Efficient cost-sensitive human-machine collaboration for off-line signature verification

Johannes Coetzer[a], Jacques Swanepoel[b] and Robert Sabourin[c]

[a,b] Department of Mathematical Sciences, Stellenbosch University, Stellenbosch, South Africa;
[c] École de Technologie Supérieure, University of Quebéc, Montréal, Canada

ABSTRACT

We propose a novel strategy for the optimal combination of human and machine decisions in a cost-sensitive environment. The proposed algorithm should be especially beneficial to financial institutions where off-line signatures, each associated with a specific transaction value, require authentication. When presented with a collection of genuine and fraudulent training signatures, produced by so-called guinea pig writers, the proficiency of a workforce of human employees and a score-generating machine can be estimated and represented in receiver operating characteristic (ROC) space. Using a set of Boolean fusion functions, the majority vote decision of the human workforce is combined with each threshold-specific machine-generated decision. The performance of the candidate ensembles is estimated and represented in ROC space, after which only the optimal ensembles and associated decision trees are retained. When presented with a questioned signature linked to an arbitrary writer, the system first uses the ROC-based cost gradient associated with the transaction value to select the ensemble that minimises the expected cost, and then uses the corresponding decision tree to authenticate the signature in question. We show that, when utilising the entire human workforce, the incorporation of a machine streamlines the authentication process and decreases the expected cost for all operating conditions.

Keywords: Human-machine collaboration, classifier combination, off-line signature verification

1. INTRODUCTION

Despite recent advances in the fields of computer vision and machine learning, well-trained human beings are still generally more proficient than machines in recognising most biometric patterns. Humans are especially proficient at recognising patterns that they encounter on a daily basis, like handwritten numerals, speech patterns and facial patterns. Humans are however not so proficient at recognising more intricate patterns, like fingerprints, irises, and retinal patterns. Machines, on the other hand, are able to efficiently process large amounts of data, without the constraints of fatigue or boredom. In the context of off-line signature verification, for example, it has recently been shown[1] that the performance of a hidden Markov model-based (HMM-based) machine is better than, but comparable to, the performance of a typical (amateur) human-being.

In order to exploit the synergy between human and machine capabilities,[2] recent emphasis has shifted towards the concept of utilising a machine to *assist* a human (or workforce of humans) in the decision-making process. Although human-machine interfaces are widely utilised in the fields of aviation[3] and medicine,[4] to our knowledge, human-machine collaboration for the purpose of *biometric* authentication has only been investigated in two recently proposed protocols.[5,6]

The *first protocol*[5] focuses on the recognition of flowers and faces. An abstraction of the object to be recognised is constructed by the machine and superimposed onto an image of the object. A human operator is then able to modify the abstraction and in so doing aids the machine in the feature extraction process. Effective human-machine collaboration therefore occurs on the *feature extraction level*. The authors show that interactive recognition is more accurate than automated classification, faster than unaided human classification, and that both human and machine performance improve with use.

Further author information: (Send correspondence to J. Coetzer)
J. Coetzer: E-mail: jcoetzer@sun.ac.za, Telephone: +27 (0) 21 808 4221
J.P. Swanepoel: E-mail: jpswanepoel@sun.ac.za, Telephone: +27 (0) 21 808 2680
R. Sabourin: E-mail: robert.sabourin@etsmtl.ca, Telephone: +1 (514) 396 8932

Document Recognition and Retrieval XIX, edited by Christian Viard-Gaudin, Richard Zanibbi,
Proc. of SPIE Vol. 8297, 82970J · © 2012 SPIE · CCC code: 0277-786X/12/$18 · doi: 10.1117/12.910460

The *second protocol*[6] investigates the problem that is also addressed in this paper, namely the authentication of static handwritten signatures on cheques (or credit card receipts) in a banking environment. This protocol proposes human-machine collaboration on the *decision level*. It is assumed that a machine and a workforce of human employees are available for signature verification. The proficiency of the machine and the human employees are determined by presenting them with a collection of genuine and fraudulent training signatures, produced by so-called guinea pig writers. By considering the true positive rate (TPR) and false positive rate (FPR) of these classifiers in ROC space, the decision of every human classifier is combined with the decision of every threshold-specific machine-generated classifier using maximum likelihood estimation (MLE).[7] By again considering the signatures from the guinea pig writers, the respective performances of the combined human-machine hybrids can be estimated and represented in ROC space.

A so-called Neyman-Pearson criterion is then specified by an operator (e.g. a bank manager), by stipulating a maximum allowable FPR (denoted by $\text{FPR}_{\max}$) for each transaction, based on its associated monetary value. In a practical, cost-sensitive environment, the financial institution has to (empirically) determine an appropriate mapping between the monetary value (associated with a specific transaction) and $\text{FPR}_{\max}$, in order to minimise the expected cost. When presented with a questioned signature linked to an arbitrary writer, the system first selects the appropriate hybrid classifier, with the highest TPR and with an FPR less than or equal to $\text{FPR}_{\max}$. The corresponding human classifier and threshold-specific machine-generated classifier are then utilised to authenticate the signature. In order to avoid model over-fitting and optimistically biased experimental results, care is taken to ensure that all of the questioned signatures belong to *different* writers than those contained in the guinea pig set. The authors show that the hybrid system outperforms the machine and *all* of the unaided human employees for *all* operating conditions − a higher TPR is therefore achieved for *every* specified value of $\text{FPR}_{\max}$. The aforementioned protocol[6] has several limitations and disadvantages: (1) There is no elegant way in which to map the monetary value, associated with a specific transaction, to an appropriate value of $\text{FPR}_{\max}$, such that the expected cost is minimised; (2) In order to maximise the TPR for all $\text{FPR}_{\max} \in [0, 1]$, only a few (very proficient) human employees are consistently prompted for decisions − many (less proficient) employees are not utilised at all; (3) For every specified value of $\text{FPR}_{\max} \in [0, 1]$, only *one* human classifier, in conjunction with a threshold-specific machine-generated classifier, is utilised.

In this paper we propose a protocol for human-machine collaboration in a cost-sensitive environment, that addresses the above-mentioned inadequacies. We specifically consider the scenario where off-line signatures (on cheques or credit card receipts), each associated with a specific monetary value, have to be authenticated. This protocol is however generic and may be applied to other cost-sensitive scenarios as well.

The remainder of this paper is organised as follows. In the next section we briefly discuss performance evaluation in ROC space. Relevant techniques for combining classifiers in ROC space are presented in section 3. In section 4 we address some key issues pertaining to classification in a cost-sensitive environment. The proposed protocol for human-machine collaboration is discussed in section 5, while section 6 outlines the data set, experimental setup and results. Section 7 concludes the paper and introduces potential future work.

2. PERFORMANCE EVALUATION IN ROC SPACE

Given a two-class classifier (dichotomiser) C_A and an instance (e.g. a signature), there are four 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*. Similarly, 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*. Let the number of instances for which the outcomes are true positive, false negative, true negative, and false positive be denoted by T_A^+, F_A^-, T_A^-, and F_A^+, respectively. The probability that classifier C_A will correctly classify a positive instance can be approximated by its TPR (t_A^+) as follows, $t_A^+ = T_A^+/(T_A^+ + F_A^-)$. Similarly, the probability that classifier C_A will erroneously classify a negative instance as positive can be approximated by its FPR (f_A^+) as follows, $f_A^+ = F_A^+/(T_A^- + F_A^+)$.

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 depicted by a single point in ROC space. The HMM-based classifier, that we

shall consider in this paper, is an example of a *continuous classifier*, since it produces continuous output to which different decision thresholds can be applied to predict class membership. The performance of such a classifier is depicted by an ROC *curve*, where FPR-TPR pairs are plotted for any number of distinct threshold values − each threshold value is therefore associated with a different *discrete* classifier. An ROC curve consequently depicts relative trade-offs between benefits (true positives) and costs/risks (false positives). Classifier performance can therefore be represented, analysed and compared in ROC space. When two different discrete classifiers are compared, the classifier of which the performance is depicted by the more 'northwesterly' point in ROC space is considered to be superior. Similarly, when two continuous classifiers are compared, the classifier with the larger area under its associated ROC curve (AUC) is generally deemed superior.

3. CLASSIFIER COMBINATION IN ROC SPACE

In the context of this paper we only utilise classifier combination on the *decision* level. When presented with the output of multiple discrete classifiers that make conditionally independent errors (e.g. human beings), the most popular classifier combination strategy is majority voting (MV). Conditional independence is important, since it guarantees that the estimated combined performance (as evaluated on a set of guinea pig writers) is a good predictor of future performance (when evaluated on *different* writers).

The output of multiple continuous classifiers (e.g. score-generating machines) can be combined using a strategy like score averaging, but the output of a few poor classifiers often impacts negatively on the combined performance. Alternatively, the discrete classifiers associated with each imposed threshold value can be considered. Since the discrete classifiers associated with a specific machine (depicted by a single ROC curve) invariably make conditionally *dependent* errors, it is ill-advised to consider MV for their effective combination.

An algorithm for combining two ROC curves using MLE was recently proposed.[7] This algorithm combines every discrete classifier on one ROC curve with every discrete classifier on the other. Two discrete classifiers, C_A and C_B, are combined using a set of MLE rules, to select either (1) the output of C_A, (2) the output of C_B, (3) their conjunction ($C_A \wedge C_B$), or (4) their disjunction ($C_A \vee C_B$). This algorithm does however assume that the classifiers are conditionally independent and that their ROC curves are convex. In practice, only the optimal 'northwesterly' section of the convex hull of the combined performances, the so-called maximum attainable ROC (MAROC) curve, is retained. This strategy has proved beneficial in the context of off-line signature verification.[8]

The iterative Boolean combination (IBC) algorithm[9] improves on the aforementioned strategy and requires no prior assumptions on the conditional independence of the classifiers or the convexity of their ROC curves. Two ROC curves are again combined by fusing every discrete classifier on one ROC curve with every discrete classifier on the other, but *ten* different Boolean functions are now considered for combining any pair of discrete classifiers, C_A and C_B, where $\wedge$, $\vee$, $\neg$, and $\oplus$ denote conjunction, disjunction, negation, and the 'exclusive OR' operator, respectively: (1) $C_A \wedge C_B$, (2) $\neg C_A \wedge C_B$, (3) $C_A \wedge \neg C_B$, (4) $\neg(C_A \wedge C_B)$, (5) $C_A \vee C_B$, (6) $\neg C_A \vee C_B$, (7) $C_A \vee \neg C_B$, (8) $\neg(C_A \vee C_B)$, (9) $C_A \oplus C_B$, and (10) $\neg(C_A \oplus C_B)$. Furthermore, this IBC-based approach also allows for the combination of more than two continuous classifiers, each associated with a specific ROC curve. The first two ROC curves are combined, thereby producing a MAROC curve. This MAROC curve is subsequently combined with the third ROC curve, and the process is repeated until the entire set of ROC curves has been considered. The *final* MAROC curve, that depicts the estimated performance of a set of optimal classifier ensembles, where each ensemble is associated with a specific decision tree, can then be utilised to evaluate *unlabelled* data. The authors suggest combining the final MAROC curve with the *original* ROC curves and repeating the process until no gain in proficiency is witnessed. More than one iteration of the algorithm may therefore be beneficial. The fusion protocol proposed in this paper (see section 5) is based on this algorithm.

4. COST-SENSITIVE CLASSIFICATION

In a cost-sensitive environment, the classifier ensemble (on a MAROC curve) with the lowest expected cost is typically selected. This can be achieved by specifying a Neyman-Pearson criterion[6] as explained in section 1. Other strategies include the construction of cost curves[10] and the modification of the actual ROC curve so that it allows for instance-varying costs.[11] The strategy employed in this paper is based on the conventional ROC representation, and utilises *iso-cost* lines with variable gradients, as we now explain.

By considering realistic prior probabilities for questioned signatures in the banking environment, i.e. $P(+) \approx 1$ and $P(-) \approx 0$, one can easily attain an expected cost very close to optimal by simply accepting every questioned signature. This will however render any proposed verification system obsolete. We therefore rather set out from the assumption that the *prior* probabilities of a questioned instance being positive or negative are equal, i.e. $P(+) = P(-) = 0.5$. Human operators are therefore instructed to be as *unbiased* as possible in this regard, while the selection of an optimal ensemble is also based on this assumption. We further assume that there is no cost associated with rejecting a negative instance, nor with accepting a positive instance, i.e. $S(-|-) = S(+|+) = 0$. The expected cost of a transaction authenticated by a classifier C_A can therefore be estimated as follows,[10]

$$E_A = 0.5 \left[S(-|+) \cdot (1 - t_A^+) + S(+|-) \cdot (f_A^+) \right]. \tag{1}$$

With the error costs ($S(+|-)$ and $S(-|+)$) fixed, the line $t^+ = Mf^+ + N(E)$ in ROC space represents the performance of all hypothetical classifiers associated with a *specific* expected cost of E, where $M = S(+|-)/S(-|+)$ and $N(E) = 1 - (2E)/S(-|+)$. For a *specific* cost-ratio M and *different* values of E, different parallel iso-cost lines can be specified. The selection of the classifier ensemble with the lowest expected cost is therefore equivalent to obtaining a line with a gradient of M that intersects a linearly interpolated version of the MAROC curve at a *single* optimal point. Figure 1 (b) illustrates the procedure for selecting the ensemble with the lowest expected cost for three different cost scenarios.

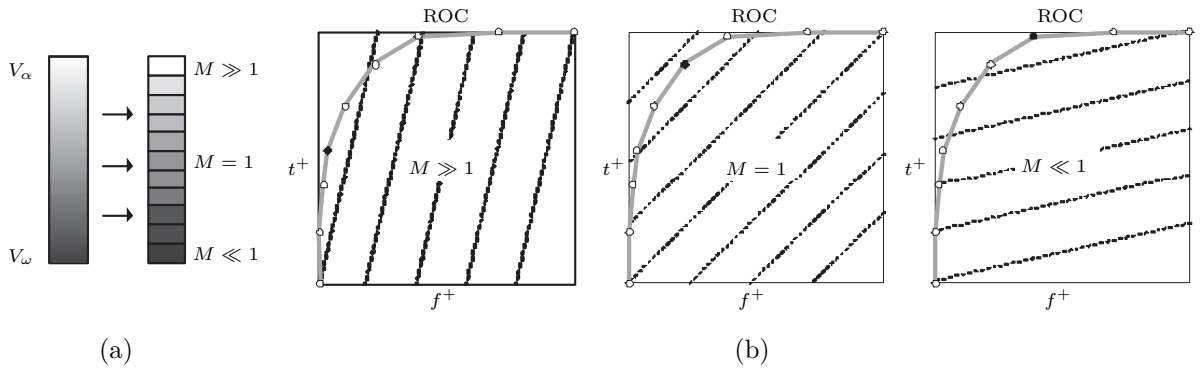


Figure 1. Cost gradient mapping and iso-cost lines. (a) Mapping of a hypothetical transaction value V to a specific cost gradient M. (b) A hypothetical MAROC curve and three different cost scenarios with corresponding iso-cost lines. The performance of the classifier ensemble with the lowest expected cost is denoted by a solid marker.

5. SYSTEM DESIGN

The proposed protocol is summarised in Figure 2. It is reasonable to assume that *each* new client will provide a number of *positive* signature samples during enrollment. The proficiency of a machine and a workforce of human employees can be estimated by presenting them with a collection of genuine *and* fraudulent signatures produced by so-called guinea pig writers − the experimental protocol is discussed in detail in section 6. An estimate of the *combined* performance of the entire human workforce is first obtained by combining the individual human decisions through MV. By again considering the guinea pig signatures, the MV decision is combined with the output of each threshold-specific machine-generated classifier by considering the ten Boolean fusion functions introduced in section 3. A set of candidate classifier ensembles and corresponding decision trees are therefore generated, for which only the MAROC curve is retained. In this way *all* the humans are guaranteed to be included in the decision process.

In consultation with the client, an appropriate mapping between the transaction value $V \in [V_\alpha, V_\omega]$ (in monetary terms) and a finite set of discrete cost gradients $\{M_i\}$, $i = 1, 2, \ldots, K$ is agreed upon, that is $V \mapsto \{M_i\}$ (see Figure 1 (a)). When an unlabelled signature (claimed to belong to a specific writer and associated with a certain transaction value) is introduced, the classifier ensemble (on the MAROC curve) and the corresponding decision tree that is associated with the *lowest* expected cost, are utilised for authentication (see Figure 1 (b)).

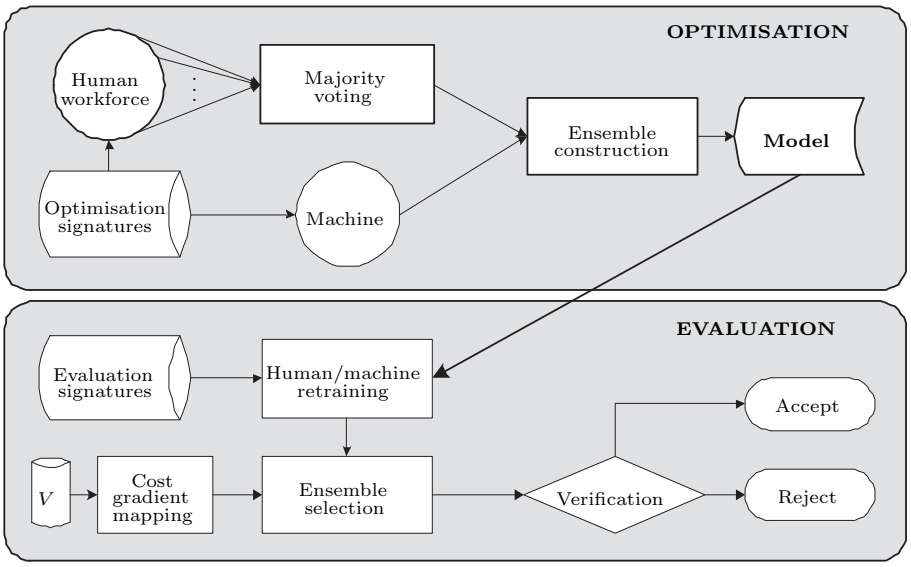


Figure 2. Schematic representation of the system proposed in this paper.

6. EXPERIMENTS

The data set considered in the subsequent experiments is segmented into an *optimisation set* (OS) and an *evaluation set* (ES). These two subsets contain signatures from *different* writers. The OS consists of guinea pig writers and the ES is representative of the general public. It is reasonable to assume that positive signatures are available for *each* writer in *both* of the aforementioned data sets and may be used to train writer-specific HMMs and serve as reference for the human workforce. *Labelled* positive and negative samples are *only* available for writers in the OS and are used to determine the proficiency of the machine and the human workforce, as well as to select optimal classifier ensembles. *Unlabelled* positive and negative samples, belonging to writers in the ES, are used to determine the generalisation potential of the system.

Data. In order to demonstrate the merit of the proposed protocol, we consider signatures from 51 writers that are randomly selected from a data set that was originally captured on-line.[12] The dynamic signatures are converted into static signature images using the pen position data and a morphological dilation operator.[13] In the subsequent experiments we only consider *skilled forgeries*. A skilled forgery is produced by someone who has access to one or more genuine samples of a writer's signature, and ample time to practise its reproduction. For each writer in the data set, there are 15 genuine 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.

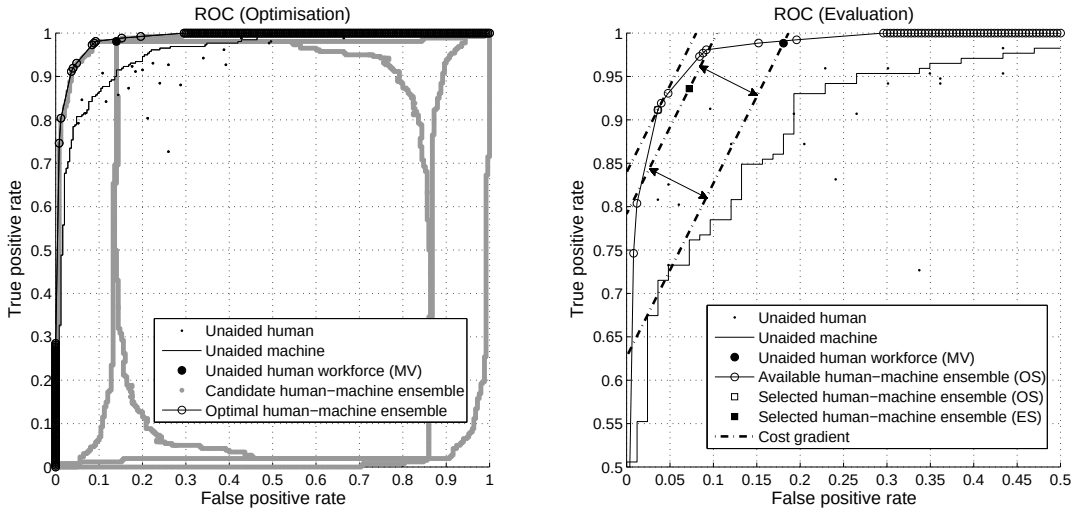
Experimental Protocol. For each writer, all of the 15 available genuine training signatures are utilised, while a test set, that consists of *only* 15 signatures, is constructed. Each test set (that is used for optimisation and evaluation) contains a randomly selected number (between 0 and 15) of skilled 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. Each classifier (human or machine) is presented with a total of $15 \times 51 = 765$ test signatures. The *total* number of genuine signatures and forgeries contained in the test sets turn out to be 432 and 333, respectively.

Twenty-three amateur human verifiers (consisting of faculty members and graduate students) are each presented with a training set (15 signatures) and corresponding test set (15 signatures) for all 51 writers in the data set. The training set and the test set for a specific writer are presented on two separate sheets of paper. Each human typically compares the test signatures, as a unit, with the corresponding training set and then decides which of the test signatures to classify as fraudulent. Each individual human verifier was instructed not to ponder over a decision, so as to simulate what a bank official is likely to do.

The same training and test signatures, that are considered for human verification, are also considered for machine verification. Based on the calculation of the discrete Radon transform, features are extracted from each signature image. These features are utilised to train an HMM for each writer. A questioned signature is matched with the appropriate HMM using Viterbi-alignment, so that a score is obtained. This score is then normalised using a strategy based on the z-norm.[13]

Since the utilised data contains signatures from only 51 writers, the experimental protocol is based on a combination of 3-fold cross validation and repetitive data randomisation, and proceeds as follows: (1) The data set is split into three equal subsets, each containing the signatures of 17 writers; (2) Each subset, in turn, is used as an evaluation set containing 17 writers, while the remaining two subsets constitute the optimisation set containing 34 writers; (3) The order of the writers is randomly shuffled, and the process is repeated 10 times. We therefore report the results from 30 trials.

Results. We first demonstrate the proposed collaboration strategy by considering a *single trial*, using the data discussed in the previous section. Figure 3 (a) depicts the estimated performances when considering the OS for a single trial, while Figure 3 (b) depicts the estimated performances when considering the ES for the *same* trial. Since the optimal human-machine ensembles generated during model optimisation are also available for *selection* during model evaluation, the same MAROC curve is shown for both scenarios. During evaluation, when considering a specific transaction value V that corresponds to a cost gradient of $M = 2$, *only* the ensemble on the MAROC curve that minimises the expected cost is selected. It is clear from Figure 3 (b) that, for the trial in question, the expected cost for the optimal human-machine ensemble is less than the expected cost for the unaided human workforce (using MV).



(a) (b)

Figure 3. Classifier performance in ROC space for a *single* trial. The optimal human-machine ensemble outperforms each individual unaided human, the unaided machine and the unaided human workforce (using MV).

Table 1 shows the average AUC (μ_{AUC}) and expected cost (μ_{EC}) for *all* 30 trials and for 50 equally distributed cost gradients ranging from 0.01 to 15. Note that μ_{AUC} is calculated by considering *every* threshold-specific machine and optimal human-machine ensemble available, while μ_{EC} is calculated by considering *only* the machine and optimal human-machine ensemble that minimises the expected cost for the specific cost gradient considered. The proposed optimal human-machine ensemble-based strategy outperforms the unaided human workforce (using MV) and the unaided machine when the AUC and EC is considered. The standard deviation (σ) and generalisation error (ϵ), for both quality performance measures, is also reported for each strategy.

In Figure 4 the average expected cost is plotted against the cost gradient for *all* 30 trials, and the results are compared for the scenarios where the unaided human workforce (MV), the unaided machine, or the optimal

	Optimisation set			Evaluation set		
	Unaided humans (MV)	Unaided machine	Optimal human-machine ensemble	Unaided humans (MV)	Unaided machine	Optimal human-machine ensemble
μ_{AUC}	0.9170	0.9501	0.9853	0.9175	0.9507	**0.9797**
σ_{AUC}	0.0102	0.0056	0.0034	0.0203	0.0108	**0.0081**
ϵ_{AUC}	-	-	-	+0.0005	+0.0006	**-0.0056**
μ_{EC}	0.0621	0.0325	0.0196	0.0619	0.0298	**0.0234**
σ_{EC}	0.0149	0.0127	0.0059	0.0205	0.0138	**0.0089**
ϵ_{EC}	-	-	-	-0.0003	-0.0027	**+0.0039**

Table 1. Performance comparison of the strategies considered in this paper for all 30 trials and for 50 equally distributed cost gradients ranging from 0.01 to 15. For both quality performance measures considered, the optimal human-machine ensemble outperforms the unaided humans (MV) and the unaided machine.

human-machine ensemble is utilised. The optimal human-machine ensemble outperforms the unaided human workforce (MV) for *all* operating conditions. It is also clear that the difference between the average expected cost for the optimal human-machine ensemble and the unaided human workforce (MV) is more pronounced in high risk scenarios, i.e. $M > 1$, and that the average expected cost for the optimal human-machine ensemble is significantly less than that for the workforce of unaided humans *and* the unaided machine in *common* risk scenarios, i.e. $M \approx 1$.

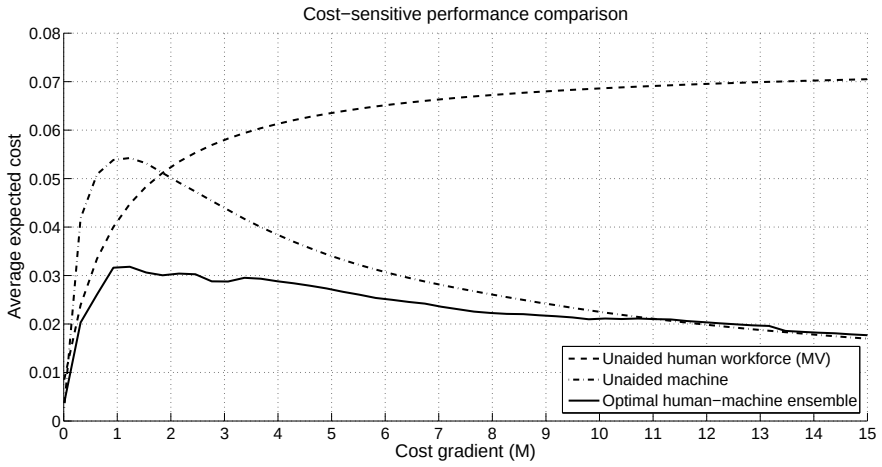


Figure 4. Performance comparison in a cost-sensitive environment, when only the *evaluation sets* are considered. The optimal human-machine ensemble outperforms the unaided human workforce (MV) for *all* operating conditions.

7. CONCLUSION AND FUTURE WORK

We showed that the proposed protocol for incorporating a machine into the authentication process can significantly improve the performance of a workforce of human employees for all operating conditions. The protocol also enables a financial institution to select an appropriate mapping between the monetary value, associated with a *specific* transaction, and a ROC-based cost gradient in a much more intuitive way than the previously proposed Neyman-Pearson-based approach,[6] so as to to minimise the expected cost. Since the human decisions are initially fused by majority voting, the relative simplicity of the protocol enables the optimal human-machine ensembles to be efficiently re-estimated should one or more employees not be available. The proposed protocol can be easily adapted to allow for the utilisation of more than one machine. The majority vote decision of the human workforce is first combined with the threshold-specific machine-generated decisions associated with the respective machines, so that *several* MAROC curves are obtained. These MAROC curves can then be combined

using the IBC strategy[9] detailed in section 3, so that a *final* MAROC curve is obtained. The same procedure, as detailed in this paper, can then be employed to authenticate an unlabelled signature.

A much less efficient, but potentially more accurate, collaboration strategy is based on the concept of first utilising the Boolean fusion functions detailed in section 3 to combine a specific human's decision with the decision of each threshold-specific machine-generated classifier, and then selecting the best human-machine hybrid. A pool of optimal human-machine hybrids, each one linked to a specific human, is therefore produced. For each predefined ROC-based cost gradient, a genetic algorithm can then be used to obtain the ensemble of human-machine hybrids that minimises the expected cost − fusion is achieved through majority voting. This strategy does not guarantee that all the humans are included in the decision process, but the number of humans utilised can be maximised by using *both* the cardinality of the ensemble *and* the expected cost as objective functions to guide the search. This concept is currently being investigated.

Finally, in Figure 4 we showed that the optimal human-machine ensemble outperforms the unaided human workforce for all cost gradients considered. For relatively large cost gradients ($M > 11$ in Figure 4), however, the unaided machine slightly outperforms the optimal human-machine ensemble. This is also the case for relatively small cost gradients, although this is not apparent from Figure 4. During model optimisation, however, the optimal human-machine ensemble outperforms *both* the unaided human workforce *and* the unaided machine for *all* operating conditions. It may be worth investigating whether this is a persistent phenomenon. If so, one may consider an additional data set for validation, in order to obtain a graph similar to Figure 4, and use said graph to find the optimal lower and upper interval bounds for M_1 and M_K, respectively.

REFERENCES

1. J. Coetzer, B. Herbst, and J. du Preez, "Off-line signature verification: a comparison between human and machine performance," in *Tenth International Workshop on Frontiers in Handwriting Recognition*, *Proc. IEEE*, pp. 481–485, 2006.
2. J. Zou and G. Nagy, "Human-computer interaction for complex pattern recognition problems," in *Data Complexity in Pattern Recognition*, M. Basu and T. Ho, eds., pp. 271–286, Springer London, 2006.
3. A. White, "The human-machine partnership in ucav operations," *Aeronautical Journal* **107**(1068), pp. 111–116, 2003.
4. D. Kragic, P. Marayong, M. Li, A. Okamura, and G. Hager, "Human-machine collaborative systems for microsurgical applications," *Int'l J. Robotics Research* **24**(9), pp. 731–741, 2005.
5. J. Zou and G. Nagy, "Visible models for interactive pattern recognition," *Pattern Recognition Letters* **28**(16), pp. 2335–2342, 2007.
6. J. Coetzer and R. Sabourin, "A human-centric off-line signature verification system," in *Ninth International Conference on Document Analysis and Recognition*, *Proc. IEEE* **1**, pp. 153–157, 2007.
7. S. Haker, W. Wells, S. Warfield, I. Talos, J. Bhagwat, D. Goldberg-Zimring, A. Mian, L. Ohno-Machado, and K. Zou, "Combining classifiers using their receiver operating characteristics and maximum likelihood estimation," in *Medical Image Computing and Computer-Assisted Intervention*, **3749**, pp. 506–514, 2005.
8. L. Oliveira, E. Justino, and R. Sabourin, "Off-line signature verification using writer-independent approach," in *Int'l Joint Conf. on Neural Networks*, pp. 2539 –2544, 2007.
9. W. Khreich, E. Granger, A. Miri, and R. Sabourin, "Iterative boolean combination of classifiers in the roc space: An application to anomaly detection with hmms," *Pattern Recognition* **43**(8), pp. 2732–2752, 2010.
10. C. Drummond and R. Holte, "Cost curves: An improved method for visualizing classifier performance," in *Machine Learning*, **65**(1), pp. 95–130, 2006.
11. T. Fawcett, "Roc graphs with instance-varying costs," *Pattern Recognition Letters* **27**(8), pp. 882–891, 2006.
12. J. Dolfing, E. Aarts, and J. van Oosterhout, "On-line signature verification with hidden markov models," in *Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*, **2**, pp. 1309–1312, 1998.
13. J. Coetzer, B. Herbst, and J. 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* **2004**(4), pp. 559–571, 2004.