Repudiation Detection in Handwritten Documents

Sachin Gupta and Anoop M. Namboodiri

International Institute of Information Technology, Hyderabad, India
sachin_g@students.iiit.ac.in, anoop@iiit.ac.in

Abstract. Forensic document verification presents a different and interesting set of challenges as opposed to traditional writer identification and verification tasks using natural handwriting. The handwritten data presented to a forensic examiner is often deliberately altered, in addition to being limited in quantity. Specifically, the alterations can be either forged, where one imitates another person's handwriting; or repudiated, where one deliberately distorts his handwriting in order to avoid identification. In this paper, we present a framework to detect repudiation in forensic documents, where we only have one pair of documents to arrive at a decision. The approach generates a statistically significant confidence score from matching two documents, which can be used to screen the documents that are passed on to an expert examiner. The approach can be extended for detection of forgeries as well.

1 Introduction

In forensic science, the primary role of handwriting analysis is in Questioned Document Examination (QDE) [1,2]. Determining the authorship of a document is the main task in QDE, where one has to decide whether a pair of documents, the *questioned document* (one whose origin is unknown) and the *reference document* (one whose origin might be known), were written by the same writer or not. However, due to the circumstances under which the documents are generated, there is a motivation for the writer to deliberately alter his natural handwriting to avoid detection. We refer to this problem as *handwriting repudiation*, as the purpose of distortion is to deny someone's involvement in the case (repudiation [3]).

The problem of detection of repudiation in QDE is different from that of traditional writer identification and verification tasks. Writer identification is the problem of identifying the writer of a document as one of the enrolled candidates. In writer verification, the writer claims an identity and we need to verify the claim. In both identification and verification problems, writer need to be enrolled into the system before hand. However, in the case of forensic documents, we only have a single questioned document, and a reference sample collected from the suspect, which in turn might not be natural. The two problems that arise here are:

– *Forgery Detection*: The problem is identical to that of verification, except that there is an additional suspicion that the writer could be an impersonator.
– *Repudiation Detection*: Given two samples of handwriting (both could be deliberately distorted), verify the claim that they are from different writers.
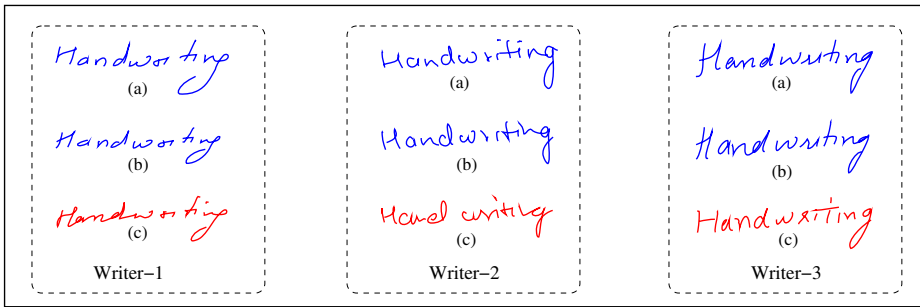
Fig. 1. (a) and (b) show natural handwriting samples from 3 writers, and (c) shows the repudiated samples from the writers

Note that in both identification and verification tasks, the users are assumed to be cooperative, and one could build statistical models for each writer from their natural handwriting. However, in the case of forgery, the questioned document need not be natural, and in repudiation, both the questioned document and the reference document could be distorted. Moreover, we have to assume that the writer is non-cooperative. Figure 1 shows examples of words from two writers in their natural form, as well as when they distort their handwriting for repudiation.

In this paper, we primarily deal with the problem of repudiation in handwritten documents. We propose a generic framework for automated analysis of handwritten documents to flag suspicious cases of repudiation.

1.1 Automatic Detection of Repudiation

Extraction of writer information from handwriting is more challenging as compared to verification based on physical biometrics traits, due to the large intra-class variation (between handwriting samples of the same person), and the large inter-class similarity (same words being written by different people). Moreover, the handwriting of a writer may also be affected by the nature of the pen, writing surface, and writers mental state. In addition, the problem of forensic document analysis is particularly difficult due to the additional problems posed by repudiation:

- During repudiation, a writer tries to change his handwriting style to be different from that of his natural handwriting. This introduces a large amount of intra-class variability that the system has to handle. Moreover, the writer need to be assumed to be non-cooperative, unlike in forgery, where the person who is being forged will be cooperative and will provide his natural handwriting in the required manner and amount.
- The content of the handwriting that is available in QDE is not in our control, and is often small in quantity. This prevents us from using the less frequent statistical properties of the handwriting for the purpose of verification of the claims.

- The cost of false match is often very high in the case of forensic documents, as it might result in erroneous conviction of an innocent person. Moreover, to use such an evidence in the court, one needs to give a statistically valid confidence measure in the result that is generated.

In spite of all these problems, it has been shown by forensic experts, that repudiation detection is possible. From the principle of *exclusion* and *inclusion*, inferred by document examiners from their experience in the field, one can't exclude from one's own writing, those discriminating elements of which he/she is not aware, or include those elements of another's writing of which he/she is not cognizant [4]. Thus the task of repudiation detection comes down to finding the discriminating elements of which, the writer is not aware of. We propose a framework (see Figure 2) that exploits the statistical similarity between lower level feature distributions in two documents to detect possible cases of repudiation. One needs to add a line of caution here that many of the clues that are used by forensic experts comes from external sources (such as background of the suspect, examination of paper material, etc.) and are not available to an automatic writer verification system. Hence any such system can only be used as an aid to a forensic expert, and not a replacement.

The prior work in this area primarily concentrates on the problems of natural handwritten documents.Comprehensive survey of work untill 1989 has been given in [5]. Based on previous research in the field, we can categorize it as text independent vs text dependent, identification vs verification, and online vs offline handwriting. One major class of text independent methods consider handwriting as texture and use various texture analysis methods, such as multi resolution gabor filter [6]. Another class of text independent methods consider handwriting as a psychomotor process and human being as a perpetual motion genertor. Schomacker et al. [7], presented a method based on density functions of lower level features extracted from connected components in handwritten images for writer identification. However, this approach requires a large amount of handwritten data to model the writer, which is not available in the case of QDE.

Macro features (such as grey level distributions, grey level threshold, contour connectivity, contour slope and slant of characters) and micro features (such as, gradient, structural and concavity) [8] are popular for writer identification and verification with natural handwriting. Such features are possible candidates for repudiation detection as well. However, in the case of QDE, as the data available is limited, it is difficult to calculate the distributions of such features, robustly.

Velocity based approaches, like FIR response of Fourier coefficients of velocity [9] and pen pressure based approaches [10] have also been tried in context of online documents. However, in the case of repudiation detection, the questioned document might not be online. Hence, features based only on velocity or pen pressure can not be used. Approaches that uses text line based features with HMM [11] to statistically model handwriting of a writer, also extract high level features and thus is not appropriate for QDE.

The work proposed by Srihari et al. [12], for writer verification using statistical inference can be used to calculate the significance of distance between a given pair of documents. In case of forgery, most of the work is done in physical biometrics [13] as
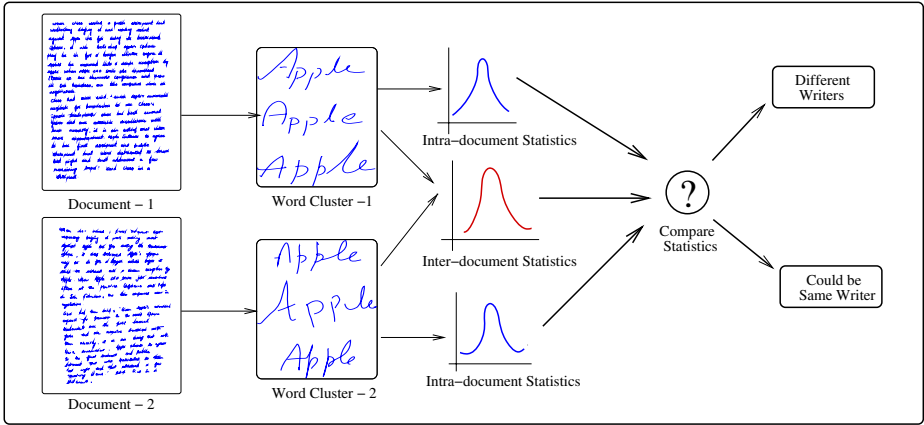
Fig. 2. Framework for detecting repudiation from handwriting

well as signature verification [14]. We are not aware of any work done on repudiation of handwritten documents till date. The approaches used for identification and verification can not be directly applied to repudiation, owing to the additional constraints involved. Moreover, all of the above approaches used for generic writer verification and identification extract discriminative elements of handwriting that can be used to discriminate between writers, while in repudiation detection, the major goal is to detect the underlying similarity between documents.

The remainder of this paper is organized as follows. The overall framework used in our solution is presented in Section 2. In Section 2.2, we will discuss implementation details with feature vector extraction. The experimental results and discussions are presented in section 3, followed by the conclusions and future work.

2   A Framework for Repudiation Detection

This section describes a generic framework for repudiation detection for questioned document examination. The primary goals of the framework are:

1. To develop a statistically significant matching score between two documents, without any additional information in the form of training data.
2. Utilize the online information that could be obtained from the reference document to improve the matching.
3. Allow the inclusion of additional features that might be extracted from the handwriting to enhance the results. This would also mean, we should not make specific assumptions about the distributions of the features, in the framework.
4. Allow the user to specify a confidence threshold, beyond which, the system will pass the documents for expert examination.

We also make the certain assumptions in our approach. The primary assumption is that the content of the questioned document and the reference document are either the

same or has significant overlap at the word level. This allows us to use text-dependent approaches to compare the words in the two documents. We also assume that the reference document is collected in the online mode (with temporal information). These assumptions are valid in the case of QDE, since the investigator can control the content and mode of the reference document being collected.

Let the two documents under consideration, questioned and reference, be denoted as $Q = \{w_i; i = 1 \cdots n_q\}$ and $R = \{w_j; j = 1 \cdots n_r\}$, where w_is are the individual words and the documents, containing n_q and n_r words respectively. Let us assume that the number of distinct words in each document is N_q and N_r respectively. The documents are then partitioned into equivalence classes, C_q^k and C_r^k, based on the words as follows:

$$C_q^k = \{w_i \in D_q | w_i \text{ is the word } k\}$$

$$D_q = \bigcup_{k=1}^{N_q} C_q^k \qquad (1)$$

This partitioning can be done using recognition-based or ink-matching based techniques. We then compute the correspondence between the sets in C_q and C_r, based on the words they denote. Note that we have assumed that there is a significant overlap between the sets, and hence some of the sets will match. Once again, we can find the correspondence based on recognition results or ink matching. Without loss of generality, we assume that the corresponding sets are C_q^k and C_r^k, $k = 1 \cdots K$.

To compute the similarity between the two documents, we first define a distance measure between two corresponding words, w_i and w_j, $d(w_i, w_j)$. This distance could be based on any set of features that are extracted from the words. We compute two distributions from these distances: i) p_w, the *within document distance distribution*; from the distances of corresponding words within the same document, and ii) p_b, the *between document distance distribution*; from the distances of corresponding words, one each from Q and R.

We now pose the computation of similarity between the two documents as that of testing the hypothesis, whether the two distributions, p_w, and p_b, come from the same population or not. The distance between the two distributions could be used in deciding, whether the two documents come from the same writer or not. We now look into the details of the distance measures and their statistical significance.

2.1 Detecting Repudiation and Forgery

One of the major problem in the case of verification by matching, is to assign a significance or confidence to the distance measure, computed between two documents. Traditionally, this is done based on a threshold, which in turn is calculated from training data. In the case of forensic documents, where training data is not available, we cannot rely on a threshold, which might be different for different writers. Hence we use a hypothesis testing based approach that computes the significance of the decision, taken at a particular level of confidence. It provides us with a formal means for verifying whether the two sample distributions, p_w and p_b come from the same population

or not. We start with the assumption that the two documents are indeed from the same writer and test the validity of this hypothesis. If the hypothesis is rejected with sufficient confidence, we can eliminate the suspect from consideration. The two-class hypothesis testing problem for forensic documents can be posed as:

$$H_0 = \text{Documents written by same writer}$$
$$H_1 = \text{Documents written by different writers,}$$

$$(2)$$

where H_0 is referred to as the null hypothesis, and H_1, the alternate hypothesis.

Several approaches have been used to compute the similarity between two distributions, the most popular of which are the Kullback-Leibler(KL) divergence or Kolmogorov-Smirnov(KS) test. Kullback-Leibler divergence or relative entropy measures the natural distance between a true probability distribution, P, and an arbitrary distribution, Q. For probability distributions P and Q of discrete variables, the KL divergence (or informally KL distance) of Q from P is given by:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}; \text{ and}$$
$$P_{KL} \qquad = e^{-\xi D_{KL}},$$

$$(3)$$

where P_{KL} gives the probability that the samples of the two distributions are drawn from the same population. Kullback Leibler distance essentially calculates divergence between distributions, and is not a distance metric, as it is neither symmetric nor satisfies triangle inequality. On the other hand, KS test determines whether whether an underlying probability distribution differs from a hypothesized distribution, based on finite samples. The KS-test also has the advantage of not making assumptions about the distribution of data and hence is non parametric. The two parameter KS-test is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples. KS test computes a simple distance measure, given by:

$$D_{KS}(P||Q) = \max_i |P(i) - Q(i)|,$$

$$(4)$$

where P and Q are the cumulative probability distribution functions and $P(i)$ and $Q(i)$ are corresponding probability values. The distance D_{KS} is, hence, the maximum absolute difference of cumulative probability on all potential values of i. The probability of similarity between two distributions is then calculated by:

$$P_{KS} = Q_{KS}(\sqrt{N_e} + 0.12 + (\frac{0.11}{\sqrt{N_e}})D_{KS})$$
$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2}|Q_{KS}(0) = 1, Q_{KS}(\infty) = 0,$$

$$(5)$$

where $N_e = N_1 N_2 (N_1 + N_2)^{-1}$ is effective number of data points (N_1 and N_2 are number of data points in two distributions respectively). The major limitation of the KS test is that it is more sensitive near the center of the distribution than at the tails. In

this paper, we have used the KS test in all the experiments. A distance metric based on a combination of KL test and KS test explained in [15], and can be used to get some improvement in the results.

The formulation of the comparison using hypothesis test, makes an implicit assumption that any two documents from the same writer will have identical distribution of features. However, this is not true due to two factors: i) natural handwriting of two documents from a writer tends to be different due to environmental and physical conditions of the writer, and ii) in case of repudiation and forgery, the writer deliberately introduces some variations even if appropriate features are extracted. Hence we modify the hypothesis test result by looking at the confidence level of the result, and choosing a threshold, α, on the confidence to decide if we should involve an expert or not.

2.2   Comparison of Words

In case of repudiated documents, the extraction of appropriate features play an important role in deciding whether we can effectively deal with deliberate distortions of handwritten data. Based on the level of details, the discriminating features of handwriting can be divided as macro level or high level features and micro level or low level features. High level features, such as alignment, slope, slant of line and words can be altered with relative ease, as the person is quite aware of these features and thus can be changed forcefully. However, lower-level features such as shape and size of primitive curves and connection between these curves can not be changed easily, as the person is habituated to write these primitive curves over a long period of time.

Although the primitive curves are used in computing the distance measure, we use words as the basic units of comparison. The words are natural units of writing and contain both primitive curves as well as their interconnections. Another major reason for choosing words as our comparison unit is that the same character is written quite differently (with respect to shape and size) within different words and this will introduce large intra class variations at character level. Individual words are segmented and clustered into clusters of same words using automatic clustering and segmentation methods. Simple features such as lower and upper profiles of the word are used to partition the documents into clusters of same words. Errors in data clustering and segmentation can be removed manually, as in case of forensic documents, manual intervention is possible as volume of data is small.

The distance between a pair of words is calculated using low-level features such as shape and size of the constituent primitive curves [16]. The primitive curves from the words are extracted by dividing the handwritten data at certain critical points in the pen trace. The critical points are often defined as the points of maximum curvature in the pen trace. However, computation of these critical points is error-prone in offline documents. We utilize the possibility of collecting online reference data to alleviate this problem. For online handwriting (reference document), the dominant points are defined based on the maximum and minimum velocity points. These dominant points form a robust basis for extraction of primitive curves. Although, the velocity of handwriting can change with change in environmental conditions, or be changed deliberately; it is observed that the critical points of velocity remains the same. Figure 3 shows the primitive extraction and comparison process, where two words (*apple*) written by the same person from both
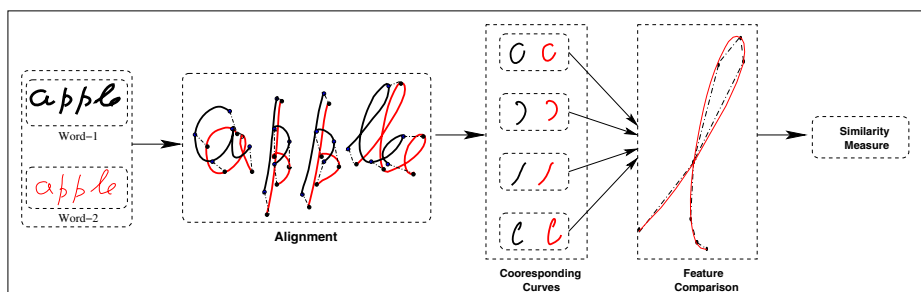
Fig. 3. Comparison between two words '*apple*'

natural and repudiated handwriting are compared. The primitive curves are defined as the portion of curves between three consecutive minimum velocity points on the stroke. Note that in case both documents are offline, critical points can be calculated using curvature, or can be computed by matching the curves to the corresponding online word.

Distance between a pair of words is calculated using a dynamic time warping based matching process. Each word is represented as a feature matrix of size $m \times n$, where, m is the number of curves in a word, and n is size of features extracted from each curve. Curves are represented using their curvature, size of connecting vectors, and higher order moments to capture the relative velocity and shape. Euclidian distance is used to calculate distance between two primitive curves. The method used is relatively simple, and could be replaced with a more comprehensive distance measure that uses various properties that are extracted from the curves.

3   Experimental Results and Analysis

The data used in our experiments was collected from 23 different writers. Each writer was asked to write three pages on A5 sized sheets in his/her own natural handwriting. In addition, three pages of data was collected from each writer, while trying to masquerade his/her handwriting style. The data was collected using iball take-note, which collects the data in both on-line and off-line forms. The data is then segmented into words and clustered into groups of same words.

As noted before, the actual significance of the distance between two documents cannot be used directly as it is based on the assumption that two documents from a writer will have identical distributions of features. Hence, a threshold on the confidence of the decision need to be identified, below which, we use the services of an expert. To present the capabilities of the system, we plot the ROC curve of the system by varying the threshold. Figure 4 shows ROC curve and the corresponding distributions of within-writer and between-writer distance distributions. The document pairs that are written by same person is considered as genuine documents. Note that this includes repudiated documents from the same writer. The genuine accept rate is the rate of acceptance (or matching) of documents that are written by same person and false accept rate is percentage of documents that are considered as matching, when they actually belong to different writers. ROC curve shows that about 82% of documents which belong

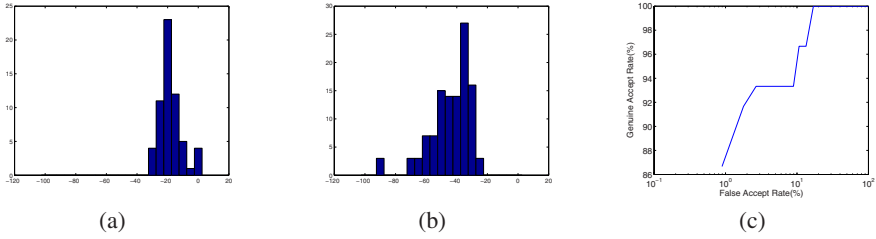(a)   (b)   (c)

Fig. 4. Histogram of Inter-writer and Intra-writer distances and the ROC curve of the system

to different writers are rejected, while keeping the genuine acceptance rate at 100%. As discussed before, this step is considered as the preliminary step for the document screening before it goes to expert, and hence we need to operate at a point, where the automatic decisions made by the system should be highly accurate. All the documents, which are not rejected can be processed further by handwriting expert.

An alternate way of presenting the results of matching a particular document pair to an expert is on the traditional nine-point scale. Forensic experts use this scale to indicate the level of match between two documents under consideration. The scale consists of: *identification*, *strong probability of identification*, *probable*, *indications*, *no conclusion*, *indications did not*, *probably did not*, *strong probability did not* and *elimination*. We can present a similar result based on the densities in the corresponding histograms in figure 4. However, due to the bias introduced by hypothesis testing (tests are done under the assumption that null hypothesis is true), the results will be confined to the values of *no conclusion*, *indications did not*, *probably did not*, *strong probability did not* and *elimination*, in the case of repudiation.

4   Conclusions and Future Work

We have introduced the problem of repudiation in handwritten documents, which is particularly relevant for forensic document examination. A statistical model for automatic repudiation (and forgery) detection, which uses the statistical significance of the distance between two distributions, is presented. Preliminary results support the validity of the model. Such an automated system can act either as a screening mechanism for questioned documents, or could provide additional insights to an expert examiner of the documents.

Preliminary investigations into the use of the model for detecting forgeries seem to be promising. However, we need to conduct extensive experiments using expert forgeries to make conclusive statements on the effectiveness. One can also experiment with a variety of features to compute the distance between two words, in order to improve the matching results.

References

1. Morris, R.: Forensic Handwriting Identification: Fundamental concepts and principles. Academic Press, London (2000)
2. Srihari, S., Huang, C., Srinivasan, H., Shah, V.: Biometric and Forensic Aspects of Digital Document Processing. In: Digital Document Processing, pp. 379–405. Springer, London (2007)

3. Fogarolo, L.: Questioned document examination using graphology (2007), http://www.graphology.it/forensic/
4. Huber, R., Headrick, A.: Handwriting Identification: Facts and Fundamentals. CRC Press, Boca Roton (1999)
5. Plamondon, R., Lorette, G.: Automatic signature identification and writer verification - the state of the art. Pattern Recognition 22(2), 107–131 (1989)
6. He, Z., Tang, Y.: Chinese handwriting based writer identification by texture analysis. In: Proceedings of International Conference on Machine Learning and Cybernetics, Shanghai, vol. 6, pp. 3488–3491 (2004)
7. Schomaker, L., Bulacu, M.: Text-independent writer identification and verification using textural and allographic features. IEEE Transactions on Patterh Analysus and Machine Intelligence, Special Issue - Biometrics: Progress and Directions 29(4), 701–717 (2007)
8. Tomai, C., Zhang, B., Srihari, S.: Discriminatory power of handwritten words for writer recognition. In: Proceedings of Int'l conf. of Pattern Recognition, Cambridge, UK, vol. 2, pp. 638–641 (2004)
9. Thumwarin, P., Matsuura, T.: On-line writer recognition for thai based on velocity of barycenter of pen-point movement. In: Proceedings of IEEE Int'l conf. of Image Processing, Singapore, pp. 889–892. IEEE Computer Society Press, Los Alamitos (2004)
10. Yu, K., Wang, Y., Tan, T.: Writer identification using dynamic features. In: Proceedings of Int'l conf. on Biometric Authentication, Hong Kong, China, pp. 512–518 (2004)
11. Schlapbach, A., Bunke, H.: A writer identification and verification system using hmm based recognizers. Pattern Analysis & Applications 10(1), 33–43 (2007)
12. Srihari, S., Bandi, K., Beal, M.: A statistical model for writer verification. In: Proceedings of the Int'l. conf. of Document Analysis and Research, Seoul, Korea, pp. 1105–1109 (2005)
13. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology. Special Issue on Image- and Video-Based Biometrics 14(1), 4–20 (2004)
14. Jain, A.K., Griess, F., Connell, S.D.: On-line signature verification. Pattern Recognition 35(12), 2963–2972 (2002)
15. Press, W., Flannery, B., Teukolsky, S., Vetterling, W.: Numerical Recepies in C: The Art of Scientific Computing. Cambridge University Press, Cambridge (1992)
16. Namboodiri, A.M., Gupta, S.: Text independent writer identification from online handwriting. In: Proceedings of Int'l workshop of Frontier in Handwriting Recognition, La Baule, Centre de Congreee Atlantia, France, pp. 23–26 (2006)