Writer Identification Using a Statistical And Model Based Approach

Diamantatos Paraskevas
Gritzalis Stefanos
Kavallieratou Ergina
University Of The Aegean
Presentation Overview

- Writer Identification-Objective
- Related work
- Skeleton Hinge distribution
- Proposed System
- Experimental Data & Results
- Conclusion
Matching unknown handwritings against a database of samples with known authorship.
Objective

✓ Improvement of existing techniques
✓ Edge-hinge distribution
✓ Edge-hinge combinations
✓ Hypothesis: Stroke width has size 1.

«Make everything as simple as possible, but not simpler.»

Albert Einstein
Edge-Direction Distribution

- Edge detection.
- Direction of edge.
- Histogram of Directions.
- Normalized to a probability distribution.
- Nearest Neighbor.

Accuracy: 35% for 250 writers.

Edge-Hinge Distribution*

✓ Statistical Feature.
✓ Outperforms all the other statistical approaches.
✓ Based on Edge-direction distribution.

Edge-Hinge Distribution*

- Edge detection.
- Directions of edges $\phi_1$, $\phi_2$ with $\phi_2 > \phi_1$.
- Histogram of Directions.
- Normalized to a probability distribution $p(\phi_1, \phi_2)$.
- Nearest Neighbor

Accuracy: 63% for 250 writers.
Edge-Hinge Distribution
Edge-Hinge Combinations*

✓ Improvement of Edge-Hinge Distribution.
✓ Similar with Edge-Hinge Distribution.
✓ Multiple length edge fragments (windows sizes).
✓ Best results on size combinations 3, 5, 7, 9.

Accuracy: 81% for 250 writers.

Skeleton-Hinge Distribution

- Improvement of previous methods.
- Similar technique.
- Skeleton information.
- Hypothesis: All stroke widths should be considered the same.
- Shorter execution time by 35%.

Accuracy: 90.8% for 250 writers.
Hypothesis: the writer acts as a stochastic generator of ink-blob shapes, or graphemes.

Codebook of Models of Graphemes*

✓ Words gets fragmented to graphemes.
✓ Grapheme codebook generation using self-organizing feature map (SOFM)
✓ Histogram of grapheme models
✓ Nearest Neighbor

System Approach

Train Data set

segmentation

SOFM train

Test Data set

segmentation

Skeleton-Hinge Distribution

Codebook Matching

SHD train features

COG train features

Nearest Neighbor

COG test features

SHD test features
Data*

- Firemaker DB
- 250 writers
- 4 pages per writer
  - 1st page: Copied text
  - 2nd page: Freestyle text
  - 3rd page: Upper case copied text
  - 4th page: Copied text with forge attempt
- Train set: 1st page
- Test set: 2nd page

Skeleton-Hinge Distribution
Experimental Results

<table>
<thead>
<tr>
<th>Fragment Length</th>
<th>Manhattan distance Performance</th>
<th>Euclidian distance Performance</th>
<th>Chi-square distance Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>80%</td>
<td>72%</td>
<td>53.2%</td>
</tr>
<tr>
<td>5</td>
<td>89.6%</td>
<td>77.2%</td>
<td>66%</td>
</tr>
<tr>
<td>7</td>
<td>90%</td>
<td>81.6%</td>
<td>69.6%</td>
</tr>
<tr>
<td>9</td>
<td>88%</td>
<td>85.2%</td>
<td>76%</td>
</tr>
<tr>
<td>3, 5</td>
<td>85.2%</td>
<td>75.2%</td>
<td>58.4%</td>
</tr>
<tr>
<td>3, 7</td>
<td>85.6%</td>
<td>75.6%</td>
<td>55.2%</td>
</tr>
<tr>
<td>3, 9</td>
<td>86%</td>
<td>74.8%</td>
<td>53.2%</td>
</tr>
<tr>
<td>5, 7</td>
<td>90%</td>
<td>78.8%</td>
<td>64.4%</td>
</tr>
<tr>
<td>5, 9</td>
<td>90.8%</td>
<td>78.8%</td>
<td>67.2%</td>
</tr>
<tr>
<td>7, 9</td>
<td>90%</td>
<td>83.2%</td>
<td>73.6%</td>
</tr>
<tr>
<td>3, 5, 7</td>
<td>86.8%</td>
<td>76.8%</td>
<td>60%</td>
</tr>
<tr>
<td>3, 7, 9</td>
<td>89.6%</td>
<td>76.8%</td>
<td>55.6%</td>
</tr>
<tr>
<td>5, 7, 9</td>
<td>90%</td>
<td>79.2%</td>
<td>68.8%</td>
</tr>
<tr>
<td>3, 5, 7, 9</td>
<td>89.6%</td>
<td>76.8%</td>
<td>60.4%</td>
</tr>
</tbody>
</table>
## Experimental Results

<table>
<thead>
<tr>
<th>Number of writers</th>
<th>Codebook Size</th>
<th>Manhattan distance Performance</th>
<th>Euclidian distance Performance</th>
<th>Chi-square distance Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>225</td>
<td>95.6%</td>
<td>91.2%</td>
<td>78.8%</td>
</tr>
<tr>
<td>150</td>
<td>225</td>
<td>96%</td>
<td>94.7%</td>
<td>86.7%</td>
</tr>
</tbody>
</table>
## Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Edge Direction Distribution</td>
<td>35%</td>
</tr>
<tr>
<td>Edge-Hinge Distribution</td>
<td>63%</td>
</tr>
<tr>
<td>Edge-Hinge Combinations</td>
<td>81%</td>
</tr>
<tr>
<td>Skeleton-Hinge Distribution</td>
<td>90.8%</td>
</tr>
<tr>
<td>System Approach(^1)</td>
<td>96%</td>
</tr>
<tr>
<td>Schomaker Approach(^2)*</td>
<td>97%</td>
</tr>
</tbody>
</table>

1 Codebook size 15 X 15  
2 Codebook size 33 X 33  

Conclusions

✓ A single statistical feature achieves high accuracy
✓ Our hypothesis proved right.
✓ Codebook of graphemes combined with skeleton hinge reached accuracy of 96%.
Thank you

«If I have seen further it is by standing on the shoulders of giants.»