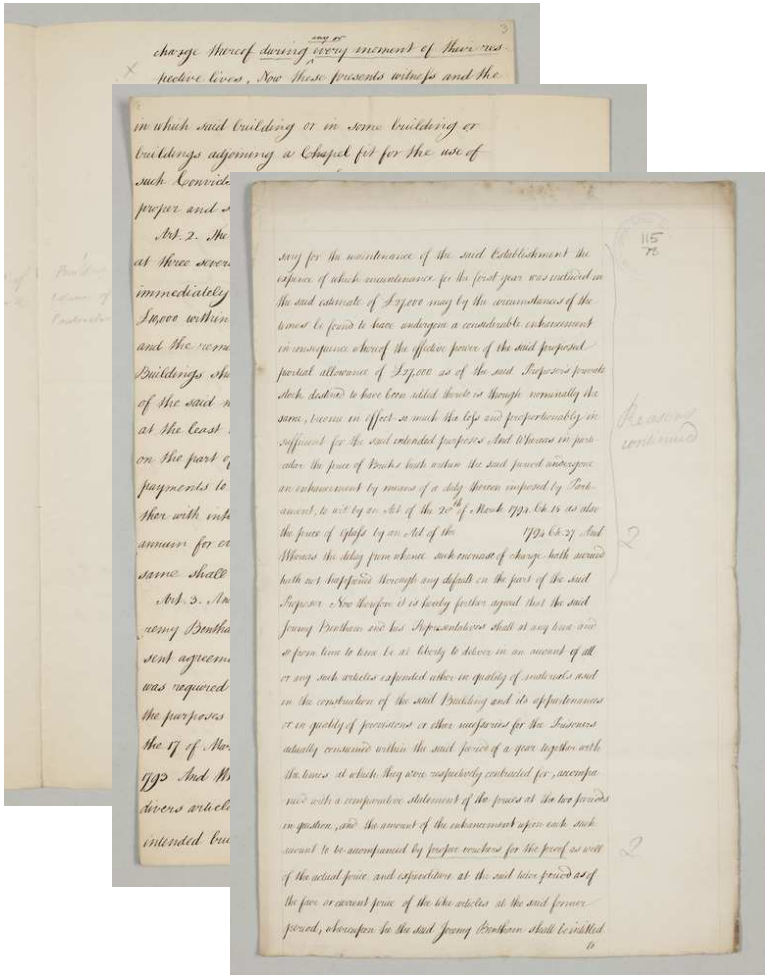
Related evaluations:

- Keyword spotting community (contests at ICFHR and ICDAR 2013–2016).
  - Query-by-example vs. Query-by-string.
  - Training-free vs. Training-based.
  - Segmentation-based vs. Segmentation-free.

Objective for ImageCLEF 2016:

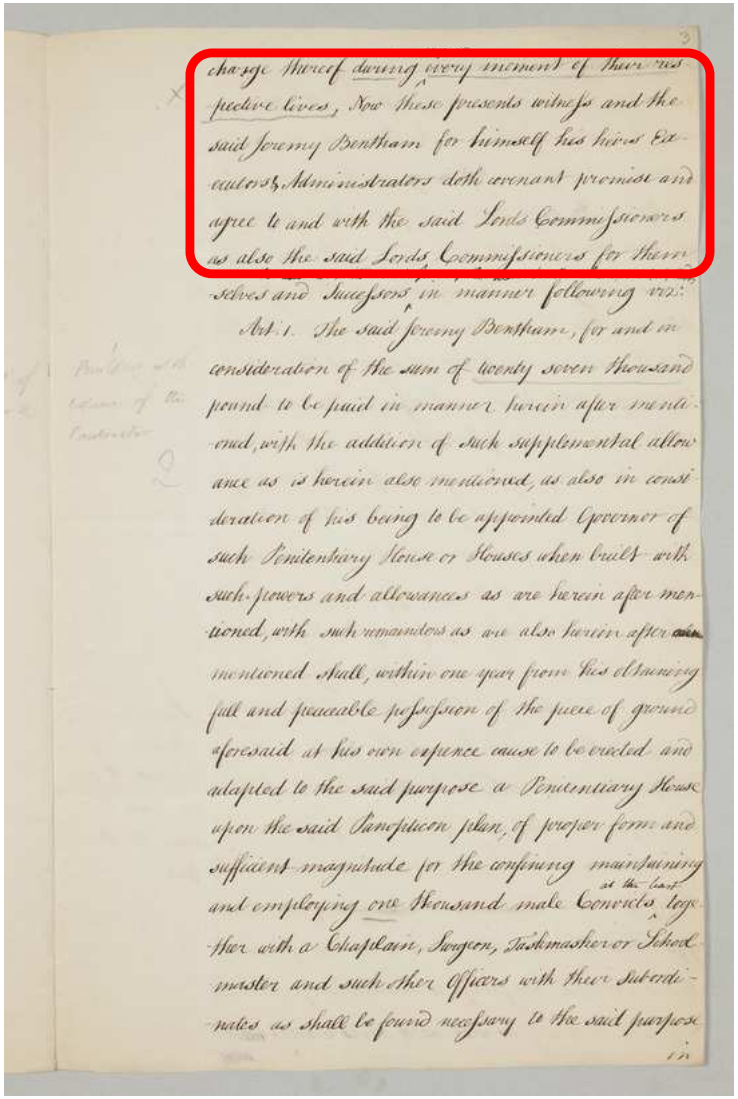
- Evaluate the performance of handwritten retrieval for multi-word query provided as a string.
  - Target a scenario close to a real application.
  - Address details related to the application.
  - Retrieval of local regions within a multi-page document.
  - Allow participation from several communities.

Description of the task



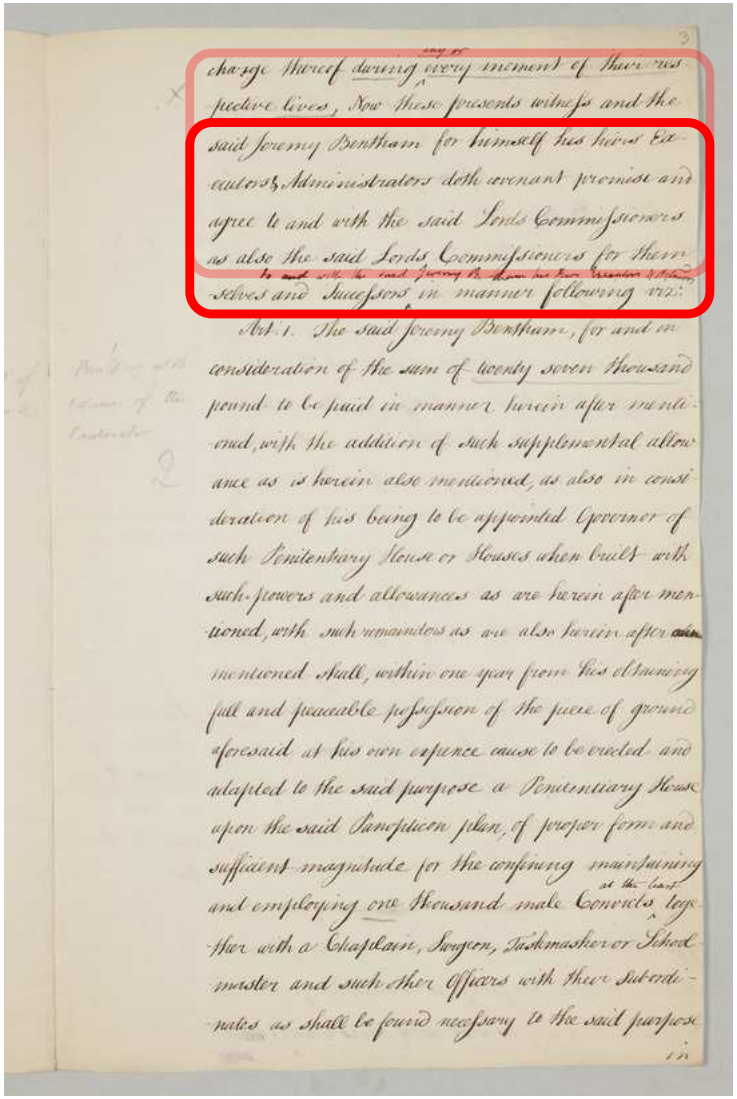
- All images are considered to be one single document.

Description of the task



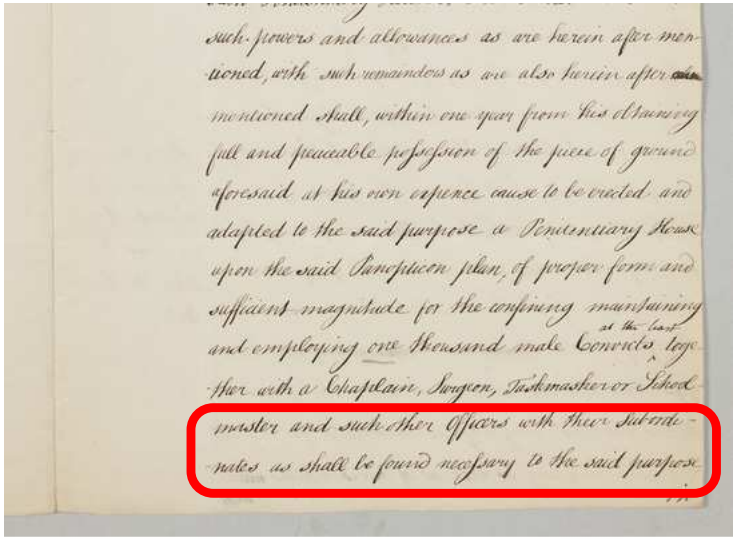
- All images are considered to be one single document.
- The elements to retrieve are small 6-line segments.

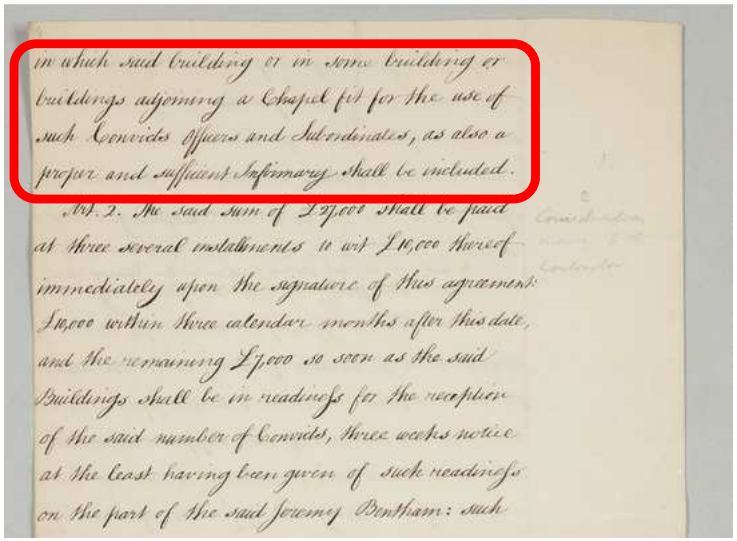
Description of the task



- All images are considered to be one single document.
- The elements to retrieve are small 6-line segments.
- Every line is the start of a segment, thus segments overlap.

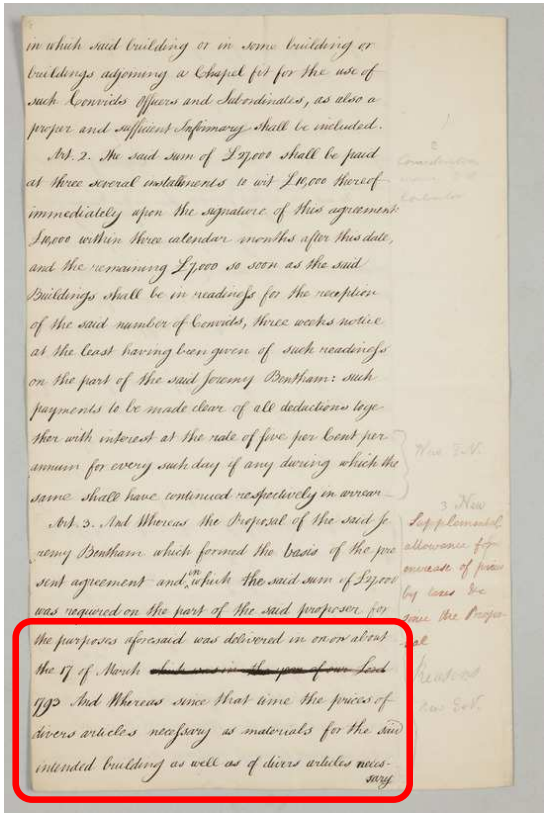
Description of the task





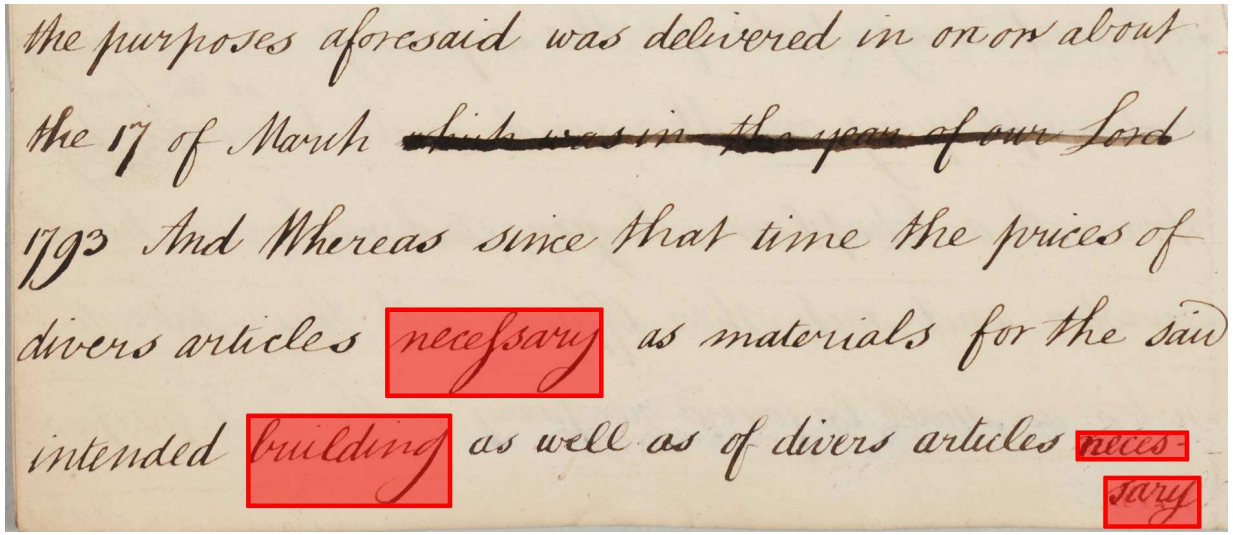
- All images are considered to be one single document.
- The elements to retrieve are small 6-line segments.
- Every line is the start of a segment, thus segments overlap.
- Segments traverse consecutive pages.

Description of the task



- All words in the query must appear in the given order.
- An occurrence of a word may be broken between two lines.
- Retrieval as scores for matched segments + word bounding boxes.

Example relevant segment for *"building necessary"*



Description of the task

Key challenges included:

- Broken words in relevant segments.
- Queries with out-of-vocabulary words.
- Queries with zero relevant results.
- Queries with repeated words.

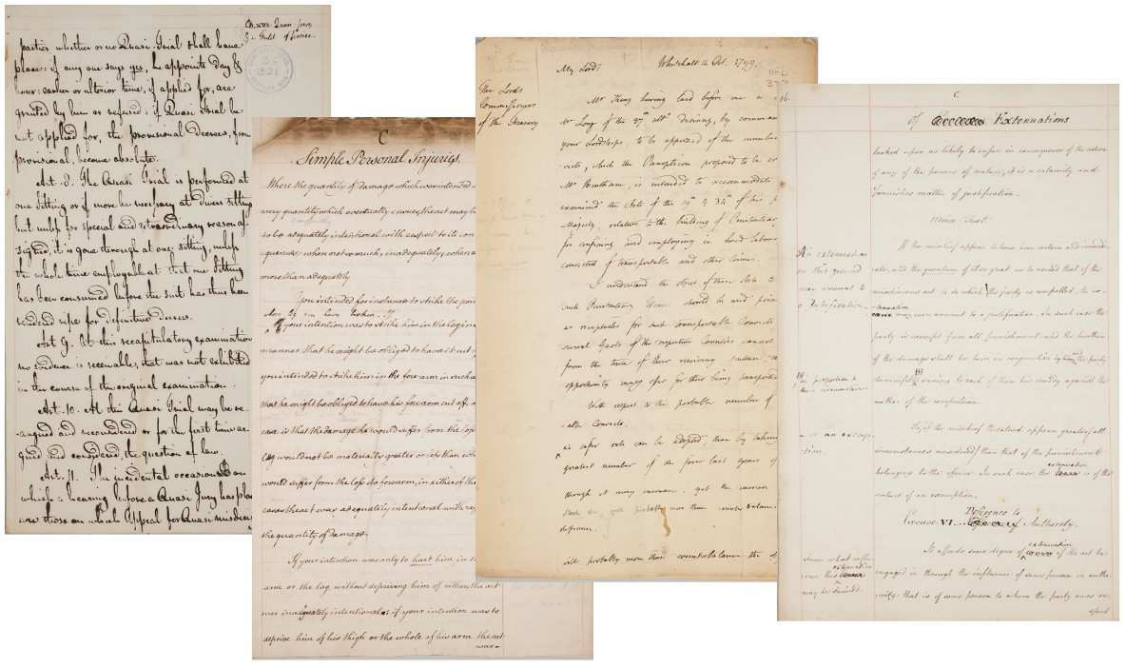
Types of participation:

- Based on training handwritten text recognition models.
- Based on n-best recognition hypothesis (no image processing).
- Query-by-example (examples from automatically segmented training words).

Outline

Dataset used in the task

Dataset of unpublished manuscripts authored by the English philosopher and reformer by Jeremy Bentham.



- 16th century.
- Multiple hands.
- Mostly English.
- Semi-automatic produced ground truth.

Dataset used in the task

	Training	Devel.	Test
Pages	363	433	99
Segments	-	10,584	2,972
Lines	9,645	10,589	3,021
Running words	75,132	91,346	20,686
Total queries	-	510	1,000
Relevant segments	-	10,367	3,493
Rel. segm. for OOV queries	-	1,268	1,083
Rel. segm. with broken words	-	736	1,032

- Provided: pages, extracted lines and n-best recognitions.
- Baseline system: HMM+GMM decoding + 100-best word scoring.

Outline

Participation

48 groups registered, 24 signed the EUA and 4 teams submitted results.

- **CITlab:** University of Rostock, Germany
  - Trained MDRNN with CTC, regular expression based decoder.
  - Handled broken words and unseen in training.
- **MayoBMI:** Mayo Clinic, USA
  - Based on provided 1-best recognition. Stemming and TF-IDF.
  - Worked on broken words but did not submit this part.
- **IIIT:** International Institute of Information Technology, India
  - Query-by-example, handled unseen words, no details given.
- **UAEMex:** Universidad Autónoma del Estado de México
  - Based on provided 1-best recognition. Longest Common Subsequence, handled unseen words.

Results summary: full set segment-based

System*	gAP†		mAP‡	
	Dev.	Test	Dev.	Test
Baseline	74.2	14.4	49.9	8.1
CITlab	95.0	47.1	89.8	39.9
IIIT	41.5	3.4	22.5	3.4
MayoBMI	25.8	2.5	23.4	2.9
UAEMex	61.1	0.3	38.5	0.4

*Best result for each team
†Global Average Precision
‡Mean Average Precision

Results summary: unseen and broken

Only queries with at least one word unseen in training:

System	gAP		mAP	
	Dev.	Test	Dev.	Test
CITlab	89.3	42.6	88.9	39.5
IIIT	13.2	1.7	17.6	2.9
UAEMex	0.2	0.0	0.9	0.2

Required that retrieved segments have a broken word:

System	gAP		mAP	
	Dev.	Test	Dev.	Test
CITlab	59.4	24.3	48.4	23.7

Outline

Conclusions

- Interest in the task was considerable, but few groups submitted.
- One group obtained impressive results, even in the unseen words and broken words challenges.
- Future improvements:
  - The proposed text processing techniques only used 1-best hypothesis, thus limiting the performance.
  - Broken words identified by detecting a hyphenation symbol, which is not always the case.
- The test set ended-up being too difficult.
- The dataset has been released[§] for evaluating with the development set, being the results comparable with this evaluation.

[§] http://dx.doi.org/10.5281/zenodo.52994

Thank you for your attention!

Questions? Comments?

More details can be found in the overview paper:
`http://ceur-ws.org/Vol-1609/16090233.pdf`