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

mauvilsa@prhlt.upv.es
Outline

1. Introduction
2. Challenge overview
3. Dataset
4. Evaluation outcome
5. Conclusions
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  - Access to fragile historical documents.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.
1. Introduction

2. Challenge overview

3. Dataset

4. Evaluation outcome

5. Conclusions
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”
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).
1. Introduction
2. Challenge overview
3. Dataset
4. Evaluation outcome
5. Conclusions
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

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

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

1. Introduction
2. Challenge overview
3. Dataset
4. Evaluation outcome
5. Conclusions
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

<table>
<thead>
<tr>
<th>System</th>
<th>( \text{gAP}^\dagger )</th>
<th>( \text{mAP}^\ddagger )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev.</td>
<td>Test</td>
</tr>
<tr>
<td>Baseline</td>
<td>74.2</td>
<td>14.4</td>
</tr>
<tr>
<td>CITlab</td>
<td>95.0</td>
<td>47.1</td>
</tr>
<tr>
<td>IIIT</td>
<td>41.5</td>
<td>3.4</td>
</tr>
<tr>
<td>MayoBMI</td>
<td>25.8</td>
<td>2.5</td>
</tr>
<tr>
<td>UAEMex</td>
<td>61.1</td>
<td>0.3</td>
</tr>
</tbody>
</table>

*Best result for each team

\( ^\dagger \) Global Average Precision

\( ^\ddagger \) Mean Average Precision
## Results summary: unseen and broken

Only queries with at least one word unseen in training:

<table>
<thead>
<tr>
<th>System</th>
<th>gAP</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev.</td>
<td>Test</td>
</tr>
<tr>
<td>CITlab</td>
<td>89.3</td>
<td>42.6</td>
</tr>
<tr>
<td>IIIT</td>
<td>13.2</td>
<td>1.7</td>
</tr>
<tr>
<td>UAEMex</td>
<td>0.2</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Required that retrieved segments have a broken word:

<table>
<thead>
<tr>
<th>System</th>
<th>gAP</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev.</td>
<td>Test</td>
</tr>
<tr>
<td>CITlab</td>
<td>59.4</td>
<td>24.3</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
2. Challenge overview
3. Dataset
4. Evaluation outcome
5. Conclusions
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: