

Similarity-Based Handwriting Copybook Style Identification

Sungsoo Yoon, Sung-Hyuk Cha, and Charles C. Tappert
Computer Science Department, Pace University
861 Bedford Road, Pleasantville, New York, 10570
{scha, ctappert}@pace.edu

Abstract *Handwriting originates from a particular copybook style such as Palmer or Zaner-Bloser that one learns in childhood. Since questioned document examination plays an important investigative and forensic role in many types of crime, it is important to develop a system that helps objectively identify a questioned document's handwriting style. Here, we propose a computer vision system that can assist a document examiner in the identification of a writer's handwriting style and therefore the origin or nationality of an unknown writer of a questioned document. We collected 33 Roman alphabet copybook styles from 18 countries. Each character in a questioned document is segmented and matched against all of the 33 handwriting copybook styles. The more characters present in the questioned document, the higher the accuracy observed.*

Key Words: *Handwriting Analysis, handwriting style*

1. Introduction

Questioned document examinations play an important investigative and forensic role in many types of crime [1, 2], and the analysis of handwritten documents has great bearing on the criminal justice system. Various automatic writer identification computer techniques, feature extraction, comparison, and performance evaluation methods have been studied (see [3, 4] for an extensive survey). Recent studies on the analysis of handwritten documents have focused on the individuality of handwriting, i.e., determining the individual writer [5-7]. Although a statistical measure of individuality was proposed in the earlier studies, it has limited capability to identify an individual writer from a large population like the two hundred million writers in the United States or the entire world population. Hence, we propose a system that identifies the copybook style of the writer rather than the individual writer.

Imagine a system that could identify the handwriting style, and therefore the origin or nationality, of an unknown writer of a questioned document. Such an intelli-

gent system could help document examiners and investigators identify the criminal, or at least reduce the suspect list, in many crime cases involving handwriting. Building and validating such an automated intelligent handwriting style identification system is a challenging computer vision application. In order to achieve this goal, a large handwriting copybook style image database must be constructed and good pattern matching algorithms must be developed and validated. Hence, we describe our ongoing handwriting style image database efforts and present a handwriting copybook style identification system.

Handwriting is usually learned by copying a formal system. Although changing somewhat over time, one's handwriting style typically originates from a particular copybook style such as *Palmer* or *Zaner-Bloser* that one learns in childhood. The problem of categorizing handwriting styles has been considered widely. Earlier approaches used neural network or hierarchical clustering techniques to find a set of handwriting families or clusters to improve the handwriting recognition performance and to make several important findings on handwriting styles [8-11]. The term "handwriting style" has been used in the literature to mean different things. Throughout this paper, however, handwriting style means a kind of copybook system and not the particular style of an individual.

There are three phases for the handwriting copybook style identification system implementation: *i)* handwriting copybook style image database construction, *ii)* feature extraction, and *iii)* similarity-based pattern matching. We collected 33 Roman alphabet copybook styles, five manuscript styles and 28 cursive styles, from 18 countries. Each character in a questioned document is segmented and matched against all of the handwriting styles.

This paper is organized as follows. Section 2 discusses the construction of the handwriting copybook style database. Section 3 discusses the features extracted and the similarity-based pattern matching algorithm. Section 4 presents the experimental results and section 5 draws conclusions of this work.

2. Handwriting Copybook Style Database

We collected 33 Roman alphabet copybook styles from 18 countries: Australia (2 manuscript styles), Austria (3 cursive styles), Belgium (2 cursive), Brazil (1 cursive), Canada (1 cursive), Chile (1 cursive), Columbia (1 cursive), Denmark (1 cursive), Ecuador (1 cursive), England (1 cursive), Germany (3 cursives), Netherlands (1 cursive), Norway (1 cursive), Peru (2 cursive), Sweden (2 cursive), Switzerland (3 cursive), United States (3 cursive and 3 manuscript), and Uruguay (1 cursive). There are a total of five manuscript styles and 28 cursive styles. These copybook images were obtained from various books and websites. Figure 1 shows 18 styles for the word ‘beheaded’ by using a synthesizing technique.



Figure 1. Some copybook handwriting styles for the word ‘beheaded.’

Currently, 20 of these styles have been segmented, and every character image, A-Z and a-z, stored in a database. The remaining copybook images are still undergoing the tedious restoration process due to the low resolution or the line removal process. Some writing systems omit a few characters – for example, ‘k’ is not in Peru’s alphabet system because ‘k’ is ‘ch’ in that system. The names for each character image have five fields (HSxx_y_zzz_vw). The field ‘xx’ is the copybook identification number, the field ‘y’ indicates whether the letter is cursive or manuscript, and the field ‘zzz’ contains the ascii number for the alphabet character. Finally, the fields ‘v’ and ‘w’ are flags for the existences of decorating head and trailing tail. For example, the name, ‘HS04_c_065_00’ indicates the cursive Austria copybook style Capital letter ‘A’ without any decorating head or trailing tail.

In order to test our system, we also collected handwritten documents whose origin is known. These include

letters from books, which claim to be from England, Norway, and Switzerland, and handwritten document samples from subjects from England, Peru, etc.

3. Copybook Style Identification Algorithm

We consider a handwriting copybook style identification problem. Given a handwriting copy style image database and a questioned handwritten document, the style identification problem can be solved by retrieving similar character images from the database for each character of the questioned document and voting the retrieved results. This method can be formalized by extracting appropriate features from the characters, defining a distance metric between a questioned character image and a character in the copybook database, and developing a voting algorithm.

3.1. Computing Character Level Features

Many computational features appear in the literature [12, 13]. Although various features have been incorporated into our system as options, we report here on the use of *gradient* features. Gradient directions are computed by the *Sobel edge detection mask* operators in the following equation [13].

$$\text{direction} = \tan^{-1} \frac{S_y(i, j)}{S_x(i, j)}$$

A sample of the gradient direction maps of a character image is shown in Figure 2 (a).

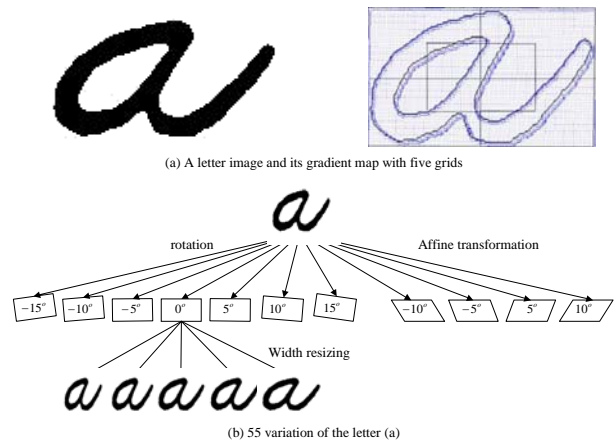


Figure 2. Gradient direction feature extraction.

It is well known that no two instances of the same handwritten letter look exactly alike. Therefore, in order to minimize the variability of a particular copybook style letter, we applied six slight rotations and four *affine* transformation operations to produce ten differently shaped letter images. To each of these modified letter images and

to the original image we applied five width-resizing operators to vary the width of the letter. As a result, a total of 55 varying images are produced for each copybook style letter. This process is depicted in Figure 2 (b).

Each of the 55 images is divided into five grids as shown in Figure 2 (a). For each grid, the frequency of gradient direction is counted. Gradient directions are quantized into eight direction ranges. As a result, a copybook style letter is represented by a two-dimensional feature vector (55x40).

3.2. Similarity-Based Pattern Matching

The similarity-based pattern-matching phase compares the gradient feature vector of each letter in the questioned document to the library of gradient feature vectors of the entire copybook style letter database. Let's denote the entire copybook style gradient feature vector as $D_{i,j}$ where i and j are indices for the m copybook styles and the 52 upper and lowercase letters, respectively. Let's denote a questioned handwritten document as Q_k where k is the index of n segmented characters. Then the problem is formalized in Figure 3. As an illustrative example, an input and corresponding output are shown in Figures 4 and 5, respectively.

The matching procedure is straightforward. Each upper or lowercase letter of the questioned document is compared to its corresponding letter in all the copybook styles. For example, the letter *W* in Figure 4 is only compared to all capital *W*s in the copybook style dictionary. For the letter comparison we use the *Euclidean* distance between two gradient direction vectors. Because a dictionary letter is represented by 55 vectors to allow some variability, the distance between a questioned document letter and one from a dictionary is the minimum distance

between the questioned letter and all of the 55 dictionary vectors, i.e.: $\text{distance} = \min_{l=1..55} (\text{euclidean distance}(D_{i,j} - Q_k))$.

Input: $D_{i,j}$ and Q_k
 where $i = 1 \dots m, j = 1 \dots 52, k = 1 \dots n$
 Output: a sorted list of i 's with distances.

Procedure:

```

for  $k = 1 \dots n$ 
  initialize Dmatrix[ $m,n$ ] = 0;
  if letter( $j$ ) == letter( $k$ )
    for  $i = 1 \dots m$ 
      Dmatrix[ $i,k$ ] = distance( $D_{i,j}, Q_k$ );
    end
  end
end
return sort with index(rowavg(Dmatrix));
    
```

Figure 3. Problem formalization and pseudo code.



Figure 4. A sample questioned handwriting.

The final output is a distance matrix as shown in Figure 5. The final decision is based on the average of the distances between the letters of the questioned document and those of each copybook style. For example, the document from Switzerland in Figure 4 can be classified appropriately by handwriting copybook style matching.

	W	i	r	s	c	h	r	e	i	b	e	n	
Switzerland1	23.4585	101.7896	89.2399	42.3244	61.4163	31.9927	151.9341	95.0172	49.3324	59.8047	77.557	57.9531	70.151658
Norway1	60.6239	77.1794	107.3982	85.9694	60.1095	70.7302	109.4171	107.6705	36.0841	77.1245	99.8826	59.0032	79.26605
Switzerland3	36.4256	103.6153	42.654	101.7364	90.5359	54.6536	147.8869	98.5662	59.4333	75.519	85.1064	72.5456	80.723183
German3	79.5436	81.8939	50.6919	84.3473	98.9446	73.0338	113.0648	112.4088	73.7017	51.858	98.2385	90.0824	83.984108
England1	75.5865	78.7899	75.0105	95.0893	97.1827	58.9992	88.3048	130.9773	77.7816	66.4929	120.1943	65.9985	85.867292
Denmark1		104.8951	67.9394	91.6346	83.6516	70.6963	129.7213	126.2796	48.2262	79.7733	121.5205	45.9902	88.211645
Peru1		92.6681	130.5025	53.0946	102.5164	59.4201	139.0174	125.7199	43.994	71.3015	102.8196	65.4165	89.679145
Belgium1		92.7983	131.5697	44.1003	101.2601	79.8797	145.6377	98.9261	48.9886	80.3063	125.4184	44.6155	90.318245
Austria2	61.86	131.1067	85.4835	109.2796	65.3644	94.0815	186.1683	115.763	62.4379	115.0544	104.2668	83.4495	101.19297
Columbia1	85.0178	78.124	146.1169	136.5091	95.0121	71.8073	173.6923	101.4297	98.1851	53.5495	87.4724	97.5557	102.03933
Brazil1	126.8338	111.2759	126.5907	103.6344	69.2081	90.5066	179.1611	120.5368	86.2276	101.9792	96.6764	133.1248	112.14628
USA1	44.8082	146.4811	120.6671	86.8788	109.4353	103.0092	176.6708	144.5657	70.6627	122.5055	142.2026	84.6051	112.70768
Ecuador1	132.4394	175.5389	130.6275	82.1152	77.5599	95.473	181.3675	95.02	119.0691	121.4527	87.4235	113.9842	117.67258
Canada1	43.4166	178.5015	113.1581	98.5721	112.4954	111.0579	179.8836	146.2421	94.2447	119.6072	145.428	129.241	122.65402
USA2	43.2158	180.054	119.1868	100.0869	114.6572	118.9346	183.5502	147.1417	95.0501	121.3252	149.9143	127.7194	125.06968
Chille1	137.8739	164.6059	124.6229	114.4116	99.6742	124.4464	202.8162	154.2122	79.0622	112.3633	156.9452	122.7928	132.8189
Netherlands1	100.9746	158.4266	144.2128	197.23	74.6089	78.386	214.2917	111.3614	172.9847	121.8061	108.3235	123.3629	133.83077
Peru2	126.3819	173.4229	168.7866	142.423	69.6703	161.5765	227.8518	119.112	99.0176	184.6229	121.7613	128.0816	143.55903

Figure 5. a sample output for Figure 4.

4. Experiments

In this section we present preliminary experimental results of the copybook style identification system. For test data we obtained ten paragraphs of known origin (Austria, England, Norway, and Switzerland) from books and Internet sources (see example in Figure 4). The number of letters in the questioned documents varied from 5 to 20.

According to our algorithm the identification of the source country is considered correct if the average of distances between the origin of a questioned document and that copybook style is the smallest. However, not all questioned documents were identified correctly on the top choice. Figure 6 shows the rank versus accuracy graph, and accuracy reached 100% when the top 10 choices are accepted. Copybook systems from England appear to be the ones most confused with other copybook systems.

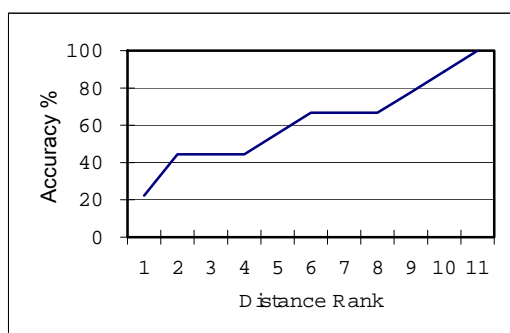


Figure 6. Rank vs. Accuracy graph.

In another illustrative example Figure 7 shows four letters (Austrian 'T', English 'M', Norwegian 'v', and Swiss 'W') with the top 15 retrieved copybook style letters and their corresponding distance scores. The system correctly identified the writing style of three of these letters and the Norwegian 'M' was identified correctly on the 4th choice.

5. Conclusion

In this paper we presented a similarity-based copybook style identification system using a collection of 33 copybook styles. Gradient direction features were extracted and a similarity-based matching algorithm was presented. Although only ten questioned documents were tested, the preliminary results are promising. From the experimental results we observe that certain letters have higher discriminating power than others. Also, from detailed examinations of the images in Figure 7 there appear to be similarities between letter shapes and geographical regions, such as North America, South America, etc.

Although we collected 33 copybook styles, there are many more Roman alphabet copybook styles throughout the world. We designed our database system so that any other copybook style can be easily appended to our database. Collecting a larger inventory of copybook styles remains as future work. Although we tested several feature sets and the gradient direction features give good results, designing better feature sets and matching algorithms are also open problems.

Acknowledgements

This work was jointly supported by the Post-doctoral Fellowship Program of Korea Science & Engineering Foundation (KOSEF) and the Research Grant Program of CSIS, Pace University.

Reference

- [1] Roy A. Huber and Alfred M. Headrick, "Handwriting Identification: Facts and Fundamentals", CRC Press, 1999
- [2] Russell R. Bradford and Ralph B. Bradford, "Introduction to Handwriting Examination and Identification", Chicago: Nelson-Hall Publishers, 1992
- [3] Rejean Plamondon and Guy Lorette, "Automatic signature verification and writer identification – the state of the art", *Pattern Recognition* 22(2):107-131, 1989.
- [4] Rejean Plamondon and Sargur N. Srihari, "On-Line and Off-Line Handwriting Recognition: a Comprehensive Survey", *TPAMI* 22(1):63-84, 2000.
- [5] Sargur N. Srihari, Sung-Hyuk Cha, Hina Arona, and Sang-jik Lee, "Establishing Handwriting Individuality using Pattern Recognition Techniques", in *Proceedings of 6th ICDAR*, Seattle, IEEE, 2001, pp 1195-1204.
- [6] Sung-Hyuk Cha and Sargur N. Srihari, "Writer Identification: Statistical Analysis and Dichotomizer". in *Proceedings of SPR & SSPR 2000*, Alicante, LNCS-Advances in Pattern Recognition, vol 1876 pp 123-132.
- [7] Sung-Hyuk Cha and Sargur N. Srihari, "Assessing the Authorship Confidence of Handwritten Items". in *Proceedings of 5th WACV 2000*, Palm Springs, IEEE, pp 42-47.
- [8] Jean-Pierre Crettez, "A set of handwriting families: style recognition," in *Proceedings of ICDAR 1995*, IEEE, pp 489-494.
- [9] Bote-Lorenzo, M. L., Dimitriadis, Y. A. & Gómez-Sánchez, E., "Allograph extraction of isolated handwritten characters", in *Proceedings of the 10th Biennial Conference of the International Graphonomics Society*. Nijmegen, 2001
- [10] Schomaker, L., Abbink, L. & Selen, S., "Writer and writing-style classification in the recognition of on-line handwriting." In *Proceedings of the European Workshop on Handwriting Analysis & Pattern Recognition*, London: The Institution of Electrical Engineers, 1994
- [11] Vuurpijl, L. & Schomaker, L., "Finding structure in diversity: a hierarchical clustering method for the categorization of allographs in handwriting." In *Proceedings of the 4th ICDAR*, 387-393. Piscataway, NJ: IEEE, 1997

[12] Oivind Due Trier, Anil K. Jain, and Torfinn Taxt, "Feature Extraction Methods for Character Recognition", Pattern Recognition 29(4):641-662, 1996

[13] S. E. Umbaugh, "Computer Vision and Image Processing", Prentice Hall PRT, 1998

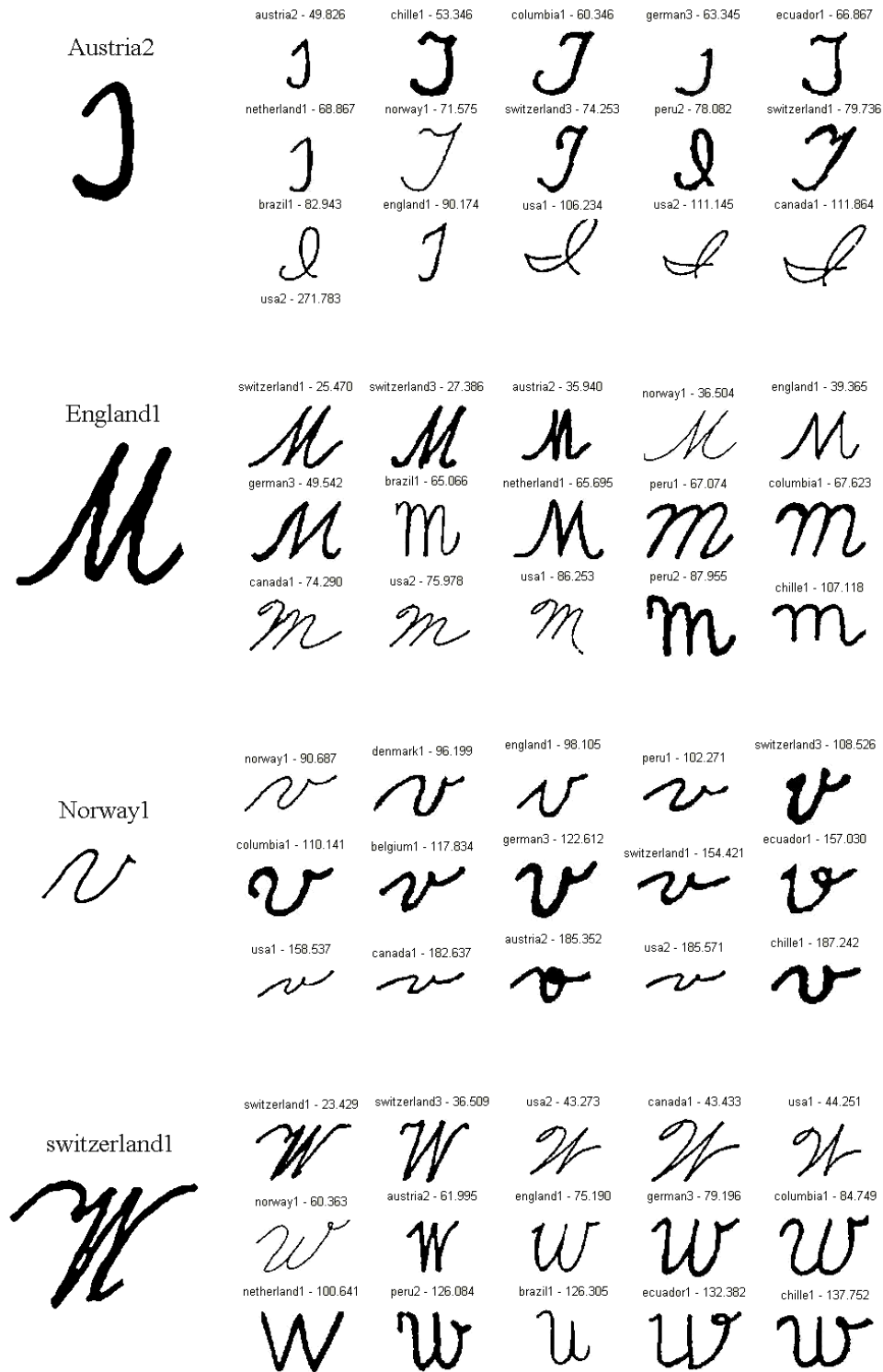


Figure 7. Questioned letters and retrieved copybook letters with distances.

