2011 International Conference on Document Analysis and Recognition

Evaluating the Rarity of Handwriting Formations

Sargur N. Srihari
CEDAR, University at Buffalo, The State University of New York, Buffalo, New York, USA
srihari@cedar.buffalo.edu

***Abstract*—Identifying unusual or unique characteristics of an observed sample in useful in forensics in general and handwriting analysis in particular. Rarity is formulated as the probability of letter formations characterized by a set of features. Modeling the distribution as a probabilistic graphical model several probabilities are inferred: the probability of random correspondence (PRC) as a measure of the discriminatory power of the characteristics, conditional PRC associated with a given sample and the probability of finding a similar one within tolerance in a database of given size. Using the most commonly occurring letter pair "th" and characteristics specified by questioned document examiners, the highest probability formation and low probability formations in a database are determined. Computational issues in scaling the methods are discussed.**

***Keywords*-handwriting;writer verification;graphical models;probability evaluation**

I. Introduction

Much of automatic handwriting recognition is concerned with determining the identity of a given letter or combination of letters by learning from example data about different forms encountered. On the other hand the goal of forensic handwriting examination is to determine as to how unusual a given structure or formation is so that it can be used to identify the writer. While an unusual, or rare, handwriting formation is central to identifying the writer, it is of little consequence and even considered as noise in recognition.

More generally, questioned document (QD) examination involves the comparison and analysis of documents, printing and writing instruments in order to identify or eliminate persons as the source. A facet of QD examination concerns handwriting comparison which is based on the premise that no two persons write the same way, while considering the fact that the writing of each person has its own variabilities [1], [2]. Individuals write differently both because they were taught differently, e.g, Palmer and D'Nelian methods which are called class-characteristics, and due to individual habits known as individualizing characteristics. Examples of such variations are seen in Figure 1. The uncertainties involved in handwriting makes it a task suitable to be characterized probabilistically.

Several types of probabilistic queries can be useful in the examination of handwriting evidence: (i) the probability of observed evidence, (ii) the probability of a particular feature observed in the evidence, (iii) the probability of finding the evidence in a representative database of handwriting exemplars. As an example, in the field of DNA evidence a statement can be made that "the chance that a randomly selected person would have the same DNA pattern as that of the sample source and the suspect was 1 in 24,000,000". In the case of fingerprints a similar statement can be made about the rarity of a particular minutiae pattern [3]. The objective of this paper is to describe as to how such probabilities can be computed for different letter formations in handwriting.

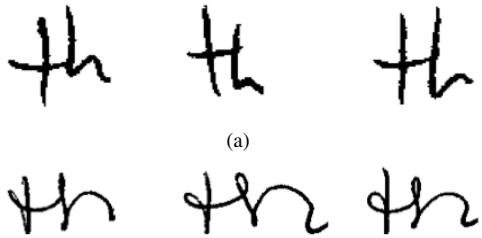


Figure 1. Samples of handwritten *th* of two writers showing two different writing styles as well as within-writer variability.

QD examiners specify handwriting characteristics (features) based on years of training [2]. However there have been very few efforts to characterize the statistical characteristics of such features, most notably [4]. On the other hand there have been efforts to compute features automatically– but the features tend to be gross approximations of the characteristics employed [5], [6] or the features do not correspond to human determined characteristics at all [7], [8].While these automated methods perform well in objective tests they do not lend support to the document examiner in testimony. Our goal here is to capture the statistics of document examiner specified characteristics so that a probability statement of rarity can be made as in other forensic domains. The need for such quantitative testimony in all of the forensic sciences has been underscored by a recent National Academy of Sciences report [9].

II. Features and Measurement Complexity

Consider the example of the most commonly encountered letter pair in the English language *th*; we consider a letter pair rather than a single letter since it is likely to be more individualistic. Characteristics for this letter pair as given by document examiners [4] is given in Table I. Thus the writing of *th* is characterized by a set of six features $X =$

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.130

Table I
CHARACTERISTICS OF *th* AS SPECIFIED BY DOCUMENT EXAMINERS.

R = **Height Relationship of *t* to *h***	L = **Shape of Loop of *h***	A = **Shape of Arch of *h***	C = **Height of Cross on *t* staff**	B = **Baseline of *h***	S = **Shape of *t***
r^0= *t* shorter than *h*	l^0 = retraced	a^0 = rounded arch	c^0 = upper half of staff	b^0 = slanting upward	s^0 = tented
r^1 = *t* even with *h*	l^1 = curved right side and straight left side	a^1 = pointed	c^1 = lower half of staff	b^1 = slanting downward	s^1 = single stroke
r^2 = *t* taller than *h*	l^2 = curved left side and straight right side	a^2 =no set pattern	c^2 = above staff	b^2 = baseline even	s^2 = looped
r^3 = no set pattern	l^3 = both sides curved		c^3 = no fixed pattern	b^3 = no set pattern	s^3 = closed
	l^4 = no fixed pattern				s^4 = mixture of shapes

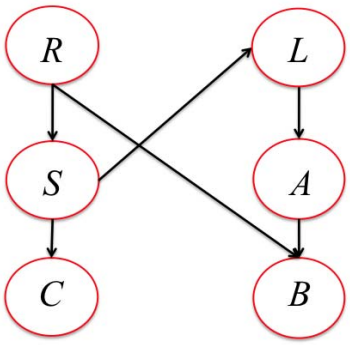


Figure 2. Bayesian network BN_{th}: this graph together with CPDs in Section III defines the probability distribution of the characteristics of *th*.

Table II
MARGINAL DISTRIBUTIONS OF FEATURES OF *th*.

Value	R	L	A	C	B	S
0	0.23	0.69	0.41	0.53	0.11	0.09
1	0.37	0.05	0.44	0.28	0.1	0.61
2	0.16	0.006	0.16	0.008	0.49	0.02
3	0.24	0.08	-	0.18	0.29	0.05
4	-	0.17	-	-	-	0.22

Table III
CONDITIONAL PROBABILITY DISTRIBUTION $P(S|R)$.

	s^0	s^1	s^2	s^3	s^4
r^0	0.21	0.48	0.02	0.1	0.19
r^1	0.04	0.68	0.04	0.05	0.2
r^2	0.07	0.71	0	0.04	0.18
r^3	0.05	0.57	0.02	0.04	0.31

$\{R, L, A, C, B, S\}$ where R takes on four possible values indicated by lower-case letters superscripted as r^0, r^1, r^2, r^3 and so on. The value is assigned to a particular writing sample, which can consist of several instances of *th*, as shown in Figures 1 and 4. For instance the three samples in Figure 1(a) will be jointly encoded as $r^1, l^0, a^0, c^3, b^1, s^2$ and the samples in Figure 1(b) as $r^2, l^2, a^0, c^1, b^0, s^2$.

In the probabilistic formulation each characteristic is considered to be a random variable. These six variables each have multinomial distributions with 4,5,3,4,4 and 5 possible values. If we assume that the variables are independent then the number of independent probabilities (parameters) to be estimated is $3+4+2+3+3+4 = 19$. On the other hand if we allow all dependencies, the number of parameters needed is $4 \times 5 \times 3 \times 4 \times 4 \times 5 - 1 = 4,799$.

When measurement complexity, i.e., the number of features and the number of discrete values for each feature, is increased the number of parameters needed grows exponentially making it impossible to have enough samples to determine the necessary parameters. Furthermore, in a classification scenario, when the sample size is fixed and finite, the average error rate over all distributions increases with measurement complexity [10].

III. PROBABILISTIC GRAPHICAL MODELS

The computational complexity and the need for samples can be managed by exploiting statistical independencies that exist between variables. Probabilistic graphical models are useful to express such independencies [11].

A Bayesian network, BN_{th}, for the distribution of the six variables in Table I is given in Figure 2. It incorporates causality such as: the shape of *t* (S) influences the shape of *h* loop (L), the shape of *h*-arch (A) influences the baseline of *h* (B), etc. This Bayesian network factorizes the distribution of *th* into component conditional probability distributions (CPDs) as

$$P(X) = P(R)P(L|S)P(A|L)P(C|S)P(B|R,A)P(S|R). \quad (1)$$

The CPDs are given in Tables II to VII, derived from data discussed in Section V.

The number of independent parameters needed to specify BN_{th} is $3 + 16 + 10 + 20 + 15 + 36 = 100$ which is far fewer than $4,799$ to directly specify the distribution.

IV. INFERRING RARITY FROM MODEL

Given the probabilistic graphical model the inference problem is that of evaluating probabilities of interest. For a distribution $P(X)$, the probabilities relating to rarity are

Table IV
CONDITIONAL PROBABILITY DISTRIBUTION $P(A|L)$.

	a^0	a^1	a^2
l^0	0.47	0.39	0.14
l^1	0.18	0.71	0.11
l^2	0.67	0	0.33
l^3	0.3	0.66	0.05
l^4	0.27	0.42	0.31

Table V
CONDITIONAL PROBABILITY DISTRIBUTION $P(L|S)$.

	l^0	l^1	l^2	l^3	l^4
s^0	0.4	0.06	0	0.34	0.19
s^1	0.8	0.05	0.01	0.04	0.11
s^2	0.54	0.15	0	0.08	0.23
s^3	0.59	0.03	0.03	0.17	0.17
s^4	0.56	0.06	0	0.08	0.3

defined: PRC or the Probability of Random Correspondence, nPRC or the PRC of at least one pair among n having the same value, conditional PRC which is the PRC of a known value X_s being found, and the corresponding conditional nPRC among n such samples [12].

A. PRC

Probability that two independent, identically distributed samples X_1 and X_2, each with distribution $P(X)$, have similar values is given by the graphical model in Figure 3(a). It is evaluated as follows:

$$\rho = P(z^0) = \sum_{X_1} \sum_{X_2} P(z^0|X_1, X_2)P(X_1)P(X_2) \quad (2)$$

where Z is an indicator variable which has the CPD in Table VIII, also given by

$$P(z^0|X_1, X_2) = \begin{cases} 1 \text{ if } d(X_1, X_2) \leq \epsilon \\ 0, otherwise, \end{cases} \quad (3)$$

d measures the difference between its arguments and the quantity ϵ represents as to how different two samples can be while they are considered to correspond (be the same). For instance $\epsilon = 0$ represents the requirement that the two values X_1 and X_2 are identical. If d is the number of characteristics that are the same then $\epsilon = 1$ would lead to X_1 and X_2 being considered to be the same if they do not differ in more than one variable.

Table VI
CONDITIONAL PROBABILITY DISTRIBUTION $P(C|S)$.

	c^0	c^1	c^2	c^3
s^0	0.85	0.06	0.02	0.06
s^1	0.47	0.37	0.01	0.15
s^2	0.69	0.31	0	0
s^3	0.83	0.1	0	0.07
s^4	0.46	0.17	0.01	0.36

Table VII
CONDITIONAL PROBABILITY DISTRIBUTION $P(B|R, A)$.

	b^0	b^1	b^2	b^3
r^0, a^0	0.03	0.17	0.53	0.27
r^0, a^1	0.18	0.12	0.41	0.29
r^0, a^2	0.13	0	0.4	0.47
r^1, a^0	0.08	0.06	0.7	0.17
r^1, a^1	0.13	0.15	0.51	0.21
r^1, a^2	0.04	0.09	0.35	0.52
r^2, a^0	0.13	0.13	0.53	0.2
r^2, a^1	0.12	0.17	0.37	0.34
r^2, a^2	0.08	0.25	0.5	0.17
r^3, a^0	0.12	0.06	0.44	0.38
r^3, a^1	0.16	0.07	0.45	0.32
r^3, a^2	0.06	0.09	0.36	0.48

Table VIII
DISTRIBUTION OF INDICATOR VARIABLE Z: $P(Z|X_1, X_2)$.

X_1, X_2	z^0	z^1
$d(X_1, X_2) \leq \epsilon$	1	0
$d(X_1, X_2) > \epsilon$	0	1

The probability that among a set of $n \geq 2$ independent, identically distributed samples $\mathbf{X} = \{X_1, .., X_n\}$, some pair have the same value within specified tolerance is given by the graphical model in Figure 3(b). The nPRC, can be written in terms of the PRC as

$$\rho[n] = 1 - (1 - \rho)^{\frac{n(n-1)}{2}}. \quad (4)$$

Note that when $n = 2$, PRC=nPRC. Since there are $\binom{n}{2}$ pairs involved this probability can be much higher than PRC. For instance, in the famous birthday paradox, while the probability of a birthday (PRC) is $1/365$, the value of nPRC for $n = 24$ is 0.5.

B. Conditional PRC

The probability that given a specific value it coincides, within tolerance, one in a set of n samples drawn from the same distribution is given by the graphical model in Figure 3(c). Since we are trying to match a specific value it depends on the probability of the conditioning value. It is smaller than nPRC and can be lower than the PRC. The exact relationship with respect to PRC depends on the distribution. The conditional nPRC is given by the marginal probability

$$p(Z = 1|X_s) = \sum_{\mathbf{X}} p(Z = 1|X_s, \mathbf{X})p(\mathbf{X}). \quad (5)$$

In the case of identical match this can be shown to be equivalent to

$$1 - (1 - P(X_s))^n \quad (6)$$

.

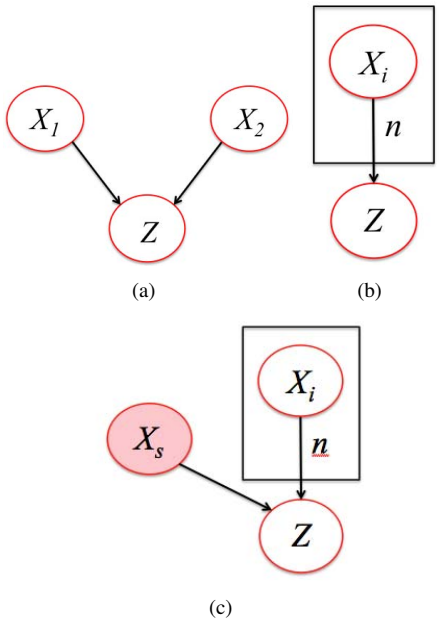


Figure 3. Graphical models for inferring rarity: (a) PRC, the probability of two samples having the same value, (b) nPRC, the probability of some pair of samples among n having the same value, and (c) conditional nPRC, the probability of finding X_s among n samples.

Figure 4. GUI for determining the features for *th* exemplars of a given writer: values are assigned manually using pull-down menus for each feature.

V. Rarity Evaluation

The CPDs for the network were computed from a data base of handwriting samples that are representative of the U.S. population [13]. The data for the CPDs are based on automatically extracting *th* samples and then assigning to each sample its characteristics manually using a tool with the interface shown in Figure 4. There were 3,125 images representing 528 authors, some of whom had just one sample while some had upto 5. Results of inference using network BN_{th} are described next.

A. Discriminatory Power of th

The PRC as given by Eq. (2) with $d(X,Y) = 0$, i.e., exact match, was evaluated to be 2.62×10^{-13}. This value can be used to compare the discriminative power of *th* to those of other letters and combinations.

B. Highest and Lowest Probabilities

Probabilities assigned by BN_{th} to each element in the database was evaluated using Eq. 1. The highest probability assigned by the model is to the feature value $\{r^1, l^0, a^0, c^0, b^2, s^1\}$ with probability 0.0304. It corresponds exactly to the features assigned to writer 100 in the database whose writing is shown in Figure 5(a). The lowest probability assigned is to $\{r^2, l^3, a^2, c^2, b^0, s^4\}$ with value 7.2×10^{-8} which does not have a corresponding element in the database. A low probability *th* is shown in Figure 5(b) $\{r^3, l^1, a^0, c^2, b^0, s^1\}$.

(a) (b)

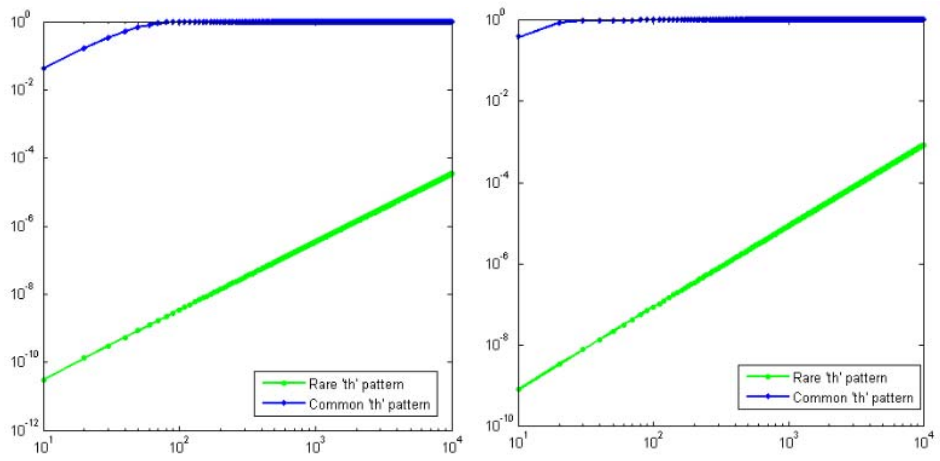
Figure 5. Examples of rarity evaluation: (a) the highest probability *th* in data set, and (b) a low probability *th*.

C. Probability of Finding Match in Database

Conditional nPRC for the two writers were evaluated using Eq. 6. Plots as a function of n for $d(X,Y) = 0$ and $d(X,Y) = 1$ are shown for the two writers in Figure 6 considering exact match and with a tolerance of mismatch in one feature. With $n = 10$ the probabilities of exact match for the two writers were 0.041 and 3.1×10^{-11} respectively, and probability allowing one mismatch were 0.387 and $7.69 \times 10 - 10$.

VI. Summary and Discussion

We have proposed a framework for evaluating the rarity of handwriting formations to be able to make probabilistic statements analogous to other forensic domains such as DNA and fingerprints. Since the probability specification involves the evaluation of a large number of parameters, we have described how probabilistic graphical models can be useful. Using document examiner specified features we constructed a Bayesian network for a commonly encountered letter pair



(a) (b)

Figure 6. Probability of finding matching entry in a database of size n for two cases: most common *th* and a rare case shown in Figure 5. Plots in (a) correspond to exact match, and (b) to match with one feature mismatch allowed.

th and showed how the number of parameters needed is much reduced.

We have also described the inferences needed to calculate rarity. They include evaluating the PRC , which is a measure of the discriminating power of a given set of characteristics, and the probability of finding it in a database of a given size. In the evaluation of Eq. 2 the number of terms in the summation was $4,800$ for the features described in Table I. With increasing measurement complexity, e.g., with full handwritten words represented by 1024 features [14], approximate inference methods will be useful.

In the methods described, the Bayesian network $BN_{th}$was manually specified. Since the number of possible letters and letter combinations can be large it is useful to automate network specification. Methods for automatically determining Bayesian networks are: (i) pairwise hypothesis testing, and (ii) searching the space of Bayesian networks (since this an NP-hard problem approximate search is used).

REFERENCES

[1] A. S. Osborn, *Questioned Documents*, Nelson Hall Pub, 1929.

[2] R. A. Huber and A. M. Headrick, *Handwriting Identification:Facts and Fundamentals*, CRC Press, 1999.

[3] C. Su and S. N. Srihari, Evaluation of Rarity of Fingerprints in Forensics, *Advances in Neural Information Processing Systems 23*, J. Lafferty and C. K. I. Williams and J. Shawe-Taylor and R.S. Zemel and A. Culotta (Eds.), NIPS 2010: 1207-1215.

[4] R. J. Muehlberger, K. W. Newman, J. Regent and J. G. Wichmann, A Statistical Examination of Selected Handwriting Characteristics, *Journal of Forensic Sciences*, 1977: 206-210.

[5] V. Pervouchine, *Discriminative Power of Features Used by Forensic Document Examiners in the Analysis of Handwriting*, PhD Thesis, Nanyang University, Singapore, 2006.

[6] A Bharadwaj, A. Singh, H. Srinivasan and S. N. Srihari, On the use of Lexeme Features for writer verification, *Proc. Int. Conf. Document Analysis and Recognition (ICDAR)*, Curitiba, Brazil, 2007: 1088-1092.

[7] G. R. Ball and S. N. Srihari, Statistical Characterization of Handwriting Characteristics using Automated Tools, *Proc. Document Recognition and Retrieval*, SPIE, 2011.

[8] S. N. Srihari, C. Huang and H. Srinivasan, On the Discriminability of the Handwriting of Twins, *Journal of Forensic Sciences*, March 2008, vol. 53(2), pp. 430-446.

[9] National Academy of Sciences, *Strengthening the Forensic Sciences: A Path Forward*, National Academy of Sciences press, 2009.

[10] G. F. Hughes, On the mean accuracy of statistical pattern recognition, *IEEE Transactions on Information Theory*, IT-14(1):55-63, 1968.

[11] D. Koller, *Probabilistic Graphical Models*, MIT Press, 2009.

[12] C. Su and S. N. Srihari, Probability of Random Correspondence of Fingerprints, *Proc. International Workshop on Computational Forensics*, Springer, 2009, pp. 55-66.

[13] S. N. Srihari, S. Cha, H. Arora and S Lee, Individuality of Handwriting, *Journal of Forensic Sciences*, 2002, 47(4): 856-872.

[14] B. Zhang, S. N. Srihari and C. Huang, Word Image Retrieval Using Binary Features, *SPIE 5296*, (2003); doi:10.1117/12.523968