

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



*Pattern Recognition
Research Lab*



LEHIGH
UNIVERSITY

Computer Science and Engineering

Computer Science and Engineering

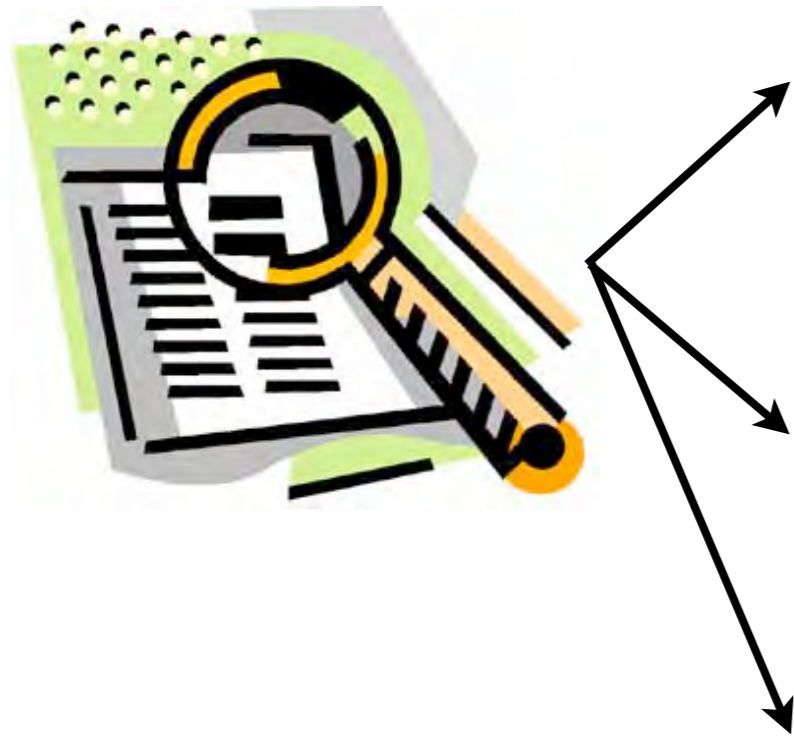
Computer Science and Engineering

Computer Science and Engineering

CSE

Introduction

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.

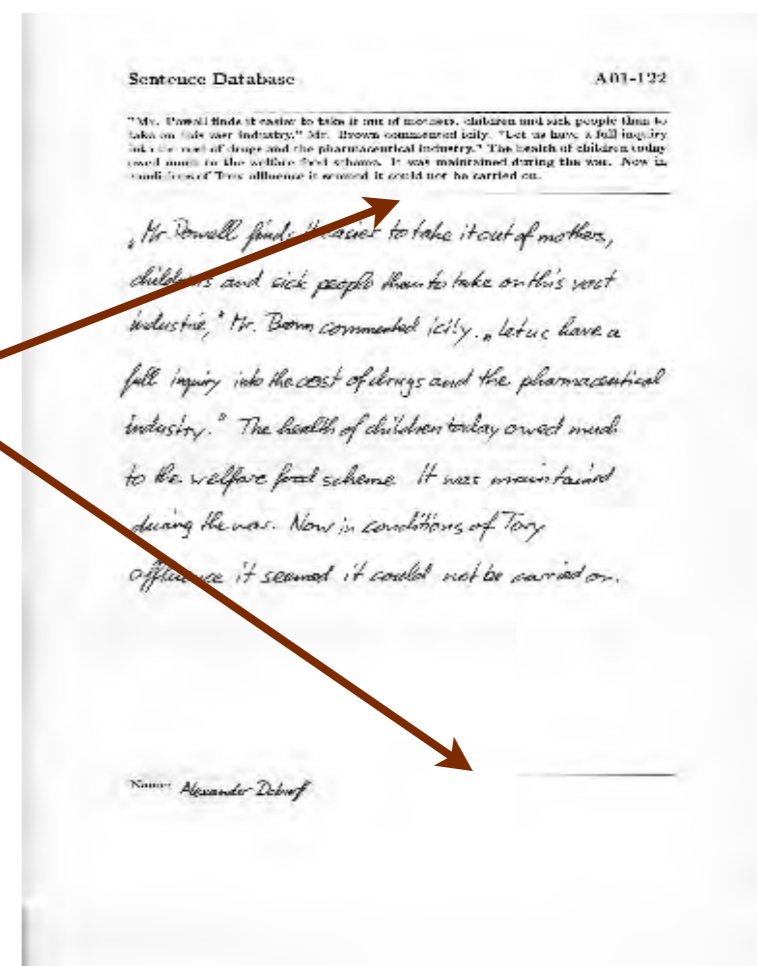
However, there are no absolute spontaneous and curation free datasets.

An Example

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.

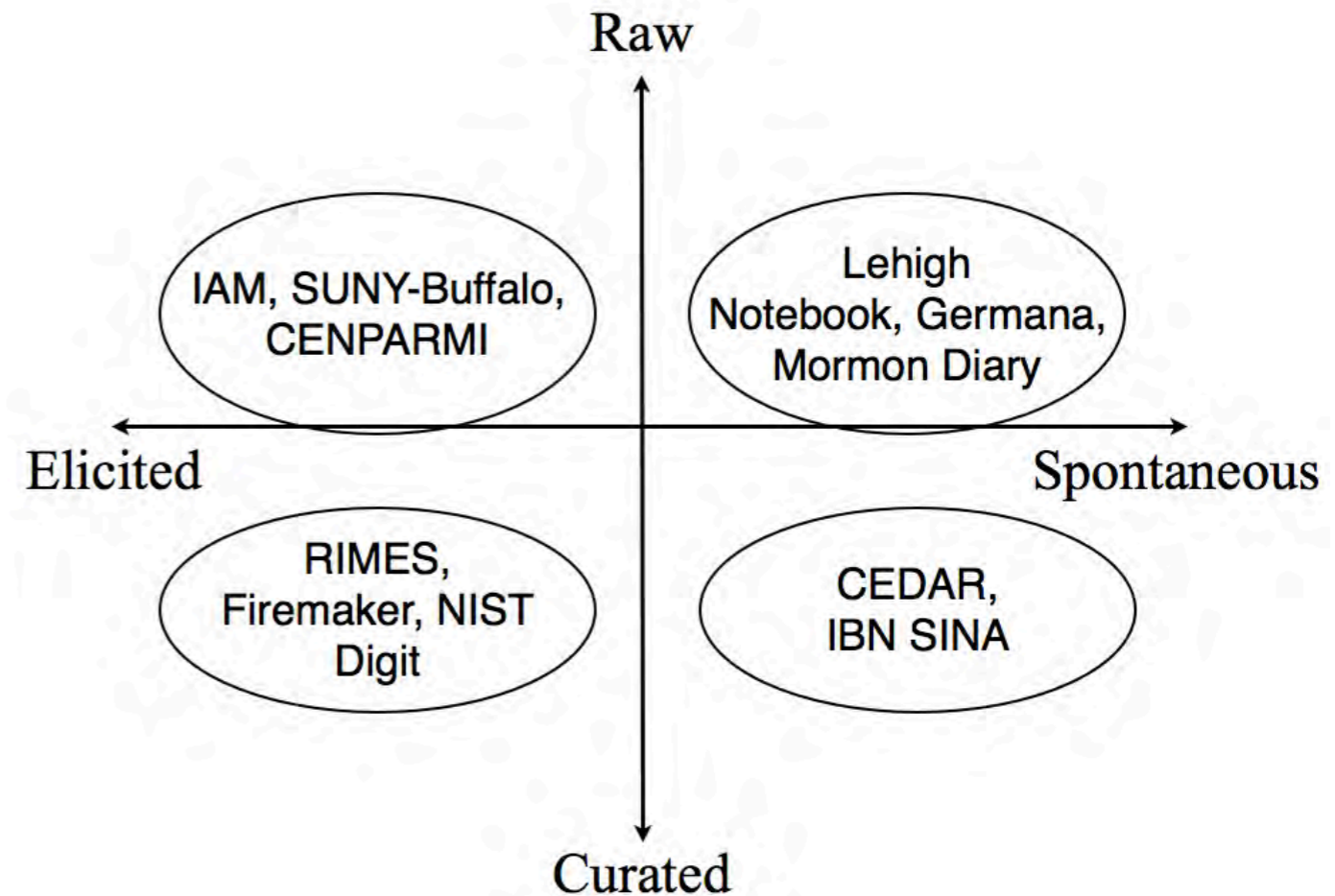


Existing Datasets

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Mormon Diary	Spontaneous	Raw	Historical Document Analysis
Germana	Spontaneous	Raw	Historical Document Analysis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

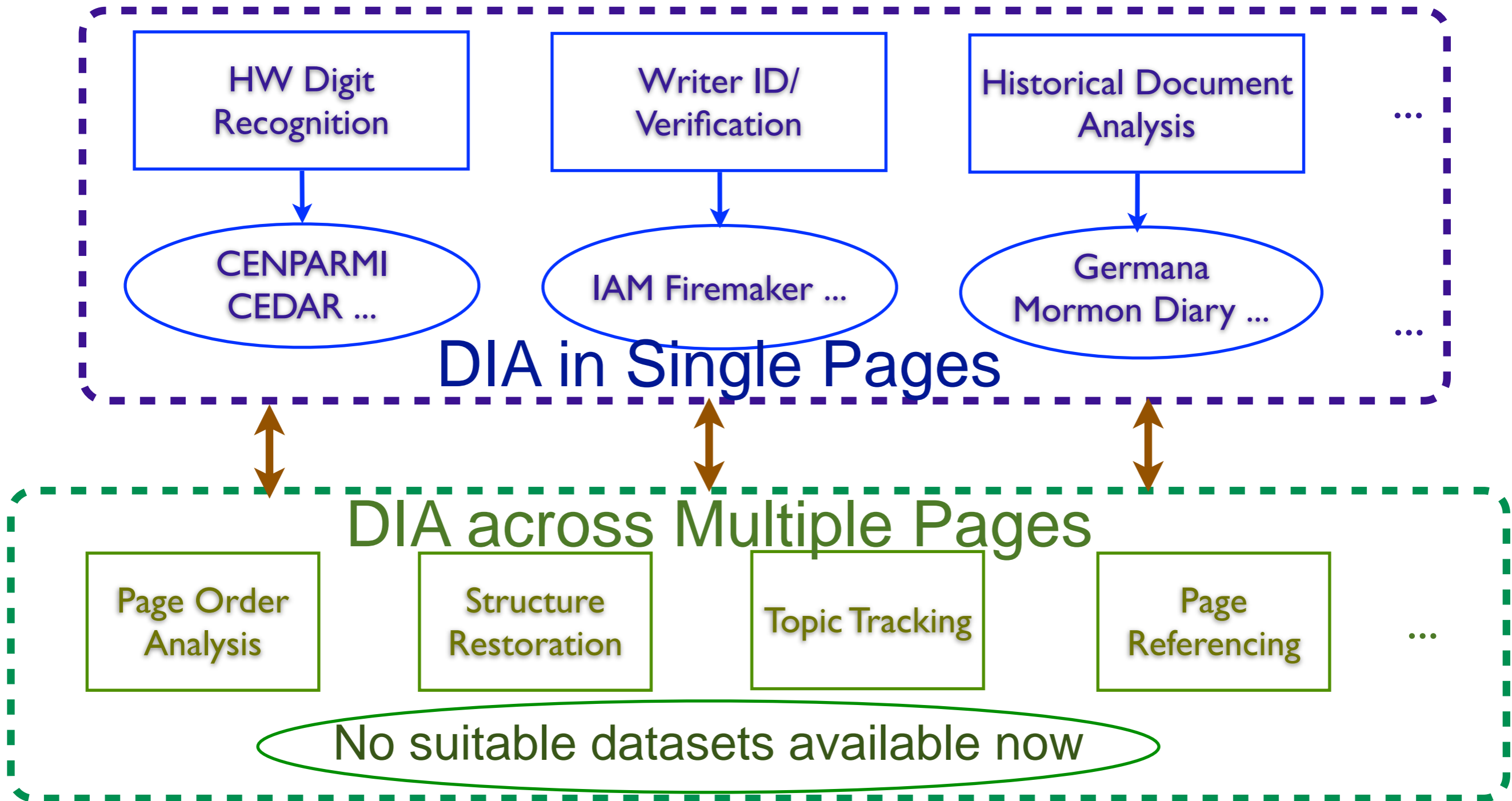
Motivation

- 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



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.

Ex: Suddenly ~~accelerated~~ **accelerated**

N-S equ. simplifies to

$$\frac{\partial^2 u}{\partial t^2} = \frac{\partial^2 u}{\partial y^2}$$

IC: $u(y) = 0$ @ $t=0$. $u(y=0, t) = V$ for $t > 0$
 $u(y \rightarrow \infty, t) = 0$ for $t > 0$

Solution: simulating behavior \rightarrow introduce similarity variables.

reduce from P.D.E \rightarrow O.D.E.

assume form of $u/V = f(\eta)$ velocity profile of some shape. $\eta = Ay^m t^n$

Substitute & determine values of η, m & A (const.) which will reduce PDE to ODE.

$$\frac{\partial f(\eta)}{\partial \eta} = \frac{df}{d\eta} \frac{\partial \eta}{\partial t} = Ay^m m t^{n-1} f'$$

In this case, $m=1, n=-\frac{1}{2}$ & $A = \frac{1}{2\sqrt{t}}$. $\rightarrow \eta = \frac{y}{\sqrt{t}}$ $\rightarrow f'' + \eta f' = 0$

B.C: $u(y=0, t) = V \rightarrow f(\eta=0) = 1$ IC: $u(y, t=0) = 0 \rightarrow f(\eta \rightarrow \infty) = 0$
 $u(y \rightarrow \infty, t) = 0 \rightarrow f(\eta \rightarrow \infty) = 0$

Solve $f'' + \eta f' = 0$ ~~from $f' = C_1 \exp(-\eta^2/2) + C_2$~~

IC: ~~$f(\eta=0) = 1$~~ $u(y, t=0) = 0$

$f(\eta \rightarrow \infty) = 0$

Physics of Semiconductor Devices
 ASSIGNMENT 9

$$\left[\left(\frac{h}{\lambda} - \frac{h}{\lambda'} \cos \theta \right)^2 + \left(\frac{h}{\lambda'} \sin \theta \right)^2 \right] c^2 + m_0^2 c^4$$

\rightarrow wrong!

$\frac{h}{m_0 c} (1 - \cos \theta)$ - I don't think you can derive the right conclusion based on the wrong equations.

$$\left(\frac{hc}{\lambda} - \frac{hc}{\lambda'} \right)^2 = \left[\left(\frac{h}{\lambda} - \frac{h}{\lambda'} \cos \theta \right)^2 + \left(\frac{h}{\lambda'} \sin \theta \right)^2 \right] c^2 + m_0^2 c^4$$

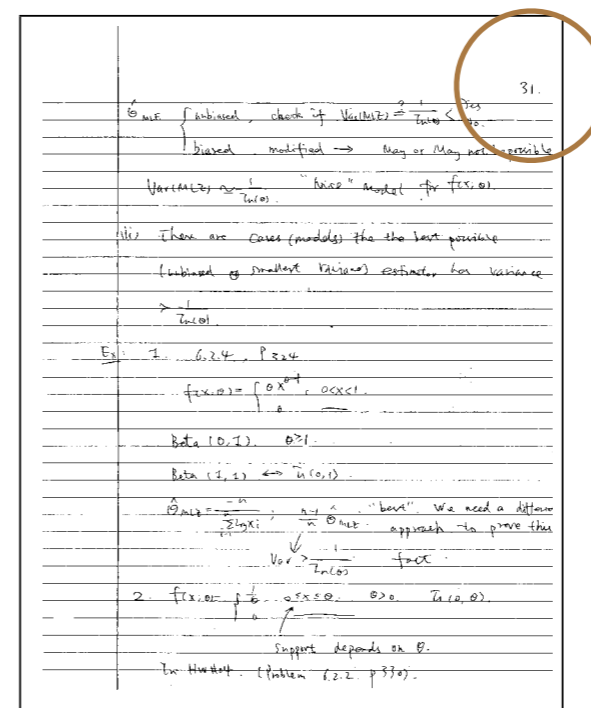
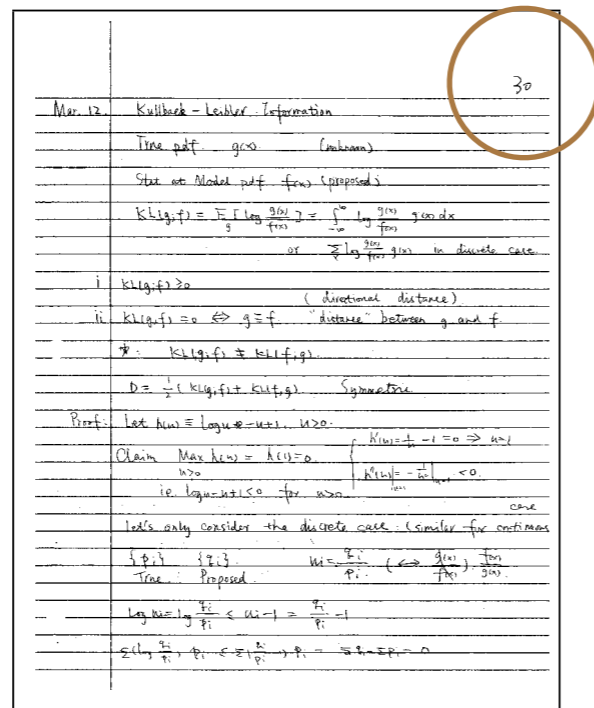
which yields $\lambda = \frac{h}{m_0 c} (1 - \cos \theta)$ - I don't think you can derive the right conclusion based on the wrong equation.

Minimum elicited and curated handwriting!



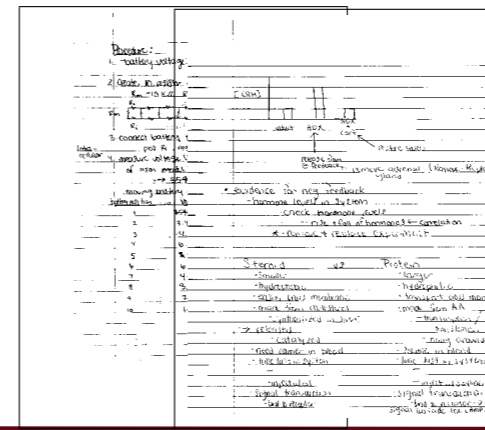
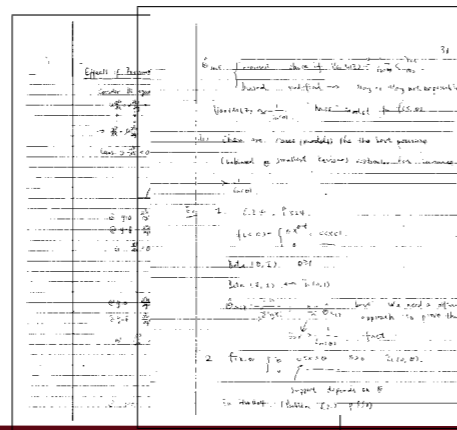
Page Order Analysis

- *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.

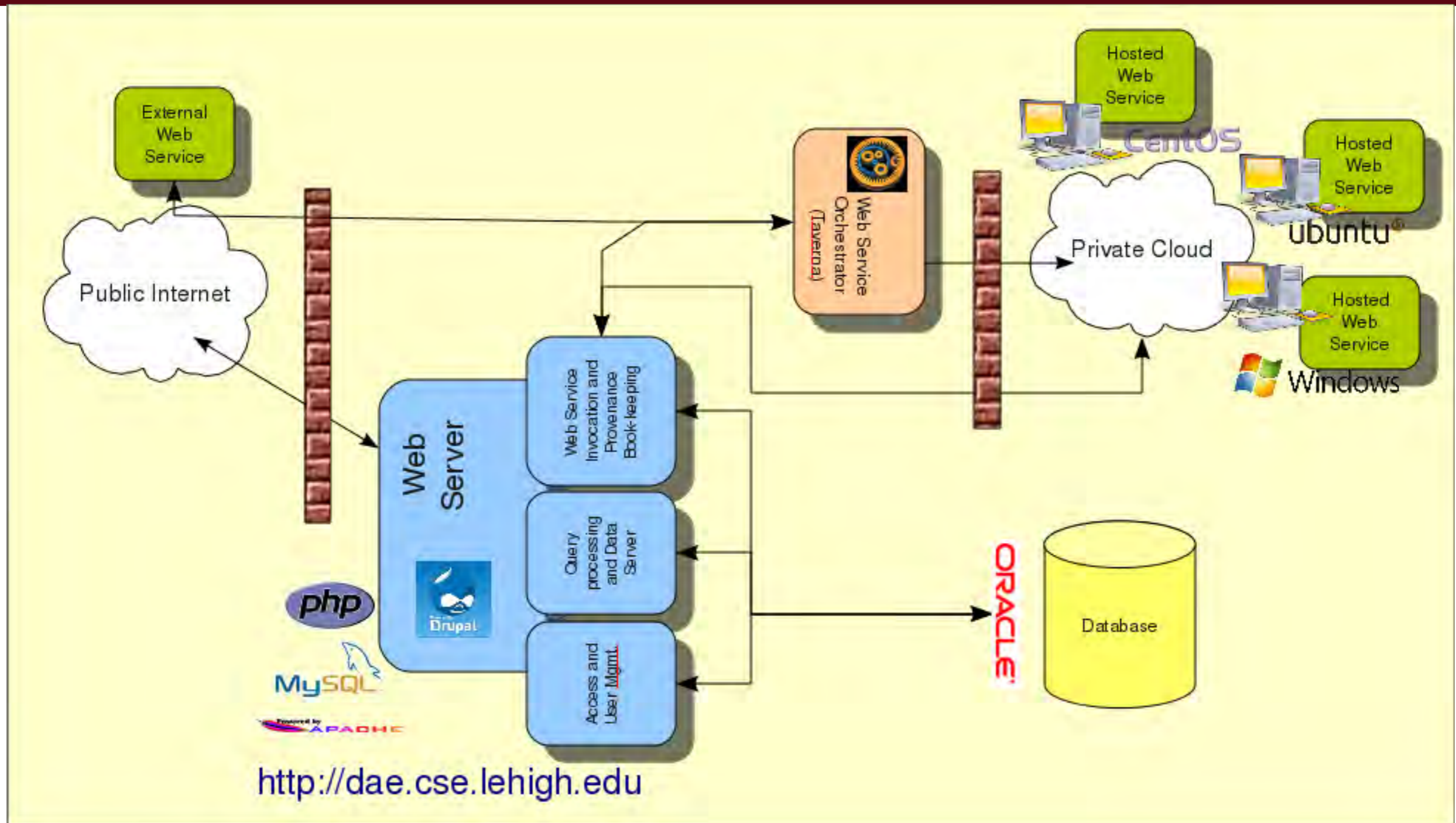


Structure Restoration

- 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



This slide is from the DRR 2011 talk by Lopresti and Lamiroy.



Document part annotation

Search this site:

admin

- Browse Data
- ▷ My Uploads
- Algorithms
- My Runs
- My account
- ▷ Create
- ▷ Admin
- Log out
- Repository

Home

1_3.JPG ★★★★☆

Submitted By: John Appleseed

Show All

Improving the Quality of Degraded Document Images

Virginia Kavallieratou and Efsthios Stamatatos

Information and Communication Systems Engineering,
University of the Aegean
83200 - Karlovassi, Greece
{v.kavallieratou, e.stamatatos}@aegean.gr

Abstract

It is common for libraries to provide public access to historical and ancient document image collections. It is common for such document images to require specialized processing in order to remove background noise and become more legible. In this paper, we propose a hybrid binarization approach for improving the quality of old documents using a combination of global and local thresholding. First, a global thresholding technique specifically designed for old document images is applied to the entire image. Then, the image areas that still contain background noise are detected and the same technique is re-applied to each area separately. Hence, we achieve better adaptability of the thresholding process to various image backgrounds.

remove noise from historical document images and improve their quality before libraries expose them to public view. Within this framework, noise is considered anything that is irrelevant with the textual information (i.e., foreground) of the document image. Image analysis systems use binarization as a standard procedure to convert a grey-scale image to binary form. An ideal binarization algorithm would be able to perfectly discriminate foreground from background, thus removing any kind of noise that obstructs the legibility of the document image. The binary image is ideal for further processing [5] (e.g., discrimination of periods from handwritten text, recognition of the contents by applying OCR techniques etc). However, in the framework of a library collection of historical and ancient documents,

Document tagging

- ✓ Paragraph
 - Element 1
- ✓ Column
 - Element 2
 - Element 3
- ✓ Title
 - Element 3
- ✓ Name
 - Element 4
- ✓ Heading
 - Element 5
 - Element 6

Document Analysis and Exploitation newsletter

Stay informed on our latest news!

User: admin

[Unsubscribe](#)

[Previous issues](#)

★★★★☆ admin

This document is quite nicely annotated and shows embedded and overlapping annotations. However, bounding boxes are not very precise.

Rating & commenting

Newsletters and discussion groups

This slide is from the DRR 2011 talk by Lopresti and Lamiroy.

XML Markup

```
<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>
```

Authorship

Notebook ID

Subject Tags

Ruling Line Specifications, etc.



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 is available via: <http://dae.cse.lehigh.edu/DAE/>. The full-color version will be uploaded soon.
- We also call for discussions on its usage.