

# Writer Identification using Innovative Binarised Features of Handwritten Numerals

Graham Leedham<sup>1</sup> and Sumit Chachra<sup>2</sup>

<sup>1</sup>*School of Computer Engineering  
Nanyang Technological University, Singapore 639798*

<sup>2</sup>*Dept. of Electrical Engineering  
Indian Institute of Technology (IIT), Roorkee-247667, India*

*asgledham@ntu.edu.sg      sumitchachra@rediffmail.com*

## Abstract

*The objective of this paper is to present a number of features that can be extracted from handwritten digits and used for author verification or identification of a person's handwriting. The features under consideration are mainly computational features some of which cannot be easily evaluated by humans. On the other hand, these features can be extracted by computer algorithms with a high degree of accuracy.*

*The eleven features used are described. All features were appropriately binarized so that binary feature vectors of constant lengths could be formed. These vectors were then used for author discrimination, using the Hamming distance measure. For this task a writer database consisting of 15 writers was created. Each writer was asked to write random strings of 0 to 9 at least 10 times. The results indicate that the combined features work well at discriminating writers and warrant further detailed investigation.*

*Although the set of features was designed for dealing with handwritten digits (as may be written on cheques), it may also be used for isolated alphabetic characters.*

## 1. Introduction

Traditionally in legal cases expert document examiners are consulted to present their expert opinion in courts of law to determine the authenticity, or otherwise, of a document. Using various measuring strategies and techniques [1] they try to prove or disprove the claimed authorship of a questioned document. They make use of both qualitative and quantitative features to draw a conclusion [5].

In recent court cases that have taken place in the United States (e.g. the case of Daubert et al v. Merrell Dow Pharmaceuticals) the scientific validity of their analysis has been questioned. There is, in fact, very little scientific proof to support their analysis.

The aim of this research is to find information about a person's handwriting, which makes it unique or at least identifiable from other people's handwriting. Not much work has been carried out specifically for finding out discrimination data for handwritten digits.

As a person's handwriting tends to change under the influence of different factors, there exists an intra-author variation in handwriting. If two individual's handwritings are distinguishable, the intra-author variation is less than the inter-author variation. By extracting features and expressing them in a quantitative form we wish to discover if it is possible to evaluate these variations and distinguish an individual writer.

The document examiner features form a set of 21 discriminating elements of handwriting [2]. The 21 features are: Arrangement, Class of allograph / alphabet, Connections, Design of allographs and their construction, Dimensions (Horizontal and Vertical), Slant and slope, Inter- and intra-word spacing, Abbreviations, Baseline alignment, Initial and terminal strokes, Punctuations, Embellishments, Legibility or writing Quality, Line continuity, Line quality, Pen control, Writing movement, Natural variations or consistency, Persistency, Lateral expansion and Word proportions.

Some of these features could be very subjective in terms of their validity and use. For example a person may find a handwriting legible and of high quality while another might find it inferior or of medium quality.

In this paper we seek to implement and investigate some features which can be extracted reliably using computational techniques.

## 2. Computational Features Extracted

We have extracted 11 features (F1-F11) from each isolated digit. We give a brief description of each feature below.

- a) **F1 - Height to Width ratio (or Aspect ratio)** – This is simply calculated as H/W.
- b) **F2 - Number of End-Points** – This is the number of pointed ends that are present in the digit (see Fig. 1). This is found by checking all the black pixels that do not have more than one 8-connected neighbours. This operation is performed on a thinned version of the original image.
- c) **F3 - Number of Junctions** – This is the number of junctions that are present in the digit (see Fig. 1). A junction is considered to be a point (black pixel) that has a minimum of three 8-connected neighbours that are black. A junction is formed if two lines simply intersect. This operation is also performed on a thinned image.
- d) **F4 - Number of Loops**
  - **F4-1 - Degree of Roundness** – For a perfect circle the degree of roundness is zero [5].
  - **F4-2 - Area\*** - Area of a loop is defined as the number of pixels (white) contained inside it (see Fig. 2(b)).
  - **F4-3 - Loop Length\*** – The loop length is found by first detecting the contours via Stentiford templates and then calculating the number of black pixels left behind in the image (see Fig. 2(c)).
  - **F4-4 - Slant** – This can give the angle at which the loop is inclined. It can be estimated in many ways. In the present work it is calculated by dividing a loop into two parts with respect to its height. After this the centres of gravity of the two halves are found. The line connecting these gravity centers is considered to represent the loop angle [5].
  - **F4-5 - Loop Fissure Length** – This feature is calculated in the case of open or incomplete loops. This is the total number of missing pixels in a loop. The feature value is normalized by dividing the number of missing pixels by the loop length calculated in feature F4-3. Normalization makes the feature value independent of the character or digit size.
- e) **F5 - Slant\*** – This can represent an approximate slant of a character or a digit. Details of a slant measurement depend on the character or digit under consideration. For example, in the case of the digit “6” slant can be taken as the tangent of an angle made by the east most point and the end point (upper end

point in case of more than one end point). Similar rules can be developed for other digits and characters. Calculating this feature explicitly for each digit or character can be very tedious and the results may not be accurate. Gradient features (F11), mentioned below, are also able to represent slant.

- f) **F6 - Zero Crossings** – For this feature the digit height is divided into the fixed number of zones (say 5 or 8) and then the zero crossing value is calculated along each of the 5 or 8 lines in the horizontal direction. Similarly this can be done for the vertical direction as well (along 4 or 8 zones). Zero crossing refers to the number of transitions from white to black or from black to white in a digit image.
- g) **F7 - Width / Height Distribution** – This is the change in the width or height of the character as a digit image is traversed across in horizontal and vertical direction along the lines in F6.
- h) **F8 - Pixel Density** – This feature is calculated by dividing bounding box of the digit into 9 equal rectangles of equal area. The number of black pixels in each box is then found and divided by the rectangle area.
- i) **F9 - Fixed Point distance and angular measure** – This feature is calculated by partitioning an image into a number of boxes. A 6x4 partitioning was chosen. The fixed point was chosen to be the first pixel (mostly white) of the bounding box of the digit. From this pixel the distance of each black pixel in all the 24 boxes was calculated and then normalized by means of dividing it by the number of black pixels within the box.

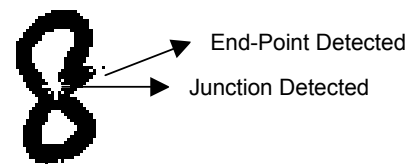
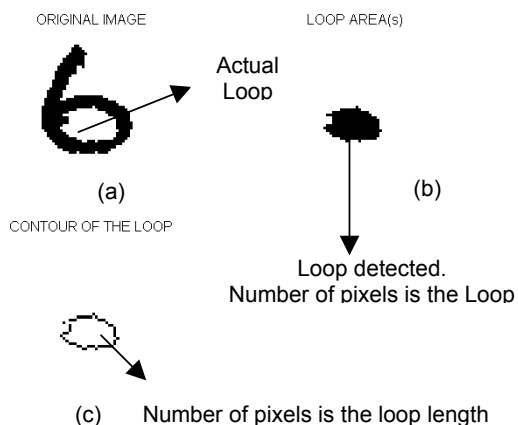


Fig. 1: End-Point(s) and junction(s) detected



**Fig. 2: Loop successfully detected**

- j) **F10 - Centre of Gravity** – This feature is calculated by first calculating the centre of gravity of the bounding box of a digit. Then the ratio of the respective coordinates of the centre of gravity with the Height and Width of the bounding box respectively is stored as a feature.
- k) **F11 - Gradient Features** – These features were obtained by dividing a digit image into 24 almost equal rectangles by imposing a 6x4 bounding box over the original image [3,4]. Then the gradients of each black pixel of the original image is calculated using a simple 3x3 Sobel operator

All the features from F1 to F11 (excluding the starred\* features) were appropriately binarized. Each of the features mentioned were binarized taking into consideration their nature (whether binary, continuous etc.) and also their ranges (eg. Range of Height Width ratio were assumed to be 0.5 to 3.0). Continuous features were binarized after selecting the bin/range in which they fall and binary features were used as they were. The process finally leads to a binary feature vector having 961 bits. Once binary feature vectors for all the digits of each writer are formed, the data can be used for the purpose of author discrimination by means of various classification techniques.

Features F4-2 and F4-3 were finally not used in our analysis since they were found to be computationally expensive.

The conversion of features is summarized in Table 1.

### 3. Digit Database

A database of handwritten numerals was collected from 15 writers. Each writer was asked to write several samples of

each digit. A healthy mix of people from various backgrounds was taken so as to make such a small database as close as possible to the real distribution and emulate the limited range of samples usually available to a document examiner in an actual court case. Each writer was asked to write the digits 0 to 9 in a random order almost 10 times, one set per line. Also for further scope of investigation the authors were asked to forge the handwritten digits of others. Hence for each person we also got 6 forged digit strings, from six different people.

All the filled forms obtained were scanned at 300 dpi resolution. Each digit string was segmented. Each digit string was further manually segmented and stored in the form of single digits. All segmented images were binarized. The images hence obtained (see Fig. 3) were used for applying feature extraction algorithms.

Feature	No. Of BITS	Multiplier	Total No. Of BITS
<b>F1</b>	20	1	20
<b>F2</b>	6	1	6
<b>F3</b>	6	1	6
<b>F4</b>	4	1	4
<b>F4-1</b>	5	1**	5
<b>F4-4</b>	12	1**	12
<b>F4-5</b>	1	1**	1
<b>F6</b>	5	12	60
<b>F7</b>	10	12	120
<b>F8</b>	5	9	45
<b>F9</b>	10+6	24	384
<b>F10</b>	5	2	10
<b>F11</b>	12	24	288

The total number of bits in the final feature vector = **961\*\***  
 \*\* Different for digit 8 since it has 2 main loops in it

**Table 1: Feature Binarization**

## 4. Results

We extracted the binary feature vector of size 961 for each digit from each writer. Since each digit was written at least 10 times by each writer we have 10 feature vectors per digit per writer. This set of 10 feature vectors per digit per writer was then divided randomly into various combinations of Standard (S) and Testing patterns (T) and percentage accuracies for writer identification, verification and forgery detection were obtained (Tables 2, 3 and 4).

A metric for binary feature vector distance computation (the Hamming distance) was used. For any given test vector we calculated its distance with each vector of each writer in the standard vector set (S). The digit test vector was finally assigned to the writer which produced the least average Hamming distance. The test set size (T) and the standard digit feature set size (S) were varied from 1 to 4 and 2 to 6 respectively. Also the accuracy calculated was found with respect to the digit's

whose test vectors were not rejected by the algorithm. Rejection of a digit test set was done if there were at least 2 such sets existing for 2 different digits having the same average Hamming distance with the standard feature set.

## 5. Conclusions and future work

A feature set was formed, which can be used for identification / verification of the authorship of handwritten documents containing digits. All the features were computational features and can easily be extracted by computer algorithms. The experiment conducted on the collected database allowed us to evaluate the discrimination power of the set. By using the Hamming distance measure and determining a threshold value for the intra-author variation a high degree of accuracy in authorship detection has been achieved. Although the set of features has been designed for handwritten digits, it may also be effectively used for handwritten characters.

However, the experiment performed has only been an evaluation of the feature set efficiency. To prove the efficiency of the set, the extraction and classification procedures need to be applied to a large database that is representative of a population. We have achieved higher levels of writer identification as compared to writer verification. This result can be explained on account of the threshold selection technique and distance metric we have used to compare binary strings. Hence, it would also be useful to implement other classification schemes and apply them to author verification / identification with aid of the designed feature set [6]. These are the current areas of research and will be reported at the conference.

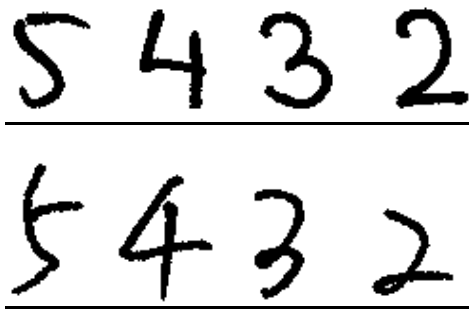


Fig.3: Example sample of segmented binary digits

## 6. References

- [1] Hilton, O., "Scientific Examination of Questioned Documents – Revised Edition", CRC Press, Inc., 1993.
- [2] Srihari S.N., Sung-Hyuk Cha and Sangjik Lee, "Establishing handwriting individuality using pattern recognition techniques", Proceedings of the Sixth International Conference on Document Analysis and Recognition, 2001, pp. 1195–1204.

[3] Srikanthan G., Lam S.W. and Srihari S.N., "Gradient-Based Contour encoding for character recognition", Pattern Recognition, Vol. 29, No.7, 1996, pp. 1147-1160.

[4] Srikanthan G. and Favata J. T., "A multiple feature/resolution approach to handprinted digit and character recognition, International Journal of Imaging and Systems and Technology, Vol. 7, Number 4, WINTER 1996, pp. 304-311.

[5] Leedham. G., "Image analysis tools for authentication and enhanced classification of handwritten script using forensic techniques – Final Report (RG25/95)", Nanyang Technological University, May 1999.

[6] Bin Zhang and Srihari S.N., "Binary Vector Dissimilarity Measures for Handwriting Identification", Proceedings of SPIE-IS&T Electronic Imaging, SPIE Vol.5010, 2003, pp. 28-38.

No. of Writers = 15		T=1			T=2			T=3			T=4		
DIGIT CONSIDERED	S=2	S=4	S=6	S=2	S=4	S=6	S=2	S=4	S=6	S=2	S=4	S=6	
0	0	26.7	26.7	10	30	33.3	20	33.3	35.6	23.3	35	36.7	
1	26.7	33.3	46.7	30	36.7	43.3	26.7	40	42.2	28.3	38.3	41.7	
2	33.3	40	46.7	40	46.7	56.7	44.4	51.1	57.8	45	50	56.7	
3	60	53.3	80	60	50	63.3	62.2	55.6	64.4	65	55	63.3	
4	46.7	73.3	66.7	50	70	63.3	48.9	66.7	60	50	68.3	66.7	
5	40	53.3	66.7	40	60	60	44.4	62.2	62.2	48.3	61.7	66.7	
6	46.7	46.7	53.3	40	53.3	56.7	40	53.3	60	36.7	50	55	
7	53.3	60	73.3	43.3	53.3	60	42.22	53.3	57.8	46.7	60	65	
8	33.3	40	46.7	36.7	43.3	43.3	40	44.4	42.2	45	46.7	46.7	
9	33.3	40	46.7	36.7	33.3	43.3	40	33.3	46.7	43.3	36.7	48.3	
Using 2 - 6	91.7, 20	86.7, 0	92.9, 6.7	86.7, 0	93.3, 0	100, 0	100, 0	100, 0	100, 0	100, 0	100, 0	100, 0	
Using 4,7 & 9	70, 33.3	71.4, 6.7	92.9, 6.7	61.5, 13.3	78.6, 6.7	86.7, 0	84.6, 13.3	86.7, 0	93.3, 0	78.6, 6.7	93.3, 0	100, 0	
All (0-9)	92.9, 6.7	92.9, 6.7	100, 6.7	92.3, 13.3	93.3, 0	100, 0	100, 6.7	100, 0	100, 0	100, 0	100, 0	100, 0	

Table 2: Accuracy of Writer Identification

No. of Writers = 15		T=1			T=2			T=3			T=4		
DIGIT CONSIDERED	S=2	S=4	S=6	S=2	S=4	S=6	S=2	S=4	S=6	S=2	S=4	S=6	
0	40	40	40	53.3	53.3	46.7	53.3	53.3	53.3	66.7	60	60	
1	40	86.7	66.7	46.7	66.7	73.3	46.7	73.3	66.7	46.7	80	73.3	
2	46.7	53.3	60	46.7	60	66.7	53.3	60	73.3	53.3	66.7	73.3	
3	53.3	80	80	60	60	66.7	66.7	80	80	66.7	73.3	80	
4	53.3	60	60	60	53.3	66.7	60	53.3	46.7	60	60	60	
5	80	40	66.7	60	60	73.3	66.7	66.7	80	80	73.3	80	
6	53.3	60	53.3	46.7	53.3	46.7	46.7	66.7	66.7	40	46.7	46.7	
7	60	60	60	53.3	66.7	46.7	60	60	46.7	53.3	53.3	60	
8	46.7	60	53.3	60	53.3	60	53.3	53.3	60	53.3	46.7	60	
9	53.3	66.7	53.3	66.7	80	66.7	73.3	73.3	60	73.3	80	60	
Using 2-6	60	73.3	80	60	60	80	66.7	80	93.3	66.7	73.3	80	
Using 4,7 & 9	60	66.7	53.3	66.7	73.3	66.7	73.3	73.3	53.3	80	66.7	66.7	
All (0-9)	80	86.7	86.7	80	80	93.3	80	93.3	86.7	86.7	86.7	80	

Table 3: Accuracy of Writer Verification

No. of Writers = 15		T=1			T=2			T=3			T=4		
DIGIT CONSIDERED	S=2	S=4	S=6	S=2	S=4	S=6	S=2	S=4	S=6	S=2	S=4	S=6	
0	60	66.7	60	60	86.7	80	80	93.3	73.3	66.7	86.7	73.3	
1	60	60	66.7	73.3	80	80	73.3	73.3	73.3	73.3	73.3	66.7	
2	60	86.7	86.7	66.7	86.7	86.7	60	86.7	86.7	60	93.3	93.3	
3	80	80	86.7	73.3	73.3	80	80	73.3	86.7	80	80	86.7	
4	60	80	93.3	73.3	80	93.3	73.3	93.3	93.3	73.3	93.3	93.3	
5	60	80	93.3	60	93.3	100	66.7	86.7	80	73.3	93.3	100	
6	53.3	86.7	80	60	100	93.3	66.7	100	100	86.7	100	93.3	
7	73.3	86.7	86.7	66.7	86.7	80	80	86.7	86.7	80	93.3	93.3	
8	73.3	73.3	80	66.7	93.3	100	73.3	86.7	100	66.7	86.7	100	
9	40	46.7	73.3	40	60	80	46.7	53.3	80	53.3	66.7	86.7	
Using 2-6	73.3	100	93.33	73.3	93.3	100	80	93.3	100	93.3	93.3	93.3	
Using 4,7 & 9	53.3	80	100	60	86.7	93.3	66.7	93.3	93.3	80	93.3	100	
All (0-9)	80	100	100	86.7	100	100	86.7	100	100	93.3	100	100	

Table 4: Accuracy of Forgery Detection