Handwriting Copybook Style Identification for Questioned Document Examination

Sung-Hyuk Cha, Sungsoo Yoon, and Charles C. Tappert

School of Computer Science and Information Systems, Pace University
861 Bedford Rd. Pleasantville, NY 10570 USA

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 English 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.

Keywords: Copybook, Handwriting Analysis, handwriting style identification

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 intelligent 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, in this paper 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 English alphabet copybook styles, five manuscript styles and 28 cursive styles, from 18 countries. Each character in a questioned document is manually 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 and feature representation. Section 3 outlines the similarity-based pattern matching algorithm. Section 4 presents some experimental results and section 5 draws conclusions of this work.

2 DATABASE OF HANDWRITING COPYBOOK STYLES

We collected 33 English 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’.

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

Currently, 20 copybook handwriting 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.

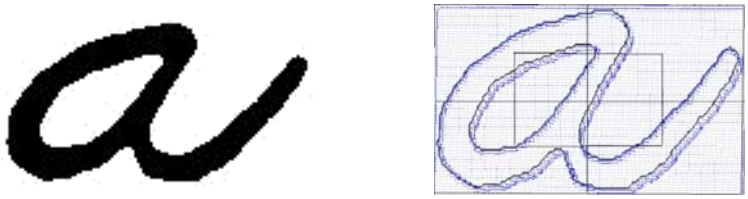
2.1 CHARACTER-LEVEL FEATURE REPRESENTATION

Many computational features appear in the literature [12, 13] mainly to solve the problem of optical character recognition. In this paper we report only simple *gradient* (slope) features because

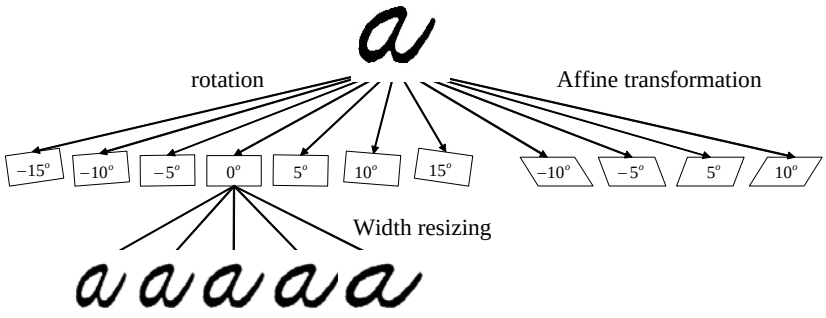
our main intent is to show the feasibility of the process. Gradient directions are computed by convolving the *Sobel edge detection mask* operators, in the following equation [13], on a character image.

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

This operator provides the approximate *x* and *y* derivatives, i.e., magnitude and direction of each pixel on a given character image. A sample of the gradient direction maps of a character image is shown in Figure 2 (a).



(a) A letter image and its gradient map with five grids



(b) 55 variation of the letter (a)

Figure 2. Gradient direction feature extraction.

It is well known that no two instances of the same handwritten letter look exactly alike. Therefore, to allow some 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 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) and stored in the database.

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 each character image, defining a distance metric between a questioned character image and a character in the copybook database, and developing a voting algorithm.

3.1 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 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 pseudo code as shown in Figure 3. As an illustrative example, an input and corresponding output are shown in Figures 4 and 5, respectively.

Input: $D_{i,j}$ and Q_k
where $i = 1 \ldots m,\ j = 1 \ldots 52, k = 1 \ldots n$

Output: a sorted list of *i*'s with distances.

Procedure:

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

Figure 3. Problem formalization and pseudo code.

Figure 4. A sample questioned handwriting.

We use a straightforward procedure. An input, Q_k, consists of *k* characters in a questioned document, e.g., the 12 letters in the sample in Figure 4. A letter, '*W*', in a questioned document is only compared to all capital '*W*' in the copybook style dictionary.

To compare a letter from a questioned document and one from a dictionary requires a distance metric. We use a simple *Euclidean* distance between two gradient direction vectors. Note that a letter from a dictionary is represented by 55 vectors allowing some variability. Hence, a distance between a letter from a questioned document and one from a dictionary is the minimum distance between a questioned letter and all 55 vectors as in the following equation.

$$\text{distance} = \min_{l=1 \cdots 55} (\text{euclideandistance}(D_{i,j,l} - Q_k))$$

Finally, the output is a distance matrix as shown in Figure 5. Sums of every distance per letter are counted for each copybook style. A handwriting sample from Switzerland, shown in Figure 4, can be classified as belonging to a Switzerland copybook style by copybook style matching.

	W	i	r	s	c	h	r	e	i	b	e	n	average
austria2_C	61.86	131.1067	85.4835	109.2796	65.3644	94.0815	186.1683	115.763	62.4379	115.0544	104.2668	83.4495	101.193
belgium1_C		92.7983	131.5697	44.1003	101.2601	79.8797	145.6377	98.9261	48.9886	80.3063	125.4184	44.6155	90.31825
brazil1_C	126.8338	111.2759	126.5907	103.6344	69.2081	90.5066	179.1611	120.5368	86.2276	101.9792	96.6764	133.1248	112.1463
canada1_C	43.4166	178.5015	113.1581	98.5721	112.4954	111.0579	179.8836	146.2421	94.2447	119.6072	145.428	129.241	122.654
chille1_C	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
columbia1_C	85.0178	78.124	146.1169	136.5091	95.0121	71.8073	173.6923	101.4297	98.1851	53.5495	87.4724	97.5557	102.0393
denmark1_C		104.8951	67.9394	91.6346	83.6516	70.6963	129.7213	126.2796	48.2262	79.7733	121.5205	45.9902	88.21165
ecuador1_C	132.4394	175.5389	130.6275	82.1152	77.5599	95.473	181.3675	95.02	119.0691	121.4527	87.4235	113.9842	117.6726
england1_C	75.5865	78.7899	75.0105	95.0893	97.1827	58.9992	88.3048	130.9773	77.7816	66.4929	120.1943	65.9985	85.86729
german3_C	79.5436	81.8939	50.6919	84.3473	98.9446	73.0338	113.0648	112.4088	73.7017	51.858	98.2385	90.0824	83.98411
netherland1_C	100.9746	158.4266	144.2128	197.23	74.6089	78.386	214.2917	111.3614	172.9847	121.8061	108.3235	123.3629	133.8308
norway1_C	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
peru1_C		92.6681	130.5025	53.0946	102.5164	59.4201	139.0174	125.7199	43.994	71.3015	102.8196	65.4165	89.67915
peru2_C	126.3819	173.4229	168.7866	142.423	69.6703	161.5765	227.8518	119.112	99.0176	184.6229	121.7613	128.0816	143.559
switzerland1_C	23.4585	101.7896	89.2399	42.3244	61.4163	31.9927	151.9341	95.0172	49.3324	59.8047	77.557	57.9531	70.15166
switzerland3_C	36.4256	103.6153	42.654	101.7364	90.5359	54.6536	147.8869	98.5662	59.4333	75.519	85.1064	72.5456	80.72318
usa1_C	44.8082	146.4811	120.6671	86.8788	109.4353	103.0092	176.6708	144.5657	70.6627	122.5055	142.2026	84.6051	112.7077
usa2_C	43.2158	180.054	119.1868	100.0869	114.6572	118.9346	183.5502	147.1417	95.0501	121.3252	149.9143	127.7194	125.0697

Figure 5. A sample output corresponding to the handwriting in Figure 4. Some styles had no '*W*'.

3.1 GRAPHICAL USER INTERFACE

Our handwriting copybook style identification prototype system provides the following GUIs (graphical user interfaces). First, a questioned document is loaded into our proposed graphical user interface as shown in Figure 6. Second, the system allows a user to crop a letter of interest using the mouse to draw an enclosing polygon with a mouse or an electronic pen device as shown in Figure 6 (a), and the system displays the resulting cropped letter as in Figure 6 (b). Third, a user enters the identity (truth) of the letter (A-Z, a-z), and the system retrieves all similar styles of the letter, ordered with their distance values to the questioned letter as shown in Figure 6 (c). The user, such as a document examiner, repeats the previous steps for several letters of interest. Finally, when the user presses the finish button, the system shows the individual letter rankings and the average rank for all examined letters as shown in Figure 6 (d). For the example of Figure 6, the letters 'W', 'I', and 'r'

were examined and the system identifies the questioned-document handwriting style as a copybook style from Switzerland.

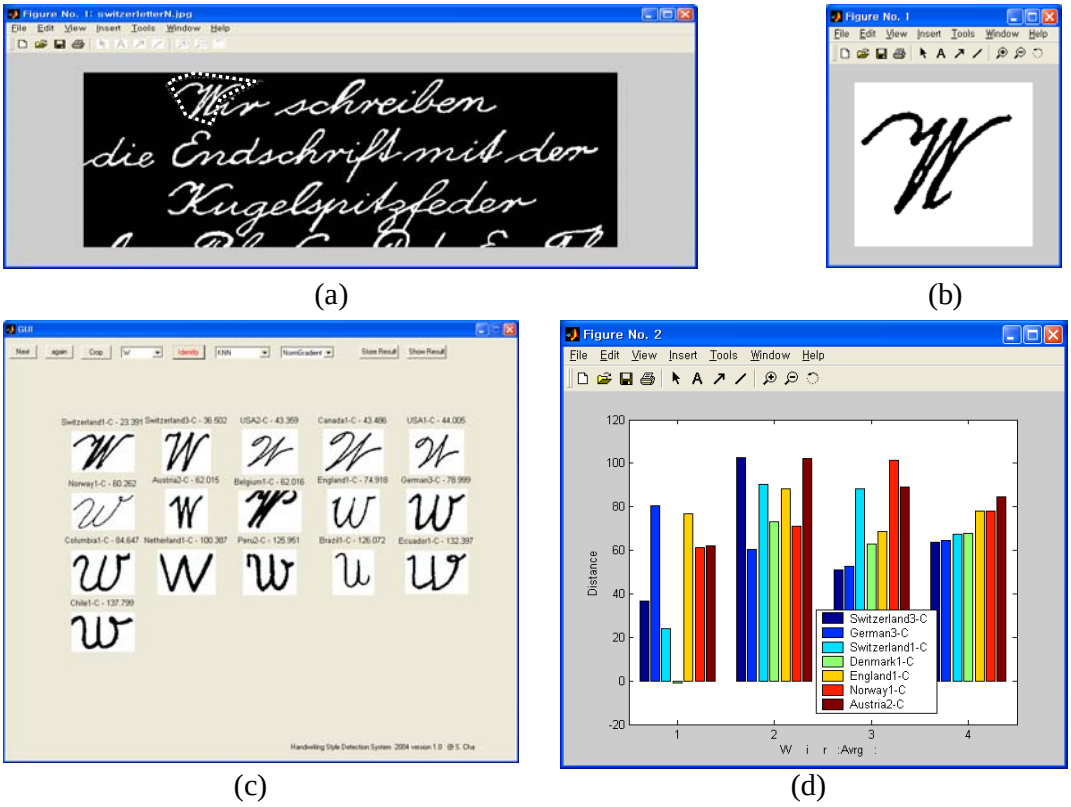


Figure 6. GUI for Computer Assisted Copybook Style Identifier.

Another feature of our prototype system allows users to select similar characters according to human vision. Figure 7 illustrates the results of the voting process on a questioned document known to be from a writer native to England. Here, for each letter of the sample document the writing systems having a similar letter are enumerated. For example, seven writing systems have a letter ‘M’ similar to the first letter in the questioned document. One can compare it with a side-by-side or superimposed view. By counting the frequency of each copybook style, we find that England copybook style has the highest voting count among all the writing systems in our database. In this brief example, each of the nine letters of the questioned document matched letters from English

handwriting styles. Hence, it is likely that the writer of the questioned document has an English writing style.

Figure 7. Output of handwriting style identification system.

4 EXPERIMENTS

In this section, we present the experimental results of our copybook style identification system. To test our system it is essential to collect handwritten documents whose style or origin is known. A total of 15 questioned documents of known origin (Austria, England, Norway, Peru, and Switzerland) were obtained from books, the Internet, and students. The number of letters in these questioned documents varied from 6 to 32.

The system is considered accurate if the average of distances between letters in a questioned document and those in the respective copybook style is the smallest. Only four documents were correctly identified. Handwritten documents from England were most confused with copybook

styles of other countries. As shown in Figure 7, alphabet letters (a, d, e, h, y) have almost no contribution to identification. A correlation-based feature selection algorithm using exhaustive search found the top ten alphabet letters containing the most information for the copybook style from England to be *D, E, F, H, R, S, j, k, l,* and *n*. Top ten alphabet letters for the Norway copybook style were *E, G, L, N, T, U, k, m, r, t* and those for the Switzerland style were *D, G, L , T, j, m, r, z*.

Figure 8 shows four letters (‘*I*’ from a style of Austria, ‘*M*’ from one of England, ‘*v*’ from one of Norway, and ‘*W*’ from one of Switzerland) with the top 15 retrieved letters from the copybooks. Except for the ‘*M*’, our system correctly identified the writing style by their shape similarity.

5 CONCLUSIONS

In this paper, we presented a similarity-based copybook style identification system using 33 copybook styles. Gradient direction features were extracted and a similarity-based matching algorithm was presented. Although only 15 questioned documents were tested, the results were promising, indicating that the described procedure can narrow the search of handwriting styles to help identify the writer of a questioned document. We also observed from the experiments that certain letters have higher discriminating power than others. Finally, an examination of the images in Figure 8 reveals a similarity of letter shapes from 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 other copybook styles can be easily appended to our database. Collecting a larger inventory of copybook styles remains as a future work. Designing better feature sets and matching algorithms are also open problems. Furthermore, collecting more testing questioned documents of known origin is essential. Since the type of pen or

quality of printing may influence the process, careful experimental testing data is also an important future work.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Roy A. Huber and Alfred M. Headrick, "Handwriting Identification: Facts and Fundamentals", CRC Press LLC, 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 Survery", TPAMI 22(1):63-84, 2000.

[5] Sargur N. Srihari, Sung-Hyuk Cha, Hina Arona, and Sangjik Lee, "Establishing Handwriting Individuality using Patten Recognition Techniques", in Proceddings of 6th ICDAR '01, Seattle, IEEE Computer Society, 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 Proceddings of 5th WACV 2000, Palm Springs, IEEE Computer Society, 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 isolatedhandwritten 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 Computer Society, 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

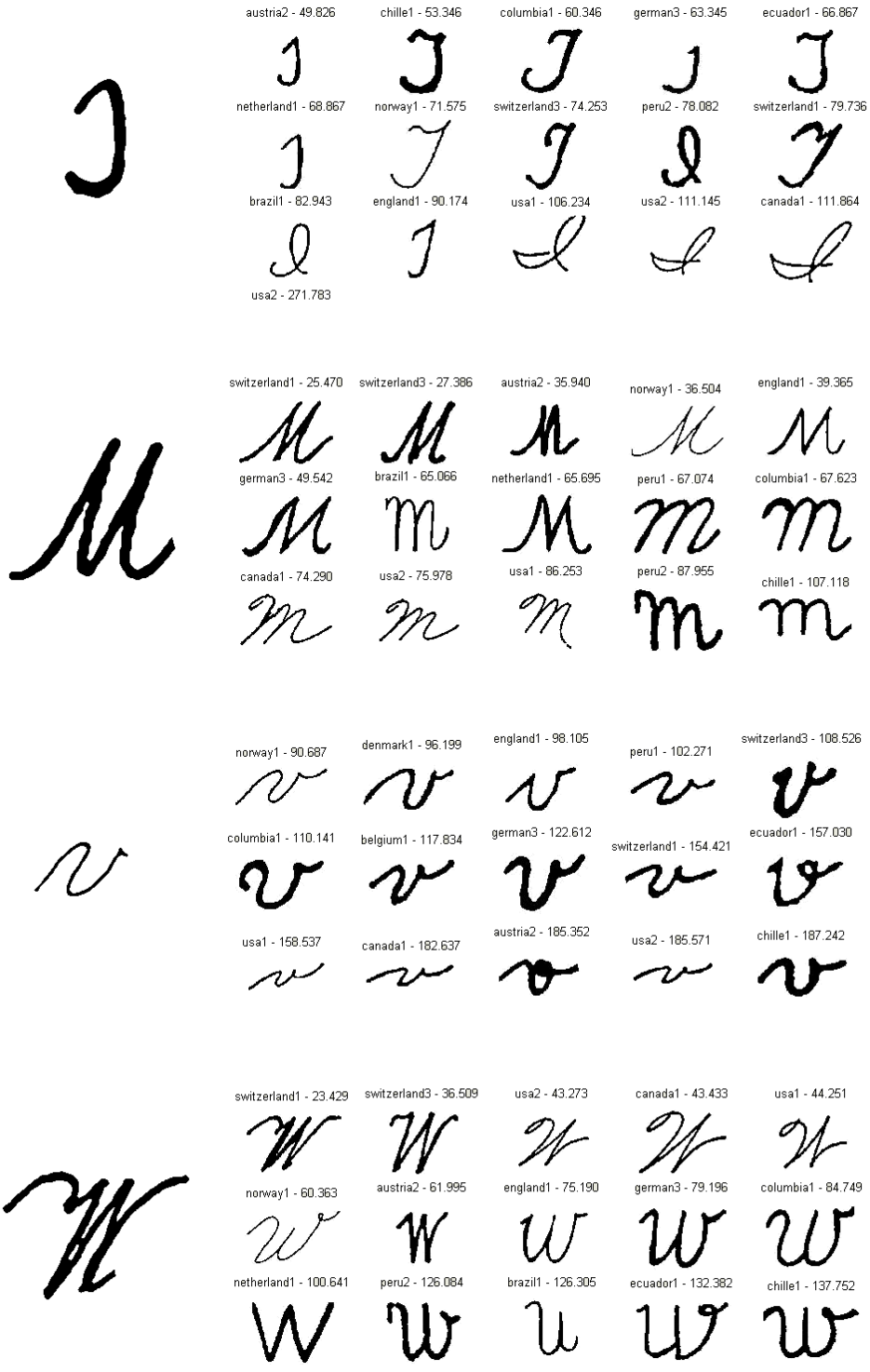


Figure 8. Questioned letters and retreieved copybook letters with distances.