A Real-World Noisy Unstructured Handwritten Notebook Corpus for Document Image Analysis Research

Jin Chen, Daniel Lopresti, Bart Lamiroy
CSE Department, Lehigh University
{jic207, lopresti}@cse.lehigh.edu, Bart.Lamiroy@loria.fr
Traditionally, document image analysis (DIA) is conducted on datasets that are prepared for research purposes.

- Handwriting Recognition: CEDAR, CENPARMI, ...
- Authorship Analysis: IAM, Firemaker, ...
- ...
Prepared Datasets

*spontaneous* or *elicited*: whether handwritten samples are affected by data collectors.

- **+**: Elicitation simplifies the data collection.
- **-**: Differs from real-world scenarios.

*raw* or *curated*: whether the post-processing of datasets excludes any type of samples, e.g., hard cases.

- **+**: Curation simplifies solutions to the problem.
- **-**: Might overestimate system performance in real life.

*However, there are no absolute spontaneous and curation free datasets.*
The IAM dataset is a large scale handwritten English dataset for handwriting recognition, writer ID, etc.

However, restrictions are applied in data collection:

• Employ pre-printed separating lines.
• Require the use of rulers and an 1.5cm spacing between lines.
• Subjects intervened if the supervisor observes limited space on page.
## Existing Datasets

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Source</th>
<th>Process</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>IAM</td>
<td>Elicited</td>
<td>Raw</td>
<td>HW Recognition, Writer ID</td>
</tr>
<tr>
<td>SUNY</td>
<td>Elicited</td>
<td>Raw</td>
<td>Writer ID</td>
</tr>
<tr>
<td>Firemaker</td>
<td>Elicited</td>
<td>Curated</td>
<td>Writer ID/Verification</td>
</tr>
<tr>
<td>NIST(SD3)</td>
<td>Elicited</td>
<td>Curated</td>
<td>Character Recognition</td>
</tr>
<tr>
<td>RIMES</td>
<td>Elicited</td>
<td>Curated</td>
<td>HW Recognition</td>
</tr>
<tr>
<td>IBN SINA</td>
<td>Spontaneous</td>
<td>Curated</td>
<td>Historical HW Recognition</td>
</tr>
<tr>
<td>CENPARMI</td>
<td>Elicited</td>
<td>Raw</td>
<td>U.S. Zip Code Recognition</td>
</tr>
<tr>
<td>CEDAR</td>
<td>Spontaneous</td>
<td>Curated</td>
<td>U.S. Zip Code Recognition</td>
</tr>
<tr>
<td>Mormon Diary</td>
<td>Spontaneous</td>
<td>Raw</td>
<td>Historical Document Analysis</td>
</tr>
<tr>
<td>Germana</td>
<td>Spontaneous</td>
<td>Raw</td>
<td>Historical Document Analysis</td>
</tr>
<tr>
<td>LU Notebook</td>
<td>Spontaneous</td>
<td>Raw</td>
<td>Various Document Analysis</td>
</tr>
</tbody>
</table>
Motivation

- Most datasets are either elicited or curated.
- Germana and Mormon Diary datasets are historical handwriting datasets that are divergent from modern handwritten datasets.

We want to reduce as much as possible the elicitation and the curation during the process of building datasets.
Problem Space vs. Dataset Space

DIA in Single Pages
- HW Digit Recognition
  - CENPARMI CEDAR ...
- Writer ID/Verification
  - IAM Firemaker ...
- Historical Document Analysis
  - Germana Mormon Diary ...

DIA across Multiple Pages
- Page Order Analysis
- Structure Restoration
- Topic Tracking
- Page Referencing

No suitable datasets available now
Lehigh Notebook Dataset

- All the notebooks were used by Lehigh students, thus ensuring minimum elicited handwriting.

- To scan notebooks, we separated pages while ensuring the page order.

- Each notebook page was scanned at 600dpi into PDF files, using a bitonal setting under plain text mode.

- All pages were converted into TIFF images with no compression, resulting in $5104w \times 6600h$.

- So far, we have collected 499 pages from nine students and aim for 100 notebooks, 3k pages from >50 students.
Lehigh Notebook Dataset

Differences from existing handwriting datasets: corrections, annotations, arrows, doodles, etc.

Minimum elicited and curated handwriting!
Page Order Analysis

• **Page order**: logical sequence of pages that ought to be interpreted sequentially.

• In real life, page order is important for understanding an unstructured document collection, e.g., a set of loose pages.
Structure Restoration

Figure 5: Some example notebook pages from the Lehigh notebook dataset.

(a) The recto side of a double-sided document.
(b) The verso side of a double-sided document.
(c) One page of a computer science course.
(d) Reading order indicated by hand-drawn arrows
(e) An example page containing handwritten annotations at the bottom.
(f) Page 30 in Notebook #2009.
(g) Page 31 in Notebook #2009.
(h) One page containing a handwritten table.
(i) One page containing a handwritten table.

Pattern Recognition Research Lab

\[ \text{Figure 5: Some example notebook pages from the Lehigh notebook dataset.} \]

(a) The recto side of a double-sided document.
(b) The verso side of a double-sided document.
(c) One page of a computer science course.
(d) Reading order indicated by hand-drawn arrows
(e) An example page containing handwritten annotations at the bottom.
(f) Page 30 in Notebook #2009.
(g) Page 31 in Notebook #2009.
(h) One page containing a handwritten table.
(i) One page containing a handwritten table.
• Structure restoration decides which pages belong to separate physical/logical units, e.g., notebooks or topics.

• In real life, it is important for machines to employ customized techniques, e.g., style-based OCR/HWR.

• It is natural to use Lehigh Notebook dataset for such tasks. We have provided notebook IDs, pre-printed ruling line specifications, etc.
The DAE Platform

http://dae.cse.lehigh.edu

This slide is from the DRR 2011 talk by Lopresti and Lamiroy.
This slide is from the DRR 2011 talk by Lopresti and Lamiroy.
Our Lehigh notebook dataset provides a basis for document analysis. One example is the whole book that contains. For machines, such analysis can also benefit style contents, or the temporal order of events those documents analysis might facilitate their understanding. For human beings, such knowledge if there are sets of pages that are coherent in terms of order indicator, notebooks but happened to contain consistent styles of page der. For example, these two pages can be from different creation dates are su. It is also important to bear in mind that neither page numbers show two consecutive pages within the same notebook. It is it can be used to test algorithms for cases where only loose contains notebooks which have obvious physical page order, logical page order. Thus, although the notebook dataset pages within a notebook is not necessarily the same as the a real-world scenario where an unstructured document collection. To this is important for a set of loose pages, which reflects discussion, we define Authorship Notebook ID Subject Tags Ruling Line Specifications, etc.

XML Markup

```xml
<metadata>
  <page_image>
    <id>page294</id>
    <path>/TIFF/lehigh1003_nb2009_page294.tiff</path>
    <hdpi>600</hdpi>
    <vdpi>600</vdpi>
    <page_element>
      <value_list>
        <value_list_item>
          <id>author</id>
          <value>subject1003</value>
        </value_list_item>
        <value_list_item>
          <id>notebookID</id>
          <value>nb2009</value>
        </value_list_item>
        <value_list_item>
          <id>creation_date</id>
          <value>2010/09/15</value>
        </value_list_item>
        <value_list_item>
          <id>Subject</id>
          <value>mathematics, statistics, linear models</value>
        </value_list_item>
        ...
      </value_list>
    </page_element>
  </page_image>
</metadata>
```
Conclusions

• We are motivated by the fact that most existing handwriting datasets are either elicited or curated.

• We aim for collecting 100 notebooks, 3k pages from 100 students. So far we have collected 18 notebooks from nine college students, in a total of 499 pages.

• Currently, the bitonal version if available via: http://dae.cse.lehigh.edu/DAE/. The full-color version will be uploaded soon.

• We also call for discussions on its usage.