ImageCLEF 2016 Handwritten Scanned Document Retrieval Task

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

mauvilsa@prhlt.upv.es













3 Dataset



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- Access to fragile historical documents.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.

Introduction

2 Challenge overview

3 Dataset

Evaluation outcome

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.

change thereof decing very mount of there were pretive loves . Now these presents where for and the

in which word building it in some building er buildings adjenning a behapit for for the use of

Ast. 2. Au at these sever.

the with into account for a second shall state 3 An

110 And H.

and p for excitations of the out of the different of the grant of the second s

All images are considered to be one single document.

have thereof derend every incount of the pective lives , Now these presents withof and the said Journy Bentham for himself has hires Est conterss Administrators dethe arenant premies and agree to and with the said Lends Comme pronores as also the said Lords Commissioners for them schoes and Junefsens in manner following viz. Ast. 1. The said foreing Bontham, for and in consideration of the sum of worky seven thousand pound to be paid in manner furcin after mante end, with the addition of such supplemental allow ance as is horisin also mentioned, as also in conse deration of his being to be appointed Operaner of such Senitentiary Hense or Houses when built with such powers and allowances as are haven after men lioned, with such remainders as are also herein after an mentioned shall, within one your from his obtaining full and praceable performen of the prese of ground aferesaid at fus even expanse onuse to be oracted and adapted to the said purpose a Senicontiany House upon the said Sancpucon plan, of proper form and sufficient magnitude for the confining maintainin and employing one thousand male Convicto tog ther with a Chaplan, Suger, Juckmasher or School master and such other Officers with their Suberde nates as shall be found necessary to the said purpose

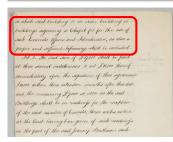
- All images are considered to be one single document.
- The elements to retrieve are small 6-line segments.

change thread densing second of the means particle lives, the how provide interfa and the stationary the thread provide interfa and the assist forming sections. It is the average assist forming section for the mean of presence and agree to and with the static forming provide as also the south section forming theory of the thread and provide in manuary of theory of the attest as a dataford in manuary of theory of the

Art. 1. The said forcing Bontham, for and consideration of the sum of woorty seven thousand pound to be paid in manner haven after mante end with the addition of such supplemental allow ance as is herein also mentioned, as also in const deration of fais being to be appointed Operaner of such Somitentury House or Houses when built with such powers and allowances as are herein after men would, with such remaraders as one also furier after an mentioned shall, within one year from his obtaining full and praceable performen of the piece of grown aferesaid at his own expense ourse to be orieted and adapted to the said purpose a Ponicentiary House upon the said Sancpucon plan of proper form and sufficient magnitude for the confining maintain and employing one thousand male Convicto log ther with a Chaplan, Suger, Jackmasher or School master and such other Officers with their Suberde nates as shall be found necessary to the said purpos

- 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.

such powers and allowances as are barein after men liened with such remainders as one also furier after an mentioned shall, within one year from his obtaining fall and placeable population of the piece of ground aforesaid at his own expense onuse to be created and adapted to the said purpose a Pendentiary House upon the said Sancplicon plan, of proper form and sufficient magnitude for the confinency maintain and employing one thousand make Convicto log Hor with a Chaplain, Songen, Jackmasher or Siked moster and such other Officers with their Suberde nates as shall be found necessary to the wait parper



- 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.

Enclosings adjourning a Chapel for the ade of week Convides officers and Antendenales, as class a proper and refugered Septemony shall be inderted Ast. 2. He said som of Ingers shell be paid at three several motationants is oit Joyers Harref immediately open the sugnations of this agriceness Buildings shall be in readings for the maplies of the wait number of Convictor, three weather notice some shall have unterund respectively in arrive Ast. 3. And Mouras the Surpered of the soid for seat agreement and which the said sum of Sym and required on the part of the said proprior the of Arach that and the you from find 175 And Rhound since that time the process of severs articles newfory is materials for the send indended boulding as well on of dires adults and

- 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"

the purposes aforesaid was delivered in on on about the of March which was in the your four Lord 1793 And Whereas since that time the prices of divers articles necessary as materials for the said intended building as well as of divers articles mees

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).

Introduction

2 Challenge overview

3 Dataset

4 Evaluation outcome

ImageCLEF 2016 Handwritten Scanned Document Retrieval Task

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.

Introduction

2 Challenge overview

3 Dataset

4 Evaluation outcome

Participation

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

- CITIab: 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

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

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

ImageCLEF 2016 Handwritten Scanned Document Retrieval Task

Only queries with at least one word unseen in training:

Svotom	gAP		mAP	
System	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

Introduction

2 Challenge overview

3 Dataset

Evaluation outcome

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.

ImageCLEF 2016 Handwritten Scanned Document Retrieval Task

^{\$}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