Handwriting Copybook Style Analysis of Pseudo-Online Data

Mary L. Manfredi, Sung-Hyuk Cha, Sungsoo Yoon, and Charles C. Tappert
Computer Science Department, Pace University
861 Bedford Road, Pleasantville, New York, 10570 USA
mmanfredi@att.com, {scha, syoon, ctappert}@pace.edu

Abstract

A common forensic problem is determining the writer of a questioned document. Identifying the copybook style of a questioned document can help reduce the suspect population as an important step towards the identification of an individual writer. This feasibility study presents a method of identifying the copybook style of a questioned document using pseudo-online data and a string edit distance. In addition, clustering analysis performed on the database reveals similarities among the copybook styles.

1. Introduction

Handwritten language continues to be a heavily used means of communication and therefore handwriting analysis and recognition also continue to be pursued. There are two types of data used in handwriting analysis – offline and online [9, 11]. Offline handwriting analysis operates on data that has been previously written and then scanned in as an image. Online handwriting analysis operates on data captured in real time as the writer is writing, for example, on a pen-enabled tablet. In addition to the static information the online data contains dynamic information which can be useful in the recognition process. Pseudo-online data is data created by tracing offline data to give it dynamic characteristics.

In questioned document analysis, handwriting is almost always in the form of offline image data. Since it has been shown, using the same underlying data, that online information is superior to offline [9], we explore the use of pseudo-online data in this feasibility study.

Handwriting originates from a particular copybook style, such as Palmer, Zaner-Bloser, or D’Nealian, that one learns in childhood. Questioned document examination plays an important investigative and forensic role in many types of crime [1, 6], and being able to identify the copybook style of a questioned document could reduce the scope of the suspect population in the identification of an individual writer.

To analyze the copybook styles of a questioned document, we create a database containing the character image data of the copybook styles. Our current database contains 19 Roman alphabet copybook styles: one manuscript style and 18 cursive styles from 15 countries. Then we obtain pseudo-online data of these copybook styles by tracing the characters of each style. This allows us to calculate a string edit distance between pairs of characters using the Stroke Direction Sequence Strings (SDSS) [3].

Using this distance metric, we develop a similarity-based copybook style identification system and also perform a cluster analysis of copybook style characters. Further details can be found in the dissertation study that this paper summarizes [8]. First, a similarity-based pattern matching algorithm allows comparison between the characters of a questioned document and those in the database. However, using only a distance metric to determine the copybook style, although partially successful, is not sufficiently accurate because many copybook styles are similar and some virtually indistinguishable. Therefore, a clustering analysis of the available copybook styles is also performed, where clustering is the unsupervised classification of patterns into groups/clusters [4, 7].

A search of the literature shows work done on the categorization of allographs using individual writer data [13]. Other literature describes the use of clustering techniques on numeral recognition [5, 10], which is similar to the clustering we perform here on alphabetic characters.

This paper is organized as follows. Section 2 describes the database creation, the pseudo-online extraction process, and the string edit distance metric. Section 3 contains the string matching classification results,
section 4 describes the cluster analysis, and section 5 draws some conclusions and mentions possible extensions of this work.

2. Copybook Style Database, Pseudo Online Data Extraction, and Distance Metric

The data accumulated for the offline database consists of 19 Roman alphabet copybook styles (Figure 1 shows sample copybook styles) from 15 countries: Austria (1 cursive style), Belgium (1 cursive), Brazil (1 cursive), Canada (1 cursive), Chile (1 cursive), Columbia (1 cursive), Denmark (1 cursive), Ecuador (1 cursive), England (1 cursive), Germany (1 cursive), Netherlands (1 cursive), Norway (1 cursive), Peru (2 cursive), Switzerland (2 cursive), United States (2 cursive, 1 manuscript). There are a total of 18 cursive styles and one manuscript style. These copybook characters were obtained from various books and websites, and were used in an earlier study [2].

Figure 1. Sample copybook styles.

The online database was created by tracing each offline alphabetic character from each of the aforementioned copybook styles using a digital pen and tablet. The tracing was performed by using the number of strokes estimated as most appropriate from an examination of the offline images, and then drawing the strokes in a standard order, using top-down and left-to-right directions. The tracing allows the capture of the dynamic characteristics of the data, including the time-ordered sequence of (x, y) points and the derived SDSS feature extractions.

Although a variety of feature extraction methods appear in the literature [12], here we use the SDSS method [3]. A stroke is defined as the set of points between a pen-down and the next pen-up. Each alphabetic character is represented as a sequence of directions (arrows) and is thus represented as an angular type string. Each direction is quantized into one of 8 directional values as shown in Figure 2, and the quantized values are used to calculate distances between characters.
Capturing the data as pseudo-online data allows string matching to be used for character matching. For string matching we use the modified Levenshtein edit distance to handle angular strings utilizing the “turn” concept in place of substitution [3].

3. String Matching Results

The similarity-based, pattern-matching technique compares the quantized SDSS feature vector of each letter in the questioned document to the database of SDSS feature vectors of the copybook style letter database. The matching procedure for classification is straightforward. Each upper or lowercase letter of the questioned document is compared to its corresponding letter in all the copybook styles. Figure 3a and 3b show a sample questioned document from Switzerland as input and the resulting distance table as output.

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>i</th>
<th>r</th>
<th>s</th>
<th>c</th>
<th>h</th>
<th>94.0815</th>
<th>115.763</th>
<th>62.4379</th>
<th>115.0544</th>
<th>104.2668</th>
<th>83.4495</th>
<th>101.193</th>
</tr>
</thead>
<tbody>
<tr>
<td>austria2_C</td>
<td>61.86</td>
<td>131.1067</td>
<td>85.4835</td>
<td>109.2796</td>
<td>65.3644</td>
<td>94.0815</td>
<td>186.1683</td>
<td>115.763</td>
<td>62.4379</td>
<td>115.0544</td>
<td>104.2668</td>
<td>83.4495</td>
<td>101.193</td>
</tr>
<tr>
<td>belgium1_C</td>
<td>92.7863</td>
<td>131.5697</td>
<td>44.1003</td>
<td>101.2601</td>
<td>79.8979</td>
<td>146.6377</td>
<td>98.9261</td>
<td>48.9866</td>
<td>80.3063</td>
<td>125.4184</td>
<td>44.6155</td>
<td>90.31625</td>
<td></td>
</tr>
<tr>
<td>brasil1_C</td>
<td>126.8338</td>
<td>111.2795</td>
<td>126.5907</td>
<td>103.6344</td>
<td>69.2081</td>
<td>90.5066</td>
<td>179.1611</td>
<td>120.5368</td>
<td>86.2276</td>
<td>101.9792</td>
<td>96.6764</td>
<td>133.1248</td>
<td>112.1463</td>
</tr>
<tr>
<td>canada1_C</td>
<td>43.4166</td>
<td>178.5015</td>
<td>113.1581</td>
<td>98.5721</td>
<td>112.4954</td>
<td>111.0579</td>
<td>179.8836</td>
<td>146.2421</td>
<td>94.2447</td>
<td>119.6072</td>
<td>145.4287</td>
<td>129.2410</td>
<td>122.654</td>
</tr>
<tr>
<td>chile1_C</td>
<td>137.8739</td>
<td>164.6059</td>
<td>124.6229</td>
<td>114.1161</td>
<td>99.6742</td>
<td>124.4464</td>
<td>202.8162</td>
<td>154.2122</td>
<td>79.0622</td>
<td>112.3633</td>
<td>156.9452</td>
<td>122.7928</td>
<td>132.8189</td>
</tr>
<tr>
<td>colombia1_C</td>
<td>85.0178</td>
<td>78.124</td>
<td>146.1169</td>
<td>136.5091</td>
<td>95.0121</td>
<td>71.8073</td>
<td>173.6923</td>
<td>101.4297</td>
<td>98.1851</td>
<td>53.5495</td>
<td>87.4724</td>
<td>97.5557</td>
<td>102.0993</td>
</tr>
<tr>
<td>denmark1_C</td>
<td>104.8951</td>
<td>67.9394</td>
<td>91.6346</td>
<td>83.6516</td>
<td>70.6963</td>
<td>129.7213</td>
<td>126.2796</td>
<td>48.2262</td>
<td>79.7733</td>
<td>121.5205</td>
<td>45.9902</td>
<td>89.21165</td>
<td></td>
</tr>
<tr>
<td>ecuador1_C</td>
<td>132.4394</td>
<td>175.5389</td>
<td>130.6275</td>
<td>67.3364</td>
<td>91.5346</td>
<td>63.6516</td>
<td>70.6963</td>
<td>129.7213</td>
<td>126.2796</td>
<td>48.2262</td>
<td>79.7733</td>
<td>121.5205</td>
<td>45.9902</td>
</tr>
<tr>
<td>england1_C</td>
<td>43.4166</td>
<td>178.5015</td>
<td>113.1581</td>
<td>98.5721</td>
<td>112.4954</td>
<td>111.0579</td>
<td>179.8836</td>
<td>146.2421</td>
<td>94.2447</td>
<td>119.6072</td>
<td>145.4287</td>
<td>129.2410</td>
<td>122.654</td>
</tr>
<tr>
<td>germ3_C</td>
<td>79.5436</td>
<td>81.8939</td>
<td>50.9619</td>
<td>84.3473</td>
<td>96.9446</td>
<td>73.0348</td>
<td>113.0648</td>
<td>112.4488</td>
<td>73.7017</td>
<td>51.8581</td>
<td>98.2365</td>
<td>90.0624</td>
<td>83.9481</td>
</tr>
<tr>
<td>netherland1_C</td>
<td>100.9746</td>
<td>158.4266</td>
<td>144.2128</td>
<td>197.23</td>
<td>74.6069</td>
<td>78.368</td>
<td>214.2917</td>
<td>111.3614</td>
<td>172.9847</td>
<td>121.8061</td>
<td>108.3235</td>
<td>123.3629</td>
<td>133.8038</td>
</tr>
<tr>
<td>norway1_C</td>
<td>60.6239</td>
<td>177.194</td>
<td>107.3982</td>
<td>85.9694</td>
<td>60.1095</td>
<td>70.7302</td>
<td>109.4171</td>
<td>107.8047</td>
<td>36.0841</td>
<td>77.1245</td>
<td>99.8826</td>
<td>59.0032</td>
<td>79.26605</td>
</tr>
<tr>
<td>peru1_C</td>
<td>92.6681</td>
<td>130.5025</td>
<td>53.0946</td>
<td>102.5164</td>
<td>59.4201</td>
<td>139.0174</td>
<td>125.7199</td>
<td>43.994</td>
<td>71.3015</td>
<td>102.8196</td>
<td>65.4165</td>
<td>89.67915</td>
<td></td>
</tr>
<tr>
<td>peru2_C</td>
<td>126.3819</td>
<td>173.4229</td>
<td>168.7966</td>
<td>142.423</td>
<td>69.6073</td>
<td>161.5705</td>
<td>227.8518</td>
<td>119.112</td>
<td>99.0176</td>
<td>184.6229</td>
<td>121.7613</td>
<td>128.0816</td>
<td>143.559</td>
</tr>
<tr>
<td>switzerland1_C</td>
<td>23.4585</td>
<td>101.7896</td>
<td>89.2399</td>
<td>42.3324</td>
<td>61.4163</td>
<td>31.9927</td>
<td>151.9341</td>
<td>95.0172</td>
<td>49.3324</td>
<td>59.8047</td>
<td>77.557</td>
<td>57.9531</td>
<td>70.15166</td>
</tr>
<tr>
<td>switzerland3_C</td>
<td>36.4256</td>
<td>103.6153</td>
<td>42.6524</td>
<td>101.7896</td>
<td>90.5329</td>
<td>54.6536</td>
<td>147.8869</td>
<td>98.5662</td>
<td>59.4333</td>
<td>75.519</td>
<td>85.1064</td>
<td>72.5456</td>
<td>80.72218</td>
</tr>
<tr>
<td>usa1_C</td>
<td>44.8082</td>
<td>146.4811</td>
<td>120.6671</td>
<td>86.8788</td>
<td>109.4353</td>
<td>103.0929</td>
<td>176.6708</td>
<td>144.5657</td>
<td>70.6627</td>
<td>122.5055</td>
<td>142.2026</td>
<td>84.6051</td>
<td>112.7077</td>
</tr>
<tr>
<td>usa2_C</td>
<td>43.2158</td>
<td>180.054</td>
<td>119.1866</td>
<td>100.0869</td>
<td>114.6572</td>
<td>118.9346</td>
<td>183.5502</td>
<td>147.1417</td>
<td>95.0501</td>
<td>121.3252</td>
<td>149.9143</td>
<td>127.7194</td>
<td>125.0697</td>
</tr>
</tbody>
</table>

Figure 2. A sample offline character image and its traced SDSS feature vector.

Figure 3. A questioned input sample (a) and output table of copybook style character matches (b).

The output table allows the identification, or partial identification (narrowing), of the copybook style of the questioned handwriting by selecting the most similar copybook style to the questioned handwriting. The last column of the figure gives, for each copybook style, the average distance over all the letters of the questioned document. In this example, the smallest average distance, 70.15, indicates that the style of the questioned document is closest to the copybook style from Switzerland1_C.
Although the accuracy on individual letters may be low, accuracy increases with the length of the questioned document as more letters contribute to the decision. However, to better handle short questioned documents and the fact that some copybook styles are so similar that they are virtually indistinguishable, we investigated a clustering analysis on the copybook styles.

4. Cluster Analysis

We applied the same modified Levenshtein edit distance to compare all copybook style characters for cluster analysis. This analysis provides insight into the degree of similarity or dissimilarity among copybook style characters. Distances are calculated between each alphabetic character and all the other styles of the same alphabetic character. For example, each uppercase A is compared against all other uppercase A’s. Clusters are then formed based on the calculated distances. Figure 4a shows an example of the dendrogram formed from the uppercase A’s using agglomerative hierarchical clustering with a single linkage [4].

Based on the dendrogram in Figure 4a, we reorganized the data based on a human visual perspective and found five groups of capital A’s as shown in Figure 4b.

![Dendrogram formed for the uppercase A’s.](image1)

![Clusters resulting from the uppercase A copybook style.](image2)

Observing the copybook style clusters for uppercase A, we see that the North and South American versions are written like the lowercase a and most of the South American versions are rather vertical in contrast to the Canadian and USA styles which are more slanted. The other copybook styles are written more in the manuscript style of capital A’s with crossbars varying as straight, curved, or looped.

Examining the Canada1 (cursive) and USA2 (cursive) on the top row of Figure 4b, we see two essentially indistinguishable letters which could result in incorrect copybook style identification. A possible extension of this study that might help alleviate this problem is mentioned in the following section.

5. Conclusions and Future Work

In this paper, we presented two computer assisted handwriting copybook style analysis techniques – a method of identifying the copybook style of a questioned document, and a clustering method that groups similar characters of the copybook styles. In contrast to the earlier study that used image matching [2], here we used pseudo online data to improve the matching. Although only 19 Roman alphabet copybook styles were studied, the results were promising.

This feasibility study presents an approach that can reduce the suspect population of a questioned document. First, an online database was created containing all the copybook style letters. Then each of the letters of a questioned document was compared against the letters in the database using the modified
Levenshtein edit distance. The comparison that gave the least overall distance was considered to be the copybook style of the questioned document. In addition, cluster analysis was performed on all the letters of the database to determine those copybook styles that were most similar using specific characteristics, in this case the SDSS and edit distance matches of each letter.

There are numerous different Roman alphabet copybook styles taught throughout the world. Collecting and incorporating most of these copybook styles into the database is necessary for completeness and further analysis. An extension of this study would be to categorize the copybook styles based on specific characteristics and allow for the identification of the cluster to which a copybook style of a questioned document belongs. Also, other dynamic characteristics, such as velocity or SPSS (stroke pressure sequence string), and other distance metrics could be investigated.

References