D-Scribe - Automatic Authorship Identification and Clustering

Measurement Science and Standards in Forensic Handwriting Analysis
D-Scribe – Automatic Authorship Identification

Table of content

- Postal roots
- Global OCR
- Use cases
- Architecture
- Structured feature extractor
- Textual feature extractor
- Allograph feature extractor
- Feature analysis
- Cluster analysis
- Potential applications
Postal recognition tasks represent a difficult proving ground

- Siemens reads everything on a letter in <1 second
- 150 Billion mail pieces per year
Region of Interest

Recognition results of chars

北京巿朝阳区望京广顺南大街嘉润花园19号写字楼B座2层

Interpretation results

(City, district, road, block, number), (Building, house, floor)
Questions to be answered by biometric handwriting detection

- Which handwriting from a group is similar to a given handwriting sample?
- How similar are two handwriting samples?
- How many authors wrote a set of documents, and how can these be organized by author?
D-Scribe – Automatic Authorship Identification
Structured Feature Extractor (1/4)

Input document

Pre-processing

Line and word segmentation

Structured feature extraction

Template Generator / OCR

Input document (200 DPI gray)

Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zeimont St. and then go to Turin and Rome and will join Col. Parry and arrive at Athens, for 16:00 on Dec. 24th. Letters should be addressed: John James Lloyd 3580. We expect Charles F. Fuller Tuesday. Dr. L. McQuaid and Robert Unwin, Esq., lift on the "X" express tonight.
Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland and I hope for good news. He will be there for a week at 1496 Zeimott St. and then go to Turin and Rome and will join Col. Parry and arrive at Athens, Greece, Nov. 27th. On Dec. 2nd, letters should be addressed: King James Blvd 3540. We expect Charles Fuller Tuesday. Dr. L. McQuaid and Robert Underhill, Esq., lift on the "Y. X." express tonight.
D-Scribe – Automatic Authorship Identification
Structured Feature Extractor (3/4)

Input document

Pre-processing

Line and word segmentation

Structured feature extraction

Template Generator / OCR

Segmented lines and words

- Our London business is good, but Vienna and Berlin are quiet. Mr. D. Lloyd has gone to Switzerland.
- And I hope for good news. He will be there for a week at 1996 Zeitoff St. and will give us further information.
- Will join our party and arrive at Athens, Harry, and our friends.
- Our Dec. 2nd. letters should be addressed: John James.
- By Nov. 35th. we expect Charles F. Fuller Tuesday, Dr. L. McQuaid, and Robert Unger, Esq., lift on the 'F.B.' express tonight.
D-Scribe – Automatic Authorship Identification
Structured Feature Extractor (4/4)

**Input document**

**Pre-processing**

**Line and word segmentation**

**Structured feature extraction**

**Template Generator / OCR**

---

W002-1 ./statistic/docfeature/n_lines 8
W002-1 ./statistic/docfeature/avg_height_of_line 69.000000
W002-1 ./statistic/docfeature/avg_length_of_line 2091.125000
W002-1 ./statistic/docfeature/avg_blackness_of_line 0.079064
W002-1 ./statistic/docfeature/avg_runlength_x_of_line 5.692302
W002-1 ./statistic/docfeature/avg_runlength_y_of_line 6.725820
W002-1 ./statistic/docfeature/avg_slope_of_line 0.012695
W002-1 ./statistic/docfeature/avg_segdist_of_line 28.944444
W002-1 ./statistic/docfeature/avg_n_words_per_line 3.125000

**Line and word features**
D-Scribe – Automatic Authorship Identification
Textual Feature Extractor (1/3)

Pre processed image

Contour direction comparison

Contour hinge comparison

Direction Co-occurrence comparison

Contour Direction PDF (Bulacu, 2007)

CD Histogram Comparison (Bulacu, 2007)

Contour Hinge PDF (Bulacu, 2007)

Pre processed image

Contour direction comparison

Contour hinge comparison

Direction Co-occurrence comparison

CH Histogram Comparison (Bulacu, 2007)

D-Scribe – Automatic Authorship Identification Textual Feature Extractor (3/3)

- Pre processed image
- Contour direction comparison
- Contour hinge comparison
- Direction Co-occurrence comparison

Direction Co-Occurrence PDF (Bulacu, 2007)

Direction COO Histogram Comparison (Bulacu, 2007)


Michael Carpenter, Matthias Schulte-Austum
Mr. Powell finds it easier to take childrens and sick people than to industry, Mr. Bann commented.

full inquiry into the cost of drugs an

(Bulacu, 2007)
D-Scribe – Automatic Authorship Identification
Key features

- Uses advanced and proven image preprocessing algorithm basis
  - Image improvements
  - Text Line extraction
  - Hand/Machine decision
  - Advanced underline and noise removal

- Simple decisions and confidences
- Portable and efficient software
  - Small footprint
  - Low runtime
  - Enables large database analysis
  - Android port available
### D-Scribe – Feature analysis

#### Used test sets

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Language</th>
<th>Documents</th>
<th>Writers</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENG</td>
<td>English</td>
<td>2535</td>
<td>507</td>
<td>5 documents per writer, 4 predefined texts + an arbitrary text</td>
</tr>
<tr>
<td>ARAB_1</td>
<td>Arabic</td>
<td>5000</td>
<td>45</td>
<td>50-55 documents per writer, large variety of documents (different background, different pens, artificial documents)</td>
</tr>
<tr>
<td>ARAB_2</td>
<td>Arabic</td>
<td>1000</td>
<td>200</td>
<td>2-3 documents per writer</td>
</tr>
</tbody>
</table>

(including unknown writer docs)
Two feature configuration sets

- **4 Features:**
  - grapheme_snn_split (500)
  - grapheme_snn_points (150)
  - hinge_improved_rotated_fragment (1536)
  - hinge_improved_broi_textline (1536)

- **6 Features:**
  - grapheme_snn_split (500)
  - grapheme_snn_points (150)
  - hinge_improved_rotated_fragment (1536)
  - hinge_improved_broi_textline (1536)
  - simple_writing_direction (12)
  - hinge_contour_approximation (144)
D-Scribe – Feature analysis
Different projections of the ENG data set

PCA

MDS
D-Scribe – Feature analysis
PCA Projections of the ENG and ARAB_2 Data Sets

Side effect:
Features can be used to separate by writing system (e.g. Arabic language from Latin)

Black: Latin, Red: Arabic

Works also with other data sets
Clustering
How effective was the clustering?

Recall

Purity
Final Cluster Quality Score

Purity is not as relevant as Recall, because the error putting different writers into same cluster is not as important as putting the same writer in different clusters.

Set of elements: \( S = \{o_1, \ldots, o_n\} \)  
Ground Truth: \( X = \{X_1, \ldots, X_r\} \)

Cluster assignment: \( Y = \{Y_1, \ldots, Y_s\} \)  
Purity: \( \text{purity} = \frac{1}{n} \sum_{y_j \in Y} \max_{x_i \in X} |y_j \cap x_i| \)

Recall: \( R = \left( \frac{|\{\text{pairs in same set in } X \text{ and in } Y\}|}{|\{\text{pairs in same set in } X\}|} \right) \)

Harmonic Mean of Purity and Recall

\[
\text{score}_\beta = \left(1 + \beta^2\right) \frac{\text{purity} \cdot R}{\beta^2 \cdot \text{purity} + R}
\]
\[
\text{score}_1 = 2 \frac{\text{purity} \cdot R}{\text{purity} + R}
\]

Recall is \( \beta \) times important than purity.
Finding Optimal Cluster Method on the ENG Data Set with 6 feature sets

The clustering parameters are a trade off between recall and purity, requiring high recall automatically means low purity and vice versa. The manhattan distance performs in all cases better than the euclidian distance. The different cluster methods have different behaviors, either single linkage or Ward’s method are considered best.
Dendrogram of Clustering with Ward's Method
Recall and Purity

Optimal according to quality score
Optimal Cut Point with different samples of data

The height in the dendrogram (intra cluster distance) is the same for different samples of the data set.
D-Scribe – Automatic Authorship Identification

Potential Applications

- OCR
  - Keyword screening
  - Machine translation (subsystem)
  - OCR combination framework (voting)

- Writer ID as a biometric

- Clustering
D-Scribe – Automatic Authorship Identification
Potential Applications

- OCR
- Writer ID as a biometric
  - Screening an questioned document against a known document
    - Document to document
  - Screening documents against a watch list
    - Document to documents
- Clustering
D-Scribe – Automatic Authorship Identification
Potential Applications

- OCR
- Writer ID as a biometric
- Clustering
  - Clustering handwritten and machine printed
  - Clustering by writing system / language
  - Clustering by author
D-Scribe – Automatic Authorship Identification
Potential Applications

- OCR
- Writer ID as a biometric
- Clustering
  - Basic Triage:
    - Eliminating non-relevant documents
    - Separating Machine Printed from Handwritten
    - Separating documents according to writing system
    - Separating documents by author
    - Focus on potentially relevant documents
D-Scribe – Automatic Authorship Identification
Contact information

Mike Carpenter
Product Manager
Logistics and Airport Solutions
2700 Esters Blvd., Suite 200B
DFW Airport, TX 75261
Phone: +1 (972) 947-7491
Mobile: +1 (817) 307-2228
E-mail: michael.carpenter@siemens.com

siemens.com/answers