Comparison of Statistical Models for Writer Verification

Sargur Srihari and Gregory R. Ball

Department of Computer Science and Engineering Center of Excellence for Document Analysis and Recognition (CEDAR) UB Commons 520 Lee Entrance, Suite 202 Amherst, NY 14228

ABSTRACT

A novel statistical model for determining whether a pair of documents, a known and a questioned, were written by the same individual is proposed. The goal of this formulation is to learn the specific uniqueness of style in a particular author's writing, given the known document. Since there are often insufficient samples to extrapolate a generalized model of an writer's handwriting based solely on the document, we instead generalize over the differences between the author and a large population of known different writers. This is in contrast to an earlier model proposed whereby probability distributions were a priori without learning. We show the performance of the model along with a comparison in performance to the non-learning, older model, which shows significant improvement.

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]. The problem consists of two handwritten documents, referred as to the known and the questioned document. The goal is to determine if the two documents were written by the same person. We present a system for performing writer verification which captures the idea of writer uniqueness. Intuitively, the system works by pairwise comparing letters of the same class between the two documents and determining their similarity to one another (by computing a similarity distance). Two conditional probability estimates are then computed based on each distance—(i) the probability of the two characters being produced by a single writer (i.e., the distance being explained by normal variation) and (ii) the chance of characters being produced by two different writers. The log-likelihood ratio (LLR) is computed and is used to determine the strength of confidence for the opinion.

A unique contribution of our method is our approach to generating the probability estimates. Our method captures the uniqueness of an individual writer by generating a probability distribution of distances between instances of character classes in the known document and the general population, as represented by a set of many samples. This intuitively captures how different a writer's sample is from the general population. While it might be desirable to generate this directly by determining the distribution of features on the writer's own samples, there will in practice often be few instances of character classes. By comparing with the general population, the uniqueness can be learned even when only a small set of characters is present.

An earlier model put forth by Srihari et al¹ addressed writer verification, but did so with static probability models. In that approach, a normal distribution was generated a priori from pairwise comparisons of different writers of the same character class. While this model performed well, it failed to capture the notion of uniqueness in the writer. In effect, the conditional probability was predicated on the odds of the difference between the known and questioned document falling in the distribution of how different writers typically differ. In our approach, we determine its placement in the distribution of how the known writer compares to writers who are not the known writer. Since this is done at runtime for many instances of available character classes (as opposed

Further author information:

E-mail: srihari@cedar.buffalo.edu, Telephone: (716) 645-6164

to simply finding their placement in the precomputed distribution), more resources are needed at runtime, but the yield is higher quality results which is typically a desirable tradeoff in writer verification.

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.² 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.,³ 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.⁴

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 microfeatures.

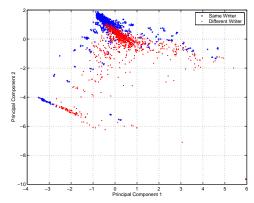


Figure 1. Principal Components of micro features reducing dimensionality to 2.

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.

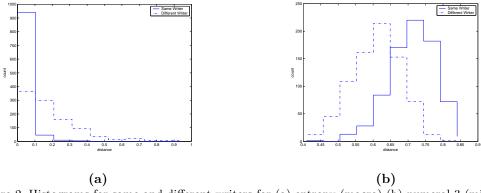


Figure 2. Histograms for same and different writers for (a) entropy (macro) (b) numeral 3 (micro).

As mentioned before, rather than calculate the distribution for the different writers a priori, we instead generate the different writer distribution at runtime, basing it one the pairwise comparisons of all characters of a class in the known document with the different writer set. Figure 4 shows the generalized distribution in the center along with the writer-specific distributions on the right and left side. This is capturing the idea of uniqueness, as illustrated in Figure 3. The distribution which includes the learned (consistent) writer is more valid for generating likelihood probabilities since it actually captures the individual uniqueness of the writer.

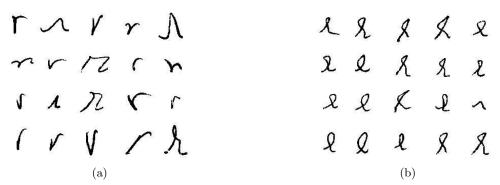


Figure 3. (a) set of character "r" by multiple writers (b) sampling of character "r" writing by a single writer

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.

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

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

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. 5.

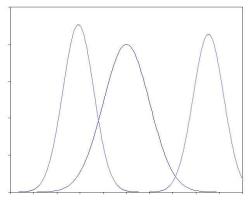


Figure 4. The center Gausian represents a generalized a priori learning. The left and right Gausians represent learned uniqueness, with the right Gausian representing a writer who has a more unique style than the left

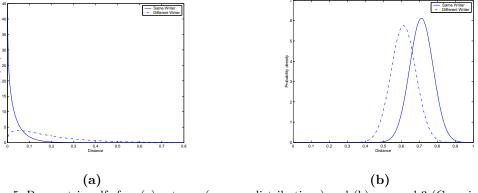


Figure 5. Parametric pdfs for: (a) entropy (gamma distributions) and (b) numeral 3 (Gaussians).

The likelihood ratio (LR), which summarizes the result, is iven 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).

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 classifer, 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.,.⁵

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, ..., c, where c is the number of writing elements considered, we compute the distance $d_i(j, k)$ between the jth occurrence of e_i in the first document and the kth 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 6. 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. 6 shows the CDF for same writer and different writer LLRs respectively.

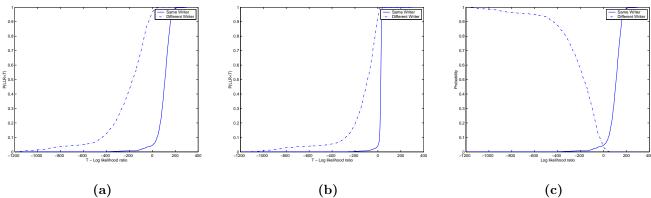


Figure 6. CDF of LLRs for same and different writer populations: (a) 12 macro and 62 micro features, (b) 12 macro and 8 micro features, and (c) CDF of LLRs for same writers and inverse CDF for different writers

3.3 Experiments

To evaluate performance, we use the CEDAR letter database.⁶ This database consists of 1,000 writers writing a handwritten document three times containing all capital and lower-case letters of the English alphabet. We created a random test set from this set which paired an equal number of same writer and different writer combinations (so as to avoid bias). The system then scored the match, evaluating a positive LLR as the system saying the authors were the same, and a negative LLR indicating they were different. The magnitude of the score is indicative of the confidence. Therefore, a system which has a larger magnitude (correct) score has more definitively stated that the authors are in fact the same or different.

4. 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. The strength of this model is that the uniqueness of a writer's sample is learned at runtime, affording better quality results.

Preliminary results consistently show a significant improvement, with all data tested so far receiving amplified confidence scores or more correct numbers. The improvement seen was on the order of at least 10%. More detailed result information will be available shortly.

REFERENCES

- [1] Srihari, S. N., Beal, M., Bandi, K., Shah, V., and Krishnamurthy, P., "A statistical model for writer verification," in [*Proc. International Conference on Document Analysis and Recognition*], 1105–1109 (August 2005).
- [2] Huber, R. and Headrick, A., [Handwriting Identification: Facts and Fundamentals], CRC Press (1999).

- [3] Srihari, S. N., Cha, S. H., and Lee, S., "Individuality of handwriting," in [Journal of Forensic Sciences], 856–872 (2002).
- [4] Zhang, B., Srihari, S. N., and Lee., S.-J., "Individuality of handwritten characters," in [Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)], 1086–1090 (2003).
- [5] Osborn, A. S., "Questioned documents," Nelson Hall Pub. (1929).
- [6] Cha, S.-H. and Srihari, S. N., "Writer identification: Statistical analysis and dichotomizer," in [Proceedings of SS&SPR 2000 WVCS Advances in Pattern Recognition], 1876, 123–132 (2000).