A Human-Centric Off-Line Signature Verification System

Hanno Coetzer
Department of Mathematical Sciences
University of Stellenbosch
Stellenbosch, South Africa
jcoetzer@sun.ac.za

Robert Sabourin
École de Technologie Supérieure
Université du Québec
Montréal, Canada
Robert.Sabourin@etsmtl.ca

Abstract

The manual signature-based authentication of a large number of documents is a laborious and time-consuming task. Consequently many off-line signature verification systems were recently developed. In this paper we propose a human-centric system, which exploits the synergy between human and machine capabilities, and show that this combined system can perform better (than humans or a machine) for almost all operating costs. The combination strategy is based on techniques in receiver operating characteristics (ROC) analysis. We conduct an experiment on a data set that contains 765 test signatures from 51 writers, and record the performance of 23 human classifiers, and that of a hidden Markov model-based (HMM-based) classifier, in ROC space. We propose that a manager (human or machine) specifies acceptable operating costs (Neyman-Pearson criterion), after which our human-centric system makes an optimal decision by utilizing the maximum attainable combined classifier.

1. Introduction

A large driving force behind the development of off-line signature verification systems is the financial benefits that the automatic clearing of cheques may have for the banking industry. Surveys of existing systems can be found in [2, 4, 6]. Banks still process millions of cheques daily. Usually, only those cheques of which the amount exceeds a certain threshold are verified manually by an operator. This is a cumbersome process that has to be completed within a limited time. In certain respects however, a human's ability to recognize patterns is superior to that of a machine. This include near instantaneous global level perception, learning from experience, novelty detection, etc. On the other hand, machines are far superior when it comes to processing speed, the management of large data sets, consistency and the ability to operate in hazardous or hostile environments. It is therefore reasonable to investigate the feasibility of a classifier with most of the above-mentioned capabilities.

In this paper we investigate the possibility of enhancing the performance of an existing off-line signature verification system [2] by also utilizing proficient human operators. The feature extraction method for this automated system is based on the calculation of the Radon transform of a signature image and each writer's signature is modelled by a ring-structured HMM. We first evaluate the performance of the above-mentioned HMM-based classifier, as well as the performance of 23 human classifiers, by conducting an experiment on a data set that contains 765 test signatures (432 genuine signatures and 333 skilled forgeries) from 51 writers. We then propose a strategy for utilizing both the human and machine classification effort by considering the best attainable combined classifier.

2. Performance evaluation in ROC space

Given a two-pattern classifier and an instance (e.g. a signature), there are 4 possible outcomes. If a positive instance (e.g. a genuine signature) is classified as positive, the outcome is "true positive" and if it is classified as negative, the outcome is "false negative". If a negative instance (e.g. a forgery) is classified as negative, the outcome is "true negative" and if it is classified as positive, the outcome is "false positive". When the number of instances for which the outcomes are true positive, false negative, true negative, and false positive, are denoted by T^+, F^-, T^-, and F^+ respectively, we can estimate the true positive rate (TPR) for the classifier as TPR $= \frac{T^+}{T^+ + F^-}$, and the false positive rate (FPR) as FPR $= \frac{F^+}{T^- + F^+}$. The two-dimensional space with the FPR on the horizontal axis and the TPR on the vertical axis is called the *ROC space*. A *discrete classifier* (e.g. a human being) produces discrete output (true or false) and the performance of such a classifier is therefore depicted by a single point in ROC space. The HMM-based classifier,

that we implement in this paper, is a *continuous classifier*, since it produces continuous output (a score in this case) to which different decision thresholds can be applied to predict class membership. The performance of such a classifier is depicted by an ROC curve, i.e. a curve in which the TPR is plotted on the vertical axis and the FPR on the horizontal axis for a number of discrete threshold values. An ROC curve therefore depicts relative trade-offs between benefits (true positives) and costs/risks (false positives).

In the signature verification context genuine signatures and forgeries are considered to be positive and negative instances, respectively. The TPR and FPR can be written in terms of the false rejection rate (FRR) and false acceptance rate (FAR) as follows, TPR $= 1 -$ FRR, FPR $=$ FAR.

The performance of classifiers can therefore be represented, analyzed and compared in ROC space.

3. Maximum attainable classifiers

3.1. Combining discrete classifiers in ROC space

In this paper we consider the classifier combination strategy proposed in [5]. Suppose that we denote the FPR and TPR for a discrete classifier, C_X, by f_X and t_X, respectively. The performance of two discrete classifiers, C_P and C_Q, are therefore represented by the points, (f_P, t_P) and (f_Q, t_Q), in ROC space. When an instance (signature) is to be authenticated, both C_P and C_Q will output either positive $(+)$ or negative $(-)$, giving 4 possible scenarios. For each of these 4 scenarios one can obtain an expression for the maximum likelihood estimation (MLE) of the unknown truth T, assuming that the two classifiers are conditionally independent (see [5]). These expressions are given in Table 1.

C_P	C_Q	Combined MLE of truth T
$+$	$+$	$t_P t_Q \geq f_P f_Q$
$+$	$-$	$t_P(1-t_Q) \geq f_P(1-f_Q)$
$-$	$+$	$(1-t_P)t_Q \geq (1-f_P)f_Q$
$-$	$-$	$(1-t_P)(1-t_Q) \geq (1-f_P)(1-f_Q)$

Table 1. Binary output for classifiers C_P and C_Q and the maximum likelihood combination.

This results in one of 4 possible MLE combination schemes, that we shall call scheme $S_{P\&Q}$ (P *and* Q), scheme S_P, scheme S_Q, and scheme $S_{P\|Q}$ (P *or* Q). These schemes are summarized in Table 2.

Using the assumption of conditional independence, it is possible to calculate attainable FPRs and TPRs for the above-mentioned combination schemes as in Table 3 (see [5]). We denote the FPR and TPR for the combined classifier C_R by f_R and t_R, respectively.

C_P	C_Q	$S_{P\&Q}$	S_P	S_Q	$S_{P\|Q}$
$+$	$+$	$+$	$+$	$+$	$+$
$+$	$-$	$-$	$+$	$-$	$+$
$-$	$+$	$-$	$-$	$+$	$+$
$-$	$-$	$-$	$-$	$-$	$-$

Table 2. Binary output for classifiers C_P and C_Q and the 4 schemes for combining them.

Scheme	f_R	t_R
$S_{P\&Q}$	$f_P f_Q$	$t_P t_Q$
S_P	f_P	t_P
S_Q	f_Q	t_Q
$S_{P\|Q}$	$f_P + f_Q - f_P f_Q$	$t_P + t_Q - t_P t_Q$

Table 3. Attainable FPRs and TPRs for the combined classifier C_R using the 4 different combination schemes.

These rates, i.e. f_R and t_R, can therefore be calculated using only the points (f_P, t_P) and (f_Q, t_Q) in ROC space.

3.2. Combining continuous classifiers in ROC space

Suppose that the performance of two continuous classifiers, C_P and C_Q, are represented by two ROC curves, $\mathbf{R}_P(j) = (f_P(\phi_j), t_P(\phi_j)),\ j = 1, \ldots J$ and $\mathbf{R}_Q(k) = (f_Q(\psi_k), t_Q(\psi_k)),\ k = 1, \ldots K$, where $\phi_j,\ j = 1, \ldots J$ and $\psi_k,\ k = 1, \ldots K$ denote the selected threshold values. It is now possible to construct JK new discrete classifiers by combining C_P and C_Q using the strategy discussed in Section 3.1. It is easy to extend this merging strategy to any number of continuous and/or discrete classifiers.

3.3. The maximum attainable ROC curve

Given the original (discrete) classifiers, and the combined (discrete) classifiers, it is possible to calculate the *maximum attainable ROC (MAROC) curve*. In this paper however, i.e. for our *human-centric* system, we *only* utilize the combined classifiers to calculate the MAROC curve. This is done in order to ensure that at least one human classifier, and at least one HMM-based classifier, are *always* involved in the decision process (see Section 7).

Suppose that (f_P, t_P) and (f_Q, t_Q) represent the performance of two combined classifiers, C_P and C_Q, and that L_{PQ} represents the line segment that links these two points in ROC space. It can be shown [7] that any point on L_{PQ} represents the performance of a classifier that can be *attained* by randomly sampling between the outputs of C_P and C_Q. Consequently, given several ROC curves and the performance of several discrete classifiers (points in ROC space), the convex hull of all of these points represents the performance that can be attained by randomly sampling between their outputs. Since the performance of a classifier C_P is deemed superior to that of C_Q when $f_P < f_Q$ and $t_p > t_Q$, the points on the top-left boundary of the convex hull represent the MAROC curve.

3.4. Classification in variable cost domains

In many real-world scenarios, e.g. where a large number of cheques have to be authenticated by commercial banks, the cost/risk associated with different types of errors is not known when the classifier is being designed and can change from time to time, or even from instance to instance. In these situations the designer often resorts to specifying the performance of the (combined) classifier in the form of an adjustable threshold and an MAROC curve. A Neyman-Pearson criterion is usually specified. This criterion requires that a manager specifies the *maximum allowable FPR*, denoted by $\text{FPR}_{\max}$. The manager then choose the point on the MAROC curve with the highest TPR and with a FPR less than or equal to $\text{FPR}_{\max}$. This point represents the performance of the *maximum attainable combined classifier*. We graphically illustrate the above concepts with a real-world example in Section 7.

4. Dolfing's data set

We conduct an experiment (Section 5) on signatures that are randomly selected from a data set that was originally captured on-line for Hans Dolfing's Ph.D. thesis [3]. Dolfing's data set contains 4800 signatures from fifty-one writers. Each of these signatures contains static and dynamic information captured at 160 samples per second. Each of these sample points contains information on pen-tip position, pen pressure, and pen tilt. Static signature images are constructed from this data using only the pen-tip position, that is the x and y coordinates, for those sample points for which the pen pressure is non-zero. A more detailed discussion of this signature acquisition method can be found in [2]. In the subsequent experiment we only consider skilled forgeries from Dolfing's data set (see Figure 1 (b), (c) and (d)). A *skilled forgery* is produced when the forger has unrestricted access to samples of the writer's actual signature.

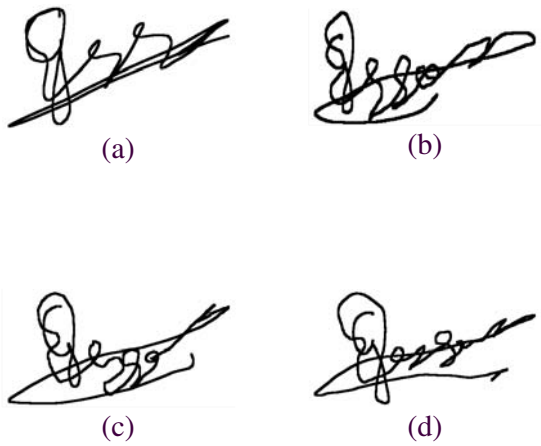

Figure 1. Signatures from Dolfing's data set. (a) Genuine signature. (b)-(d) Skilled forgeries.

For each writer in Dolfing's data set, there are 15 training signatures, 15 genuine test signatures and 60 skilled forgeries available, with the exception of two writers, for whom only 30 skilled forgeries are available.

5. Experimental protocol

For each of the 51 writers in Dolfing's data set we construct a test set that consists of *only* 15 signatures. For each of the 51 writers, all the available training signatures (15 per writer) are used. Each test set contains a randomly selected number (any number between 0 and 15) of skilled forgeries. The remaining test signatures are randomly selected from the 15 genuine test signatures for the writer in question. It is therefore possible that a specific test set contains *only* genuine test signatures or *only* skilled forgeries. A verifier (human or machine) is therefore presented with a total of $15 \times 51 = 765$ test signatures. The *total* number of genuine test signatures and forgeries turns out to be 432 and 333 respectively.

Human verification. Twenty-three human beings are each presented with the signatures from all 51 writers in the data set. These human classifiers consist of faculty members and graduate students. We present each individual with a training set (15 signatures) and a corresponding test set (15 signatures) for all 51 writers. The training set and the corresponding test set for a specific writer are presented on

two separate sheets of paper. A human being typically compares the test signatures, as a unit, with the corresponding training set and then decide which of the test signatures to reject. Each individual human classifier was instructed not to ponder over a decision, so as to simulate what a bank official is likely to do.

Machine verification. The same training and test signatures, that are considered for human verification, are also considered for machine verification (see [2]).

6. Results

When the HMM-based system (continuous classifier), that is proposed in [2], is implemented on randomly selected test signatures (see Section 5) from Dolfing's data set, the ROC curve (solid line with bullets) in Figure 2 is obtained. (The system achieves an equal error rate of approximately 12%.) The performance of the 23 humans (discrete classifiers) are indicated by circles in ROC space.

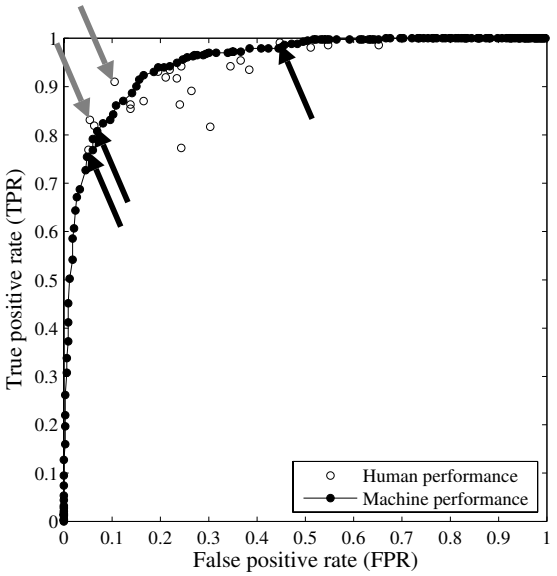

Figure 2. The performance of the 23 human classifiers and the HMM-based classifier.

In Figure 2 it can be seen that only 5 human classifiers, i.e. the circles above the ROC curve indicated by arrows, perform better than the HMM-based classifier. Only 2 of these classifiers, i.e. the circles above the ROC curve indicated by *grey* arrows, perform significantly better. By conducting an appropriate significance test, it can be shown [1] that the HMM-based classifier outperforms a "typical" human being.

7. Example: The banking environment

We assume that a number of bank officials (who undergo periodic proficiency tests) are responsible for signature-based cheque authentication, and that an automated system is available to assist them. During each proficiency test, the officials have to authenticate a number of signatures according to the protocol discussed in Section 5. We also assume (for the purpose of this discussion) that the human performance, reported in Section 6, is indicative of the bank officials' performance for a specific proficiency test.

The HMM-based classifiers (each one associated with a different threshold value) and the human classifiers can now be combined using the strategy discussed in Sections 3.1 and 3.2 (see Figure 3). The attainable performance of the combined classifiers are denoted by grey bullets, while the maximum attainable combined performance is represented by the MAROC curve (solid line with squares).

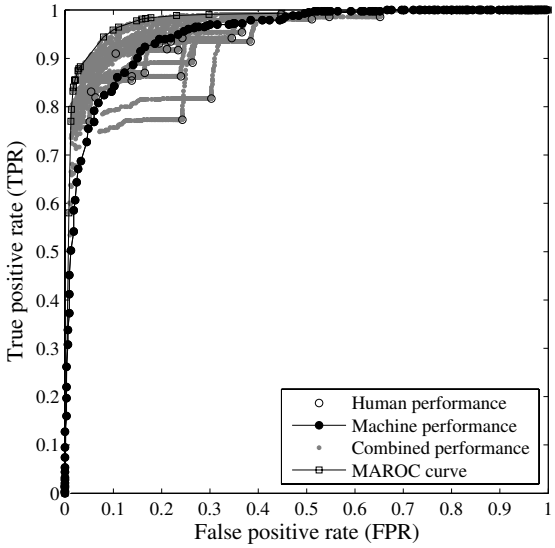

Figure 3. The maximum attainable combined performance of the 23 human classifiers and the HMM-based classifier.

From Figure 3 it is clear that the maximum attainable combined classifiers outperform the HMM-based classifier, and the most proficient human classifiers, for most operating costs. A manager (human or machine) may, for example, specify $\mathrm{FPR}_{\max}$ according to the amount of the cheque. For larger amounts, $\mathrm{FPR}_{\max}$ may be lowered. This implies that a lower threshold is applied, which makes it more difficult to accept the cheque.

Suppose that $\mathrm{FPR}_{\max}$ is fixed at 0.135 as indicated by

the vertical grey line in Figure 4. The best possible performance of an existing HMM-based classifier is given by (f_A, t_A), where $t_A = 0.870$, and that of a human by (f_B, t_B), where $t_B = 0.910$. It is however possible to employ the strategy proposed in Section 3.1 to obtain a superior combined performance of (f_C, t_C), where $t_C = 0.965$. This performance is attainable by combining the outputs of an HMM-based classifier and a human classifier, of which the respective performances are given by (f_D, t_D) and (f_E, t_E). The relevant signature is simply displayed on a computer screen, after which the above-mentioned human classifier is prompted to make a decision. It is possible to further improve the combined system's performance by also prompting a second human classifier. According to [7] a classifier with a performance of (f_H, t_H), where $t_H = 0.973$, can be attained by randomly sampling between the outputs of two combined classifiers, of which the respective performances are given by (f_C, t_C) and (f_F, t_F). This implies that the outputs of two HMM-based classifiers and two human classifiers, of which the respective performances are denoted by (f_D, t_D), (f_E, t_E), (f_G, t_G), and (f_B, t_B), are combined.

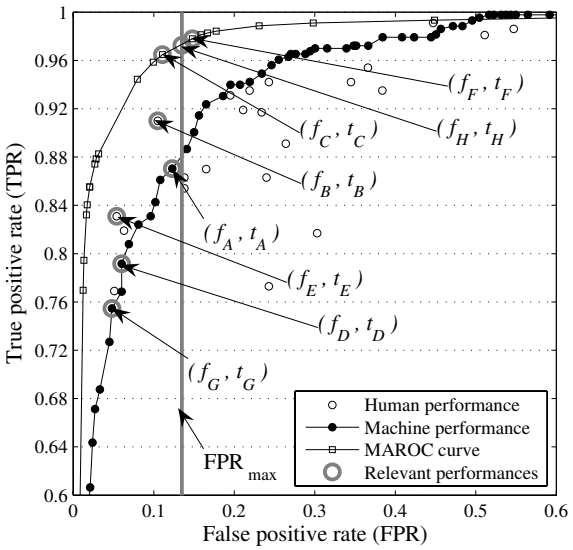

Figure 4. Strategy for combining the human and machine classification effort.

8. Conclusion and future work

In this paper we investigated the feasibility of utilizing both human and machine classifiers for the purpose of signature verification in a banking environment. By conducting an experiment on 765 off-line signatures we clearly demonstrated that it is possible to obtain combined classifiers that outperform the most proficient humans classifiers, as well as an HMM-based classifier, for most operating costs. It is advisable to show the statistical significance of the above-mentioned improvement in performance by also conducting a McNemar test. Due to time constraints, this was not carried out. We assumed that the performance of the human classifiers (academics and students), that was reported in Section 6, can be generalized to bank officials. Although it is reasonable to assume that bank officials are more proficient in authenticating cheques, the principles demonstrated in this paper still apply.

References

[1] J. Coetzer, B.M. Herbst and J.A. du Preez, "Off-line Signature Verification: A Comparison Between Human and Machine Performance", *Proc. 10th Int'l Workshop on Frontiers in Handwriting Recognition*, La Baule, France, 2006, pp. 481-485.

[2] J. Coetzer, B.M. Herbst and J.A. du Preez, "Offline Signature Verification Using the Discrete Radon Transform and a Hidden Markov Model", *Eurasip Journal on Applied Signal Processing - Special Issue on Biometric Signal Processing,* H. Bourland, I. Pitas, K.K. Lam, and Y. Wang, eds., 2004, vol. 2004, no. 4, pp. 559-571.

[3] J.G.A. Dolfing, "Handwriting Recognition and Verification. A Hidden Markov Approach.", *Ph.D. Thesis*, Eindhoven University of Technology, 1998.

[4] J.K. Guo, D. Doermann, and A. Rosenfeld, "Forgery Detection by Local Correspondence", *Int'l J. Pattern Recognition and Artificial Intelligence*, 2001, vol. 15, no. 4, pp. 579-641.

[5] S. Haker, W.M. Wells III, S.K. Warfield, I. Talos, J.G. Bhagwat, D. Goldberg-Zimring, A. Mian, L. Ohno-Machado, and K.H. Zou, "Combining Classifiers Using Their Receiver Operating Characteristics and Maximum Likelihood Estimation", *Medical Image Computing and Computer-Assisted Intervention*, 2005, vol. 3749, pp. 506-514.

[6] R. Plamondon and S.N. Srihari, "On-Line and Off-Line Handwriting Recognition: A Comprehensive Survey", *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2000, vol. 22, no. 1, pp. 63-84.

[7] M.J.J. Scott, M. Niranjan, and R.W. Prager, "Realisable classifiers: Improving operating performance on variable cost problems", *Proc. 9th British Machine Vision Conf.*, P.H. Lewis and M.S. Nixon, eds., Southampton, 1998, vol. 1, pp. 305-315.