Document Analysis

Exercise 3 : Keyword Spotting / Week 3

Andreas Fischer

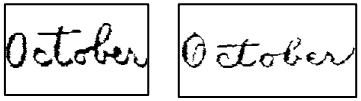
andreas.fischer@unifr.ch

UNI
FR

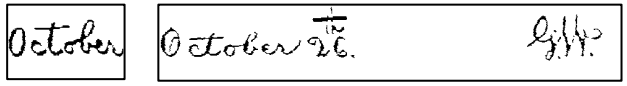
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Outline of the Exercise

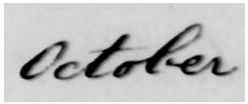
- Week 1
  - Dissimilarity between preprocessed keyword images and preprocessed word images
  - Output: ordered list of words IDs

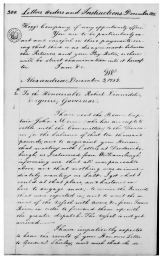


- Week 2
  - Dissimilarity between preprocessed keyword images and preprocessed text line images
  - Output: ordered list of text line IDs



- **Week 3** (ambitious - it is **optional**)
  - Dissimilarity between keyword images and automatically extracted text line images
  - Output:
    - List of text lines together with their bounding box
    - Ordered list of text line IDs

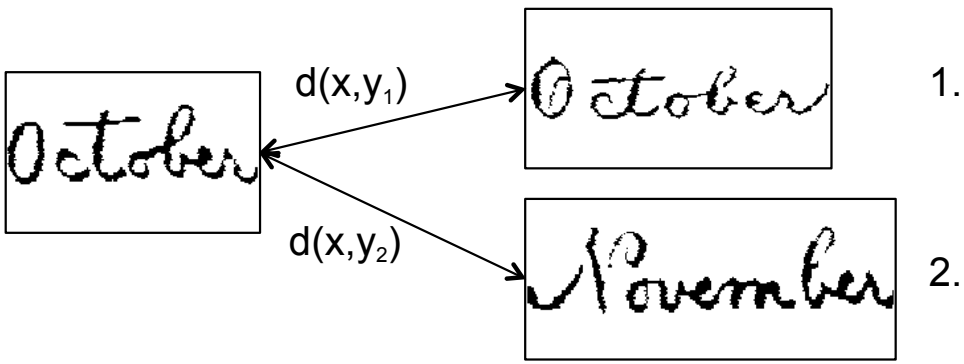


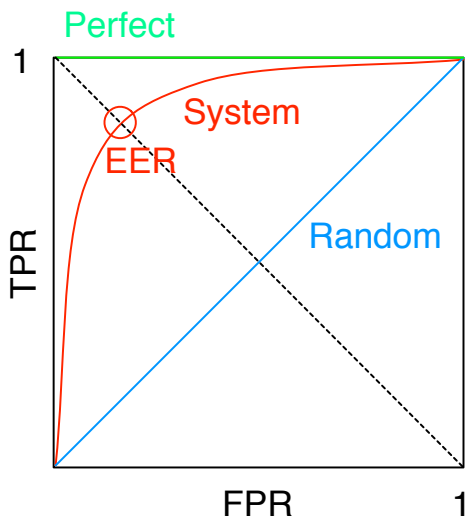


UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Week 1

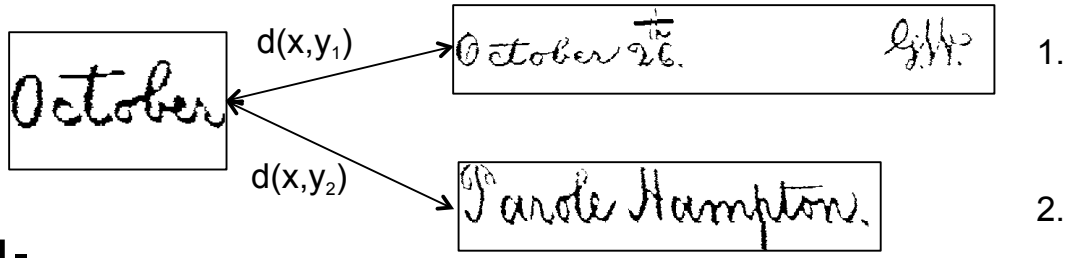
- Tasks: for **both data sets** and **each keyword image**
  - Compute a dissimilarity to all word images and order the word images accordingly
  - Compute the Receiver Operating Characteristic (ROC) Curve and the Equal Error Rate (EER)

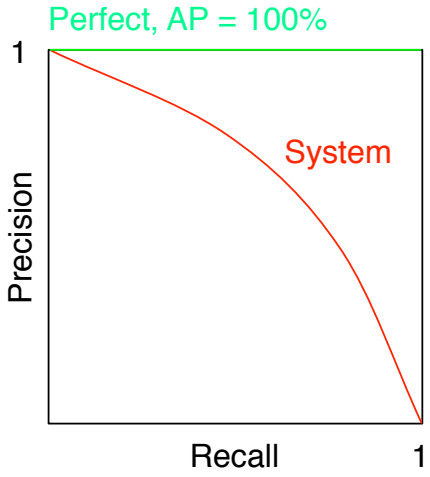




Week 2

- Tasks: for **both data sets** and **each keyword image**
  - Compute a dissimilarity to all text line images and order the text lines accordingly
  - Compute the ROC Curve, the EER, the Recall-Precision Curve, and the Avarage Precision (AP)
- Feedback from groups
  - Which **methods** did you work on?
  - What **problems** did you encounter?
  - Do you already have **results**? Are they promising?

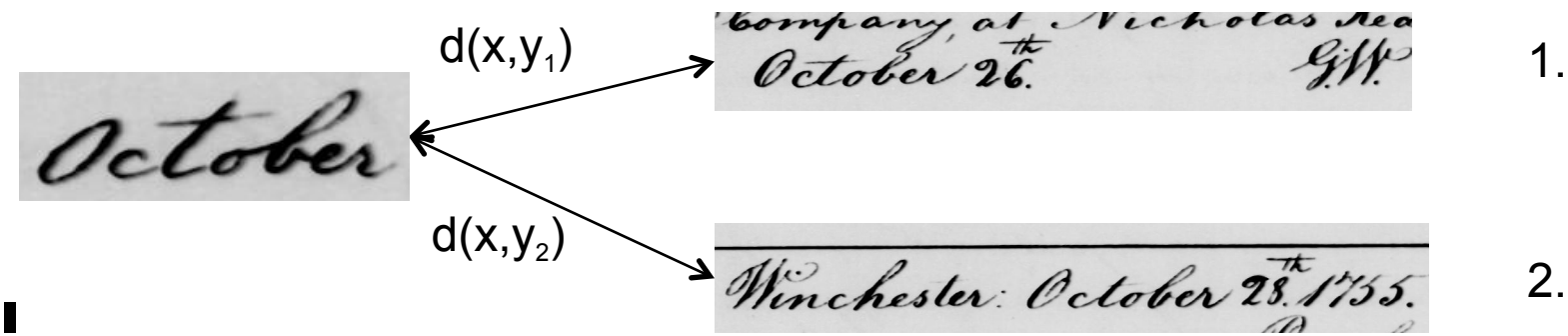




UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Week 3

- Data: available for download on Ilias
  - Keyword images (same as in weeks 1+2, no preprocessing)
  - Page images (no preprocessing)
- Tasks: for **both data sets** and **each keyword image**
  - Extract text line images from the page images
  - Compute a dissimilarity to all text line images and order the text lines accordingly
  - Visually inspect the top 50 results and compute recall and precision



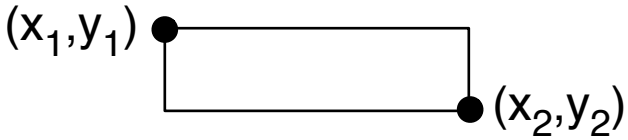
Output Week 3

- List of text lines together with their bounding boxes (x_1,y_1,x_2,y_2):

  "WashingtonDB_Lines.txt"

  270-01 645 213 1014 267

  …



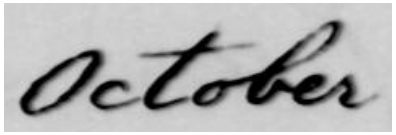
- Ordered list of text line IDs for each database and keyword:

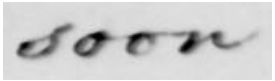
  "WashingtonDB_O-c-t-o-b-e-r.txt"

  271-11

  …

- Recall and precision for the top 50 results are part of the report.





UNI FR
UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

Deadline

- **May 5, 15:00**
- Hand in a ZIP file **“HansMuster_JaneDoe_JohnDoe.zip”** via Ilias that contains:
  - Your source code
  - Report with descriptions and figures (PDF)
  - Output files (list of IDs and bounding boxes)
- Exercises will be accepted if at least tasks 1 and 2 (preprocessed images) have been carefully addressed.
- There will be an evaluation and discussion of your results.
  - Who can solve task 3?
  - Which method achieves the best results for tasks 1-3?

Want More? Available Research Projects

- Document Analysis:
  - Keyword Spotting (hierarchical approach)
  - Signature Verification (confidence modeling)
  - Document Classification (graph-based approach)
- Human Machine Interaction:
  - Intelligent Document Annotation (learn from user corrections)
- Biomedical Applications:
  - Automatic Detection of Heart Rhythm Disorders (ECG signals, HMM-based)
- Combinatorial Optimization:
  - Graph Edit Distance & the Quadratic Assignment Problem (formal and applied)
  - Parallelization of Graph Edit Distance (Hadoop-based)