System that Identifies Writers

Sung-Hyuk Cha and Sargur N. Srihari
Center of Excellence for Document Analysis and Recognition
State University of New York at Buffalo
{scha,srihari}@cedar.buffalo.edu

Since the writer identification plays an important investigative and forensic role in many types of crime, various automatic techniques, feature extraction, comparison and performance evaluation methods have been studied (see [1] for the extensive survey). It has been practiced based on the hypothesis that people's handwritings are as distinctly different from one another as their individual natures, as their own finger prints [2,3]. However, relatively little study has been carried out to demonstrate its scientific and statistical validity and reliability as forensic evidence [4] or to answer the question whether one can build a machine that can identify writers. For this reason, we present a simple model to establish the individuality of handwriting based on Hilton's model. Hilton calculated the *odds* by taking the likelihood ratio statistic that is the ratio of the probability calculated on the basis of the similarities, under the assumption of identity, to the probability calculated on the basis of dissimilarities, under the assumption of non-identity [3,5].

There exist various parametric and non-parametric techniques to solve the multiple category classification problem or simply called *polychotomizer* where the number of classes is finite and small [6]. As the number of classes is too large to observe all (U.S. population), these techniques are of no use and the problem is seemingly insurmountable. For this reason, we suggest to transform a large and intractable polychotomizer to a simple *dichotomizer*, a classifier that places a pattern in one of only two categories: distance data between two writings of the same author and those of two different authors. In this model, one need not observe all classes and still allows the inferential classification. We state the problem as follows; given two randomly selected handwritten documents, the *writer identification* problem is to determine whether the two documents were written by the same person with two types of confusion error probabilities. To illustrate, suppose there are three writers, $\{W_1, W_2, W_3\}$. Each writer provides three documents and two scalar value features extracted per document. Fig. 1 (a) shows the plot of two features from documents for every writer and Fig. 1 (b) represents the transformed plot in the two dimensional feature distance domain.

Using eleven feature distance values, we trained an *artificial neural network* and obtained 97% overall correctness.

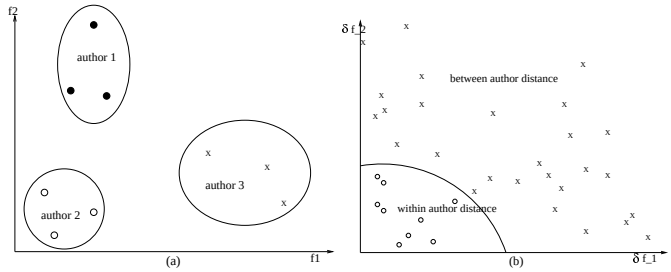

Figure 1: *Transformation from Feature domain to Feature distance domain*

In this experiment, 571 people provided three sample handwritings resulting in $1,713$ within and $1,464,615$ between author distance data. The type I and II error rates are 2.1% and 3.5%, respectively. Type I error occurs when two documents are determined to be written by two different authors even though they were written by one author and type II error is vice versa.

Acknowledgments This research is funded by National Institute of Justice (NIJ): Award Number 1999-IJ-CX-K010.

References

[1] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification - the state of the art," *Pattern Recognition*, vol. 22, no. 2, pp. 107–131, 1989.

[2] R. R. Bradford and R. B. Bradford, *Introduction to Handwriting Examination and Identification*. Nelson-Hall Publishers: Chicago, 1992.

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

[4] J. Travis, "Forensic document examination validation studies." Solicitation: http://ncjrs.org/pdffiles/sl297.pdf, October 1998.

[5] O. Hilton, "The relationship of mathematical probability to the handwriting identification problem," in *Proceedings of Seminar No. 5*, pp. 121–130, 1958.

[6] R. O. Duda, D. G. Stork, and P. E. Hart, *Pattern Classification and Scene Analysis*. John Wiley & Sons, Inc., 2nd ed., 2000.