

A Statistical Model For Writer Verification

Sargur N. Srihari

Matthew J. Beal

Karthik Bandi

Vivek Shah

Praveen Krishnamurthy

Department of Computer Science and Engineering, University at Buffalo

{srihari, mbeal, kkr2, vashah, pk35}@cse.buffalo.edu

Abstract

A statistical model for determining whether a pair of documents, a known and a questioned, were written by the same individual is proposed. The model has the following four components: (i) discriminating elements, e.g., global features and characters, are extracted from each document, (ii) differences between corresponding elements from each document are computed, (iii) using conditional probability estimates of each difference, the log-likelihood ratio (LLR) is computed for the hypotheses that the documents were written by the same or different writers; the conditional probability estimates themselves are determined from labelled samples using either Gaussian or gamma estimates for the differences assuming their statistical independence, and (iv) distributions of the LLRs for same and different writer LLRs are analyzed to calibrate the strength of evidence into a standard nine-point scale used by questioned document examiners. The model is illustrated with experimental results for a specific set of discriminating elements.

1. Introduction

Writer verification is the task of determining whether two handwriting samples were written by the same or by different writers, a task of importance in Questioned Document Examination [QDE]. This paper describes a statistical model of the task which has three salient components: (i) discriminating element extraction and similarity computation, (ii) modeling probability densities for the similarity values, conditioned on being from the same or different writer, as either Gaussian or gamma, and determining the log-likelihood ratio (LLR) function and (iii) computing the strength of evidence. Each of the components of the model are described in the following sections.

2. Discriminating elements & similarity

Discriminating elements are characteristics of handwriting useful for writer discrimination. There are many discriminating elements for QDE, e.g., there are 21 classes of discriminating elements [1]. In order to match elements between two documents, the presence of the elements are first recognized in each document. Matching is performed between the same elements in each document. Although the proposed model is general we describe here a set of discriminating elements as an example. The model can be used with any other set of features.

Elements, or features, that capture the global characteristics of the writer's individual writing habit and style can be regarded to be macro-features and features that capture finer details at the character level as micro-features. For instance macro features are gray-scale based (entropy, threshold, no. of black pixels), contour based (external and internal contours), slope-based (horizontal, positive, vertical and negative), stroke-width, slant and height. Since macro features are real-valued absolute differences are used for similarity.

For micro features of characters a set of 512 binary-valued micro-features corresponding to gradient (192 bits), structural (192 bits), and concavity (128 bits) which respectively capture the finest variations in the contour, intermediate stroke information and larger concavities and enclosed regions, e.g., [2] are used. Since micro features are binary valued several binary string distance measures can be used for similarity of characters, the most effective of which is the correlation measure [3].

Discriminability of a pair of writing samples based on similarity value can be observed by studying their distributions when the pair arise from either the same writer or from different writers. Considering the 62 micro features, Fig. 1 is the plot obtained after performing Principal Component Analysis (PCA) and reducing the dimensionality to 2. which shows that the same and different writer classes are fairly separable using micro-features.

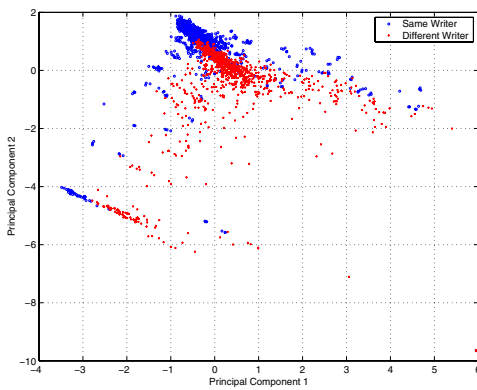


Figure 1. Principal Components of micro features reducing dimensionality to 2.

3. Probability densities and likelihood

The distributions of similarities conditioned on being from the same or different writer is used to compute likelihood functions for a given pair of samples. Several choices exist: assume that each density is Gaussian and estimate the Gaussian parameters, assume the density is gamma and estimate its parameters. The Kullback Leibler (KL) divergence test can be performed for each of the features to estimate whether it was better to model them as Gaussian or as gamma distributions. The gamma distribution is better model since distances are positive valued whereas the Gaussian assigns non-zero probabilities to negative values of distances.

Similarity histograms corresponding to same writer and different writer distributions for numeral 3 (micro features) and for entropy (macro feature) are shown in Fig. 2. Table 1 gives the KL test result values, in bits, for each macro feature. A training set size of 1000 samples was chosen for each of same and different writers. The test set size was 500 for each. As can be seen values for gamma are consistently lower than values for Gaussian thereby indicating that gamma is a better fit.

3.1 Parametric models

Assuming that similarity data can be acceptably represented by Gaussian or gamma distributions, probability density functions of distances conditioned upon the same-writer and different-writer categories for a single feature x have the parametric forms $p_s(x) \sim N(\mu_s, \sigma_s^2)$, $p_d(x) \sim N(\mu_d, \sigma_d^2)$, for the Gaussian case and $p_s(x) \sim Gam(a_s, b_s)$, $p_d(x) \sim Gam(a_d, b_d)$ for the gamma case. The Gaussian and gamma density functions are as follows.

$$\text{Gaussian: } p(x) = \frac{1}{(2\pi)^{1/2}\sigma} \exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Table 1. KL test results for macro features

Macro Feature	Same Writer		Different Writer	
	Gamma	Normal	Gamma	Normal
Entropy	0.133	0.921	0.047	0.458
Threshold	3.714	4.756	2.435	3.882
No. Black Pixels	1.464	2.314	2.151	2.510
External contours	2.421	3.517	2.297	2.584
Internal contours	2.962	3.373	2.353	2.745
Horizontal Slope	0.050	0.650	0.052	0.532
Positive Slope	0.388	1.333	0.173	0.315
Vertical Slope	0.064	0.664	0.054	0.400
Negative Slope	0.423	1.385	0.113	0.457
Stroke width	3.462	6.252	3.901	4.894
Average Slant	0.392	1.359	0.210	0.362
Average Height	3.649	4.405	2.558	2.910

$$\text{Gamma: } p(x) = \frac{x^{a-1} \exp(-x/b)}{(\Gamma(a)) \cdot b^a}$$

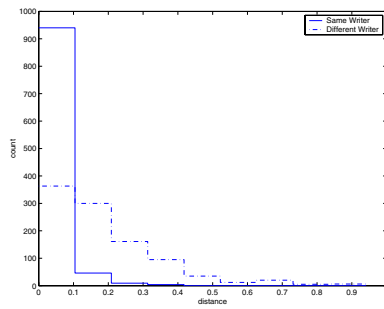
Estimating μ and σ from samples using the usual maximum likelihood estimation the parameters of the gamma distribution are calculated as $a = \mu^2/\sigma^2$ and $b = \sigma^2/\mu$. Conditional parametric pdfs for the numeral 3 (micro-feature) and for entropy (macro feature) are shown in Fig. 3. The parameters for the macro distributions (for a training set of size 1000) are given in Table 2.

The likelihood ratio (LR), which summarizes the result, is given by $LR(x) = p_s(x)/p_d(x)$. The log-likelihood ratio (LLR), obtained by taking the natural logarithm of LR, is more useful since LR values tend to be very large (or small).

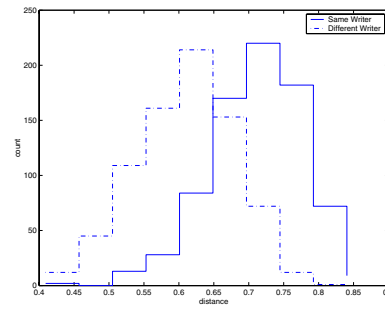
The error rates (percent misclassification) for a test set of size 500 using macro features are given in Table 3. The average error rate is lower for gamma over Gaussian although for one of the two classes (same writer) the Gaussian does better.

Table 3. Error rates for macro features

Macro Feature	Same Writer		Different Writer	
	Gamma	Normal	Gamma	Normal
Entropy	21.30	13.00	23.20	38.40
Threshold	2.60	2.60	53.40	60.00
No. Black Pixels	22.19	9.80	22.40	39.60
External contours	30.10	6.20	18.60	46.80
Internal contours	28.30	8.80	33.00	56.60
Horizontal Slope	13.44	5.40	25.20	34.40
Positive Slope	10.59	3.60	16.80	31.20
Vertical Slope	11.60	5.60	23.20	31.60
Negative Slope	14.46	3.00	23.10	37.00
Stroke width	23.20	23.20	0.00	31.60
Average Slant	9.97	3.00	18.60	31.40
Average Height	17.43	5.00	22.40	40.80

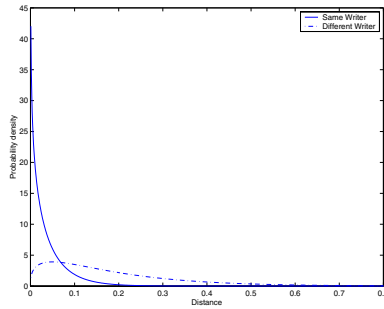


(a)

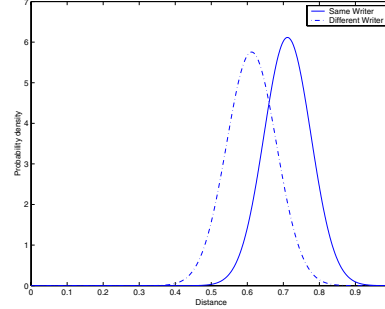


(b)

Figure 2. Histograms for same and different writers for (a) entropy (macro) (b) numeral 3 (micro).



(a)



(b)

Figure 3. Parametric pdfs for: (a) entropy (gamma distributions) and (b) numeral 3 (Gaussians).

3.2 The multivariate case

In the case where the document is characterized by more than one feature we assume that the writing elements are statistically independent. Although this is strictly incorrect the assumption has a certain robustness in that it is not an overfitting of the data. The resulting classifier, also known as Naive Bayes classification, has yielded good results in machine learning. Moreover, in the earliest QDE literature, there is reference to multiplying the probabilities of handwriting elements e.g., [4].

Each of the two likelihoods that the given pair of documents were either written by the same or different individuals can be expressed, assuming statistical independence of the features as follows. For each writing element $e_i, i = 1, \dots, c$, where c is the number of writing elements considered, we compute the distance $d_i(j, k)$ between the j th occurrence of e_i in the first document and the k th occurrence of e_i in the second document for that writing element. We estimate the likelihoods as

$$L_s = \prod_{i=1}^c \prod_j \prod_k p_s(d_i(j, k))$$

$$L_d = \prod_{i=1}^c \prod_j \prod_k p_d(d_i(j, k)).$$

The log-likelihood ratio (LLR) in this case has the form

$$LLR = \sum_{i=1} \sum_j \sum_k \ln p_s(d_i(j, k)) - \ln p_d(d_i(j, k)).$$

The two cumulative distributions of LLRs corresponding to same and different writer samples are shown in Figure 4. As the number of features considered decreases the separation between the 2 curves also decreases. The separation gives an indication of the separability between classes. The more the separation the easier it is to classify. In order to calibrate the system we analyze the distribution of LLRs for each feature and use the CDF for same writer LLR values and inverse CDF for different writer LLR values. Fig. 4 shows the CDF for same writer and different writer LLRs respectively.

4. Evaluation of strength of evidence

In order to present the result in the form of the strength of evidence [5] it is useful to represent the LLR scores on a scale ranging from -1 to 1 (-1 representing a confident different writer case and 1 representing a confident same writer case). It would be inappropriate to state all results with $LLR > 0$ as same writer and all results with $LLR < 0$ as different writer. Instead of using binary decisions a range of interpreted results is always better and more practical from the

Table 2. Gaussian and gamma parameters for 12 macro features.

Feature	Same Writer		Different Writer		Same Writer		Different Writer	
	Gaussian Parameters				Gamma Parameters			
	μ_s	σ_s	μ_d	σ_d	a_s	b_s	a_d	b_d
Entropy	0.0379	0.044	0.189	0.162	0.752	0.050	1.355	0.139
Threshold	1.603	2.025	12.581	35.430	0.627	2.559	0.126	99.779
No. Black Pixels	22761	28971	107061	89729	0.617	36875	1.424	75204
External contours	1.828	2.965	9.135	7.703	0.380	4.810	1.406	6.496
Internal contours	2.626	2.348	5.830	5.144	1.251	2.100	1.285	4.538
Horizontal Slope	0.013	0.014	0.072	0.066	0.930	0.014	1.179	0.061
Positive Slope	0.014	0.023	0.112	0.081	0.392	0.037	1.890	0.059
Vertical Slope	0.016	0.016	0.101	0.083	1.041	0.015	1.492	0.068
Negative Slope	0.008	0.013	0.060	0.050	0.381	0.021	1.416	0.042
Stroke width	0.235	0.431	0.968	1.185	0.297	0.791	0.667	1.451
Average Slant	1.730	2.674	12.402	8.955	0.419	4.133	1.918	6.465
Average Height	1.809	1.920	8.458	7.147	0.888	2.037	1.400	6.039

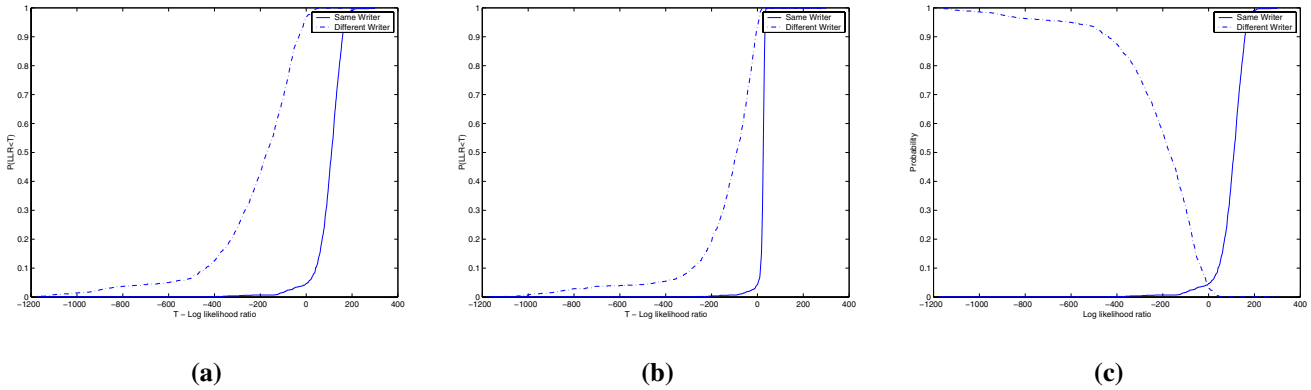


Figure 4. CDF of LLRs for same and different writer populations: (a) 12 macro and 62 micro features, (b) 12 macro and 8 micro features, and (c) CDF of LLRs for same writers and inverse CDF for different writers

QDE point of view. Based on the Tippett plot [6] (inverse CDF plot) and the CDF plot we develop a scheme for calibrating the LLR scores. The CDF and inverse CDF for same and different writer LLRs (considering 12 macro features and 62 micro features) is shown in Fig. 4.

In the case of characters since the number of matches is variable resulting in multiple instances of each feature. Since the number of features is unbound it is necessary to do some averaging to bound the value. Assume a set of m features (macro features and 62 characters 0-9, a-z, A-Z represented by micro features). If for the i^{th} feature we get k_i LLR values, $LLR_{i1}, LLR_{i2}, \dots, LLR_{ik_i}$ then their average is

$$LLR_{ave}(i) = \frac{1}{k_i} \sum_{j=1}^{k_i} LLR_{ij}.$$

For each feature i the CDF and inverse CDF for same and different writers are obtained from the distribution

of $LLR_{ave}(i)$. For the same writer case we obtain $P_{same_i}(LLR < LLR_{ave}(i))$ from the CDF of same writer LLR for that feature. For the different writer case we obtain $P_{diff_i}(LLR > LLR_{ave}(i))$ from the inverse CDF of different writer LLR for that feature.

Assuming m features are available we compute the geometric means $P1 = \prod_{j=1}^m (P_{same_j})^{1/m}$ and $P2 = \prod_{j=1}^m (P_{diff_j})^{1/m}$ to make the calibration independent of the number of features present. Finally we compute the calibration score as $Score = P1 - P2$ which lies in the interval $[-1, 1]$. QD examiners use a nine-point opinion scale: identify, highly probable, probable, indicative did, no conclusion, indicative did not, probably did not, highly probable did not and eliminate. Scatter plots of scores obtained are shown in Fig. 5 for 500 same and 500 different writer cases. Observing the histograms of scores for same and different writers the score range is divided into nine zones.

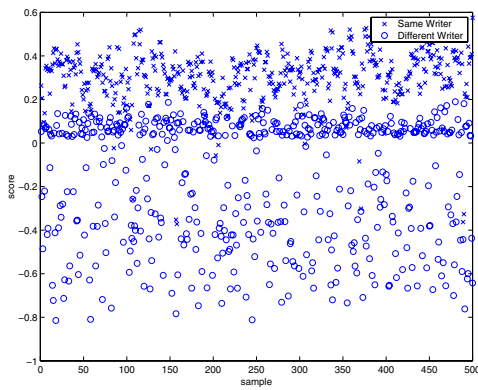


Figure 5. Scatter plots of scores for same and different writer sets after calibration: same writer values have high scores and low scores indicate different writers.

Table 4 shows the distribution of scores into the nine zones. Zones 1-4 represent the *same writer*, Zone 5 represents *no conclusion* and Zones 6-9 represent *different writers*. For a given set of test pairs accuracies are calculated as follows. Let s_1 = number of same writer cases falling into same writer zones and s_2 = number of same writer cases not falling into the 'no conclusion' zone. Then same writer accuracy is s_1/s_2 . Similarly, let d_1 = number of different writer cases falling into the different writer zones and d_2 = number of different writer cases not falling into the 'no conclusion' zone. Then different writer accuracy is d_1/d_2 . Results are based on a test set of 500 same and 500 different writers. Based on the zones obtained 2.2 % of same writer and 4.8 % of different writer cases fell into the 'no conclusion' zone. Same writer accuracy was 94.6 % while different writer accuracy was 97.6 %.

5. Summary and Conclusion

A statistical model for writer verification has been proposed with the following components: (i) extracting characteristics from the questioned and known documents and computing corresponding differences, (ii) likelihoods for the two classes are computed assuming statistical independence of the distances— where the conditional probabilities for the differences are estimated using parametric probability densities which are either Gaussian or gamma, (iii) log-likelihood ratio (LLR) of same and different writer are computed, and (iv) cdfs of the LLRs are used to calibrate the LLR values into a nine-point scale so as to present the strength of evidence. Results using the model with a test-bed representing 1,000 pairs of handwriting samples has been presented.

Table 4. Calibration of score

Zone	Opinion	Same (%)	Different (%)
1	Identified as same (> 0.5)	2.0	0.0
2	Highly probable same (> 0.35 & < 0.5)	34.6	0.0
3	Probably did (> 0.2 & < 0.35)	48.8	0.0
4	Indications Did (> 0.15 & < 0.2)	7.2	2.2
5	No Conclusion (> 0.12 & < 0.15)	2.2	4.8
6	Indications did not (> -0.05 & < 0.12)	3.6	45.4
7	Probably did not (> -0.3 & < -0.05)	0.6	12.6
8	Highly probable did not (> -0.65 & < -0.3)	1.0	27.8
9	Identified as different (< -0.65)	0.0	7.2

Acknowledgement

This project was supported in part by Grant Number 2004-IJ-CX-K050 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position or policies of the US Department of Justice.

References

- [1] R. Huber and A. Headrick, *Handwriting Identification: Facts and Fundamentals*. CRC Press, 1999.
- [2] S. N. Srihari, S. H. Cha, and S. Lee, "Individuality of handwriting," in *Journal of Forensic Sciences*, 2002, pp. 856–872.
- [3] B. Zhang, S. N. Srihari, and S.-J. Lee., "Individuality of handwritten characters," in *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)*, 2003, pp. 1086–1090.
- [4] A. S. Osborn, "Questioned documents." Nelson Hall Pub., 1929.
- [5] C. Champod, "The inference of identity of source: Theory and practice," in *The First International Conference On Forensic Human Identification In The Millennium, London, UK.*, October 1999, pp. 24–26.
- [6] F. L. Tippett C. F., V. J. Emerson and S. Lampert, "The evidential value of the comparison of paint flakes from sources other than vehicles," in *Journal of Forensic Sciences Society*, vol. 8, 1968, pp. 61–65.