Fusion of biometric systems using Boolean combination: an application to iris-based authentication

Eric Granger*, Wael Khreich and Robert Sabourin

Laboratoire d'imagerie, de vision et d'intelligence artificielle,
École de technologie supérieure
1100, rue Notre-Dame Ouest,
Montreal, QC H3C 1K3, Canada
E-mail: eric.granger@etsmtl.ca
E-mail: wael.khreich@livia.etsmtl.ca
E-mail: robert.sabourin@etsmtl.ca
*Corresponding author

Dmitry O. Gorodnichy

Video Surveillance and Biometrics Section,
Science and Engineering Directorate,
Canada Border Services Agency,
14 Colonnade Dr., Ottawa, ON K2E 6T7, Canada
E-mail: Dmitry.Gorodnichy@cbsa-asfc.gc.ca

Abstract: To improve accuracy and reliability, Boolean combination (BC) can efficiently integrate the responses of multiple biometric systems in the ROC space. However, BC techniques assume that recognition systems are conditionally-independent and that their ROC curves are convex. These assumptions are rarely valid in practice, where systems face complex environments, and are designed using limited enrollment data. In recent research, the authors have introduced an Iterative BC (IBC) technique that applies all Boolean functions iteratively, without prior assumptions. In this paper, IBC is considered for fusion of different commercial biometric systems at the decision level. Performance of IBC is assessed for biometric authentication applications in which the operational response of unimodal iris-base systems are combined. Experiments performed with four different commercial systems using anonymised data collected by the Canada Border Services Agency indicate that IBC fusion *with interpolation* can significantly outperform related BC techniques and individual systems.

Keywords: biometrics; classification; information fusion; ROC; receiver operating characteristics; BC; Boolean combination; iris modality.

Reference to this paper should be made as follows: Granger, E., Khreich, W., Sabourin, R. and Gorodnichy, D.O. (2012) 'Fusion of biometric systems using Boolean combination: an application to iris-based authentication', *Int. J. Biometrics*, Vol. 4, No. 3, pp.291–315.

Biographical notes: Eric Granger received a PhD in Electrical Engineering in 2001 from the École Polytechnique de Montréal, Canada. He is an Associate Professor in the Department of Automated Manufacturing Engineering, École de technologie supérieure, in Montreal. His research interests include adaptive pattern recognition, computational intelligence, online and incremental learning, ambiguity and novelty detection, and multi-classifier systems, with applications in biometrics and in computer and network security.

Wael Khreich received his PhD in Engineering in 2011 from École de technologie supérieure, Montreal, Canada. He is currently conducting postdoctoral research in natural language processing and machine translation at NLP technologies in Montreal. His research interests are in statistical machine learning, incremental learning, adaptive classification, and decision fusion in multi-classifier systems, with applications in computer and network security, biometrics authentication, and machine translation.

Robert Sabourin received his PhD in Electrical Engineering in 1990 from École Polytechnique de Montréal, Canada. In 1983, he joined the Department of Automated Manufacturing Engineering of the École de technologie supérieure, Université du Québec, in Montréal where he is currently Full Professor. His research interests include evolutionary computation and adaptive biometric systems.

Dmitry O. Gorodnichy received his PhD in Computing Science in 2000 from University of Alberta and PhD in Mathematics in 1997 from Ukrainian Academy of Sciences in Kiev. He is Senior Research Scientist with the Science and Engineering Directorate of the Canada Border Services Agency, where he leads the Biometrics and Video Surveillance Section. His current interests are in the theory and applications of video recognition and, in particular, face recognition in video, especially in the context of border security.

1 Introduction

The biometric recognition of individuals based on their behavioural or physiological traits, such as the face, finger print, iris, signature and voice, provides a powerful alternative to traditional authentication schemes (e.g., passwords and identification cards) presently applied in a multitude of security and surveillance systems (Jain et al., 2006; Kung et al., 2004). There are three types of applications in biometric recognition. With *verification applications*, an individual that is enrolled to the system identifies himself and provides a biometric sample. Then, the biometric system seeks to authenticate that the sample corresponds to the specific individual. In contrast, with *identification applications*, an individual provides a biometric sample, and the system seeks to determine if the sample corresponds to one of the individuals enrolled to the system. Finally, *surveillance or screening applications* differ slightly from identification in that the sampling process is performed covertly,

and they seek to determine if a given biometric sample corresponds to a restrained list of individuals of interest.

A unimodal biometric system captures raw samples of individuals using some sensor. Then, the system performs signal conditioning and segmentation to isolate samples of interest, from which invariant and discriminant features are extracted. During classification, the resulting features are assembled into patterns, and matched against the biometric model of one (verification) or of all (identification and surveillance) individuals enrolled to the system. Classification scores indicate the likelihood that the pattern corresponds to individuals, and is employed to provide application-specific decisions. For verification applications, the system accepts or rejects the authenticity, and for identification and surveillance applications, the system outputs a list of the most likely or of all possible matching identities, respectively.

Regardless of the application, biometric recognition may be modelled in terms of user-specific detection problems (Bengio and Mariéthoz, 2007), each one implemented using one or more pattern classifiers with thresholds applied to classification scores (Barreno et al., 2008; Bergamini et al., 2009; Oh and Suen, 2002). In practice, the accuracy of state-of-the-art neural and statistical classifiers applied to detection may decline because they face complex pattern recognition environments that change during operations, and because they are designed a priori using limited and imbalanced training data. The underlying data distribution corresponding to individuals enrolled to a biometric system may be complex due to inter- and intra-class variability and noise. In addition, the enrolment process typically involves some form of quality control or supervision to capture training samples, and classifiers used for detection are often trained using very few high-quality samples, contributing to a growing divergence between the biometric model of an individual and the underlying data distribution. Finally, recognition accuracy tends to decrease with the number of individuals enrolled to the biometric systems applied to identification and surveillance. These issues emerge, for instance, with current iris biometrics technology (Bowyer et al., 2008), which is intensively used to support the expedited traveller programmes of many governmental agencies – Canada Border Services Agency, USA Department of Homeland Security, UK Home Office, etc. (Gorodnichy et al., 2011).

Evidence from several studies suggest that the accuracy and reliability of a biometric system can be improved by integrating the evidence obtained from multiple different sources of information (Bergamini et al., 2009; Jain et al., 2005; Kittler et al., 1998, 2006; Snelick, 2005), including multiple samplings for a same biometric trait using different sensors, multiple different biometric traits, multiple instances and multiple samplings for a same biometric trait using a same sensor, or multiple feature extraction and classification algorithms processing a same biometric sample (Jain et al., 2006). Low-quality samples trigger a failure to enroll, and may prompt the user to provide more training samples. Various studies have also shown that poor quality biometric samples lead to a reduction in the accuracy during operations. Fusion controlled by quality measures has been shown to offer a significant gain in accuracy, but falls outside the scope of this paper.

Biometric sources of information are typically integrated at the feature, score and decision levels (Tulyakov et al., 2008). Since features extracted from sensor measurements contain richer information content about a biometric modality,

feature-level fusion should provide the higher level of accuracy, although commercial systems rarely reveal their feature patterns. The combined feature patterns may also be incompatible and increase system complexity (Zhang et al., 2008). Techniques for score-level fusion are commonly employed in biometrics when scores generated by the different commercial systems may be accessed (Jain et al., 2005; Snelick, 2005). The main limitations are the impact of score normalisation methods on the overall decision boundaries, and the availability of representative training samples. Despite reducing information, techniques for decision-level fusion may provide a simple and robust framework for combination, regardless of the specific type of biometric modality and system. Disadvantages include the limitations placed on decision boundaries due to the restricted operations that can be performed on binary decisions, and the need for independent data to design combination rules.

Boolean Combination (BC) has recently been investigated to combine the decisions of several crisp or soft detectors in the ROC space (Fawcett, 2006), where the performance of detectors is commonly characterised. These threshold-optimised decision-level combination techniques (Tao and Veldhuis, 2008) optimise operation points with respect to performance. In fact, ROC-based fusion is achieved by optimising the combination of decision thresholds, since these thresholds correspond to operation points. BC techniques have been shown to outperform well-known decision-level techniques (like majority voting) in the Neyman–Pearson sense (in terms of the detection rate for any false alarm rate). Using BC based on AND or OR functions has been shown to improve accuracy over the Maximum Realisable ROC (MRROC) technique alone, and over detection systems based on a single best classifier (Haker et al., 2005; Oliveira et al., 2008; Scott et al., 1998; Tao and Veldhuis, 2008). However, BC techniques found in literature assume that the classifiers are conditionally-independent, and that their corresponding ROC curves are smooth and proper. These idealistic assumptions are rarely valid in real-world biometric applications, where classifiers are designed using limited and imbalanced training data.

In previous research by the authors (Khreich et al., 2010), the IBC technique has been proposed for efficient BC of responses from multiple soft, crisp, or hybrid detectors. It iteratively combines the ROC curve produced by various detectors using all Boolean functions, and does not require any prior assumption regarding the independence of detectors and the convexity of ROC curves. Applying IBC to the responses of a multiple-HMM system has been shown to provide a significantly higher level of accuracy than related techniques in literature on intrusion detection data, specially when the HMMs are trained with limited and imbalanced data (Khreich et al., 2010). Although its iterative process does not necessarily provide an optimal set of combinations, its time complexity is linear with respect to the number of classifiers. Therefore, IBC represents a versatile information fusion technique for biometrics, where the ROC curves result from a wide range of biometric systems designed with different traits, sensors, feature sets, classifiers, training data and/or user-defined parameters.

In this paper, the IBC technique is considered for decision-level fusion of information produced by different commercial (black-box) biometric systems designed for iris-based authentication. It is assumed that the user does not have direct access to features or scores, but may control the discrimination (i.e., adjust decision thresholds) employed to produce decisions. Under this scenario, it is

assumed that a same biometric sample is processed independently by all biometric systems, each one using different data processing algorithms. During the enrolment process, an iris scanner produces a single sample per individual that is independently used by each unimodal biometric system to design biometric (iris) models. During operations, when an unknown individual presents himself to the iris scanner and provides a sample, each system produces a classification score, and thresholds are applied to produce decisions for each enrolled individual. Simulations are performed with anonymised data sets collected by the Canada Border Services Agency (CBSA) for large-scale evaluation of state-of-the-art iris biometrics systems. The performance obtained using four individual unimodal biometric systems is compared to that of decision-level techniques that combine the responses of these systems with BC and IBC. The impact on performance of increasing the population of individuals enrolled to biometric systems is also assessed.

The next section briefly reviews techniques for fusion of biometric sources according to various levels of a biometric system. Techniques for decision-level fusion of responses in the ROC space are also described. Then, the IBC technique is presented in Section 2.3. Section 3 presents the experimental methodology used for proof-of-concept computer simulations, including details on the iris data and evaluation protocols. Finally, simulation results are presented and discussed in Section 4.

2 Fusion of biometric information

2.1 Levels of fusion and techniques

To improve accuracy and reliability of biometric recognition, the responses from different systems can be combined at various levels according to pre- and post-classification techniques (Kuncheva, 2004; Tulyakov et al., 2008). Pre-classification fusion occurs at the sensor (raw biometric data) and feature levels, while post-classification fusion occurs at the score, rank and decision levels. Since sensor-level fusion is closely related to the specific sensor types and corresponding signal processing methods, it will not be discussed in this paper.

Biometric systems that perform information fusion at an early stage of processing are believed to be more accurate than those that perform fusion at a later stage. Features extracted from sensor measurements contain richer information content about a biometric modality than output scores or decisions from a classifier. Combining the features before prior to classification through, e.g., concatenation, should provide higher level of accuracy than other levels. However, fusion at the feature level is difficult to achieve in practice because proprietary Commercial Off-The-Shelf (COTS) systems do not typically divulge their feature patterns. In addition, the size of combined feature patterns may also increase system complexity, making it more difficult to design the classifier. Moreover, the feature sets of different systems may be incompatible, and infeasible to combine them on a common basis (Bouchaffra and Amira, 2008; Zhang et al., 2008).

Pre-classification fusions techniques are based on the generation of Ensembles of Classifiers (EoCs), where base classifiers are trained on different data subsets selected using, e.g., data-splitting, cross-validation, bagging (Breiman, 1996), and boosting

(Freund and Schapire, 1996) techniques. Classifiers may also trained on different feature subsets using, for instance, the random subspace method (Ho, 1998). Static ensemble selection attempts to select base classifiers from a pool based on various accuracy or diversity measures (Brown et al., 2005; Kuncheva and Whitaker, 2003), and then combining their responses. An alternative consists in combining classifier responses and then selecting EoCs according to accuracy or diversity over independent validation data (Banfield et al., 2003; Ruta and Gabrys, 2005). However, since combination is performed before EoC selection, its success depends on the chosen combination.

At the score level, information about the biometric modality is reduced from a feature pattern to a scalar classification score. Scores contain rich information about the biometric modality, and it is relatively easy to access and integrate scores generated by the different COTS systems. Consequently, techniques for post-classification combination at the classification score level are prevalent in biometric fusion (Jain et al., 2005; Snelick, 2005). Three approaches are commonly employed for fusion of scores obtained from different unimodals biometric systems (Bergamini et al., 2009) – transformation-based, density-based, and classification-based fusion. In *transformation-based fusion*, weighted individual scores are normalised to generate a single scalar score, which is then used to produce a final decision (Jain et al., 2005). These static fusion techniques are often implemented with a weighted sum or product of z-normalised scores, or using min-, max-, median-, mean-score techniques. The main disadvantage is that overall decision boundaries are influenced by score normalisation methods. In addition, normalisation of scores may be a difficult task for heterogeneous classifiers or across modalities.

Density-based fusion relies on the estimation of the joint densities of Genuine and Imposter scores, and is usually implemented using statistical likelihood ratio tests – the product of likelihood ratios and logistic regression are common density-based fusion techniques (Nandakumar et al., 2008). Finally, in *classification-based fusion*, each input pattern is labeled as either Genuine or impostor, and the individual scores produced by each biometric system are input as features to a global two-class classifier (Fierrez-Aguilar et al., 2005; Roli et al., 2002; Verlinde et al., 2000) (also known as stacked or meta-classifier). The limitation of density-based and classification-based fusion is the availability of sufficient representative training samples for accurate modelling of score distributions or to guarantee low generalisation error. For example, classification-based fusion are prone to overfitting if a representative and independent validation set is not available to estimate global parameters. Normalisation may also be an issue as these fusion techniques may affect the matching score densities.

When a biometric system is applied to identification applications, the system output may be viewed as a ranking of the enrolled identities – the set of possible matching identities sorted in decreasing order of confidence. Rank level fusion seeks to consolidate the output ranks by individual biometric systems in order to derive a consensus rank for each identity. Fusion at the rank level is mostly suitable for multi-class classification problems, where the correct class is expected to appear in the top of the ranked list. Logistic regression and Borda count (Ho et al., 1994; Van Erp and Schomaker, 2000) are among the more representative techniques at this level. Rank-level techniques simplify the combiner design since normalisation is not required.

Many COTS systems only provide access to the final binary decision. At the decision level, information content is further reduced from scores to binary decisions prior to being combined. Consequently, it is less investigated in literature than other levels of fusion. Despite only exploiting a compact set of operation points, such techniques may provide a simple and robust framework for combination that is independent of biometric modality and system. Technique for decision-level fusion include majority voting, Bayesian, and Dempster-Shafer techniques. A representative technique for decision-level fusion is the majority vote (Kittler et al., 1998). It counts the number of decisions from individual classifiers, and chooses the majority of the decision as its output. The weighted majority voting version assigns different weights according to the performance of individual classifiers, transforming output values from labels to continuous scores. Some potential issues that appear with decision level fusion include the possibility of ties, therefore the number of classifiers must be greater than the number of classes. Moreover, some Dempster-Shafer techniques requires a large number of training samples and some independent validation data to design combination rules.

In this paper, decision-level fusion is considered to combine different commercial biometric systems. It is assumed that the user does not have direct access to features or scores, but may adjust decision thresholds (or discrimination) employed to produce binary decisions. The main disadvantage of decision-level fusion is the limitations placed on decision boundaries, because the operations are restricted to thresholding and Boolean functions. The next subsections describe some powerful techniques for decision-level fusion in the ROC space (Fawcett, 2004) that may address this limitation.

2.2 Decision-level fusion in the ROC space

Assume that each detector is implemented using one or more 1- or 2-class classifiers. The performance of a user-specific detector may be characterised in the ROC space (Fawcett, 2006). A crisp detector outputs a binary decision and produces a single operational data point in the ROC space, while a soft detector assigns scores to the input samples, which can be converted to a crisp detector by thresholding the scores. A ROC curve is obtained by varying the threshold that discriminates between Genuine and Impostor classification scores. These scores are converted into a compact set of operational points, which indirectly convey information about the score distributions.

Given the responses of a detector for a set of test samples, the true positive rate (tpr) is the proportion of positives correctly classified over the total number of positive samples. The false positive rate (fpr) is the proportion of negatives incorrectly classified (as positives) over the total number of negative samples. A ROC curve is a parametric curve in which the tpr is plotted against the fpr. In practice, an empirical ROC curve is obtained by connecting the observed (tpr,fpr) pairs of a soft detector at each threshold. The Detection Error Trade-off (DET) space resembles the ROC space, but it plots the False Match Rate (FMR), where $\text{fpr} = \text{FMR}$, vs. the False Non-Match Rate (FNMR), where $\text{tpr} = 1 - \text{FNMR}$.

Each operation point on the ROC curve corresponds to a particular threshold applied to the scores. When the optimal operation points are obtained on a ROC, the thresholds of scores are also obtained. The operation points are tunable, and

can be optimised with respect to performance. Given two operation points, say a and b, in the ROC space, a is defined as *superior* to b if $\text{fpr}_a \leq \text{fpr}_b$ and $\text{tpr}_a \geq \text{tpr}_b$. If a ROC curve has $\text{tpr}_x > \text{fpr}_x$ for all its operation points x then, it is a *proper* ROC curve. In practice, an ROC plot is a step-like function which approaches a true curve as the number of samples approaches infinity. Therefore, it is not necessarily convex and proper. Concavities indicate poor local performance that may provide diverse information.

The area under the ROC curve (AUC) or the partial AUC has been largely suggested as a robust scalar summary of classifiers performance (Walter, 2005). The AUC assesses ranking in terms of class separation – the fraction of positive–negative pairs that are ranked correctly. For instance, with an $\text{AUC} = 1$, all positives are ranked higher than negatives indicating a perfect discrimination between classes. A random classifier has an $\text{AUC} = 0.5$, and both classes are ranked at random.

Given two or more crisp or soft detection systems, their decisions may be combined according to selected thresholds and Boolean functions. Decision-level techniques for fusion in the ROC space have been successfully applied in many real-world detection systems for biometrics (Tao and Veldhuis, 2008), bio-informatics (Haker et al., 2005) and intrusion detection (Khreich et al., 2009), etc., because they hold several advantages. For one, while common decision-level techniques combine responses directly on a fixed threshold, ROC-based combination of soft detectors involve sweeping the entire range of tpr and fpr, allowing for a flexible selection of the desired operating performance. A change of conditions, such prior class probabilities or costs of errors, lead to a shift in the optimal operating point on the composite convex hull, but the overall convex hull does not change. Fusion in the ROC space is not influenced by asymmetries in Genuine and impostor distributions, and normalisation of scores is not required because ROC curves are invariant to monotonic transformation of thresholds. The rest of this section briefly describes techniques for combination of detectors in the ROC space.

2.2.1 Maximum realisable ROC

The convex hull of an empirical ROC curve (ROCCH) is the smallest convex set containing its operation points or vertices, i.e., the piece-wise outer envelope connecting only the superior points of an ROC with line segments. It may be used to combining detectors based on a simple interpolation between the corresponding responses (Provost and Fawcett, 2001; Scott et al., 1998). In practice, this is achieved by randomly alternating detector responses proportionately between the two corresponding vertices of the line segment on the convex hull where the desired operational point (fpr) lies. This approach has been called the Maximum Realisable ROC (MRROC) (Scott et al., 1998) since it represents a system that is equal to, or better than, all the existing systems for all Neyman-Pearson criteria. However, the MRROC discards responses from inferior detectors which may provide diverse information for an improved performance. It only considers the responses of potentially 'optimal' detectors that lie on the facet of the ROCCH.

2.2.2 Boolean combination

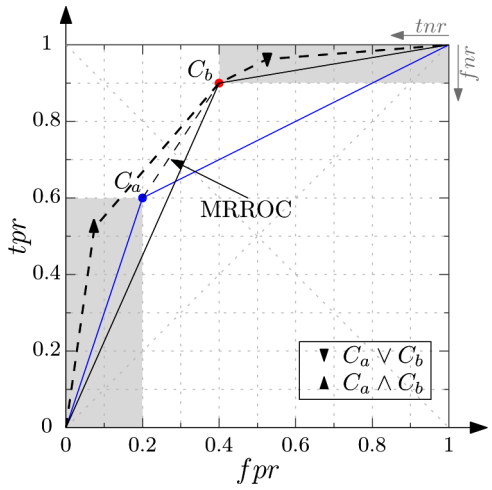
More recently, BC has been investigated to combine the decisions of several crisp detectors (1 operation point) or soft detectors (curve with n operation points) in

the ROC space. Also referred to by Tao et al. as *threshold-optimised decision-level combination techniques* (Tao and Veldhuis, 2008), BC techniques seek to optimise operation points (and decision thresholds) with respect to performance. Although considered to be decision-level techniques, such techniques can be seen as exploitings information from both score (or thresholds) and decision levels for fusion of detectors. They adaptively tune themselves between the two levels of fusion to improve overall performance (Tao and Veldhuis, 2008).

Given two soft classifiers, the corresponding ROC curves are combined for improved performance by optimising the combination of decision thresholds, since these thresholds correspond to operation points. BC involves applying a threshold to each score from the first curve and to each score from the second. The resulting decisions are then combined, and the thresholds that most improve the convex hull are selected. Note that BC techniques requires control over the decision thresholds applied to individual soft classifiers. If one of the two classifiers is soft, all its scores (or thresholds) are varied and combined with the operation point of the other crisp classifier, and then the best combinations are selected. However, with two crisp classifiers, BC techniques are purely decision-level as Boolean functions are applied to the two corresponding points in the ROC space, and the best combination is then selected.

The Boolean conjunction (AND) and disjunction (OR) fusion functions were first introduced in BC techniques for combining crisp detectors (Daugman, 2000) that are conditionally-independent. As illustrated in Figure 1, the AND function decreases the fpr at the expenses of the tpr, providing more conservative performance than that of the original detectors. Analogously, the OR function

Figure 1 Illustration showing the combination of two crisp detectors that are conditionally-independent, C_a and C_b, using the AND and OR functions. The performance achieved using these function for combination surpasses that of the MRROC. The shaded regions represent the expected performance of combination when there is an interaction between the detectors (see online version for colours)



Source: Fawcett (2004)

increases the tpr at the expenses of the fpr, providing more aggressive performance than that of the original detectors. Depending on detector interaction within the ROC space, these rules may produce a new composite convex hull that is superior to original detectors. Fusion with the OR function is especially useful in presence of outliers (Haker et al., 2005; Oxley et al., 2007; Shen, 2008; Tao and Veldhuis, 2008), where Genuine biometric samples deviate from the Genuine distribution.

The conditional independence assumption among the detectors simplifies the computation. In this cases, the combination rules depend only on the tpr and fpr (Black and Craig, 2002; Fawcett, 2004). In the more realistic conditionally-dependent case, the performance of the composite crisp detectors depends on the positive and negative correlations between detectors (Black and Craig, 2002). In order to avoid the restrictive conditional assumption among detectors, the combination functions were extended to include all Boolean functions (Barreno et al., 2008). By ranking these combinations according to their likelihood ratios, an optimal combination is possible but due to the doubly exponential explosion of combinations, a global search for the optimal rules is impractical. Applying all Boolean functions, via an exhaustive brute-force search to determine optimal combinations leads to an exponential explosion of combinations, which is prohibitive even for a small number of crisp detectors (Barreno et al., 2008).

Several authors have proposed using Boolean AND and OR functions to combine soft detectors (Haker et al., 2005; Oxley et al., 2007; Shen, 2008; Tao and Veldhuis, 2008). For a pair-wise combination, the fusion function is applied to each threshold on the first ROC curve with respect to each threshold on of the second curve. The optimum threshold, as well as the Boolean fusion function, is then found according to the Neyman-Person test (Neyman and Pearson, 1933). That is, for each value of the fpr, the point which has the maximum tpr value is selected, along with the corresponding thresholds and Boolean function to be used during operations. Using BC with either AND or OR functions has been shown to improve the accuracy over the MRROC technique, over some conventional score- and decision-level techniques, and over individual systems (Haker et al., 2005; Oliveira et al., 2008; Scott et al., 1998; Tao and Veldhuis, 2008) in the Neyman-Person sense.

Most research has addressed the problem of BC under the assumption that systems are conditionally-independent, and that their corresponding ROC curves are smooth and proper. In this ideal case, when both conditional independence and convexity assumptions are fulfilled, the AND and OR combinations have been proven to be optimal, providing a higher level of performance than the original ROC curves (Barreno et al., 2008; Haker et al., 2005; Thomopoulos et al., 1989). In real-world biometric applications, where systems are designed using limited and imbalanced data, ROC curves may be improper and large concavities will appear. When either one of the assumptions is violated, AND and OR functions will not improve performance for inferior points that correspond to concavities.

2.3 Iterative Boolean Combination

In recent research, the authors proposed the Iterative Boolean Combination ($\mathrm{BC_{ALL}}$) technique (Khreich et al., 2010) for efficient combination of responses from multiple soft and crisp detectors in the ROC space. IBC represents a versatile

information fusion technique for biometrics, where the binary decisions may come from a wide range of systems designed with different traits, sensors, feature sets, classifiers, training data and/or user-defined parameters.

This decision-level combination technique exploits *all* Boolean functions iteratively, and requires no prior assumptions regarding the independence of detectors and the convexity of ROC curves. At each iteration, $\mathrm{IBC_{ALL}}$ selects the combinations that improve the convex hull and recombines them with the original ROC curves until the MRROC ceases to improve. This process implicitly accounts for the effects of correlation among detectors. In Khreich et al. (2010), IBC was successfully applied to the fusion of multiple-HMM systems for host-based intrusion detection. It has been shown to provide a significantly higher level of accuracy than related techniques in literature on real-world data, specially when the HMMs are trained with limited and imbalanced data. Although this iterative process does not provide an optimal set of combinations, its time complexity is linear with respect to the number of classifiers, and does not suffer from the exponential explosion (Barreno et al., 2008).

The main steps of $\mathrm{BC_{ALL}}$ are presented in Algorithm 1. It combines the responses of two detectors using all Boolean functions, prior to applying the MRROC to select the thresholds and Boolean functions that most improve the ROCCH. The $\mathrm{BC_{ALL}}$ technique inputs a pair of ROC curves defined by their decision thresholds, T_a and T_b, and the labels for the validation set. Using each

Algorithm 1: $BC_{ALL}(T_a, T_b, labels)$: Boolean combination of two ROC curves

Input: Thresholds of two ROC curves, T_a and T_b (or their responses Ra and Rb), and true labels *labels* (of validation set)

Output: ROCCH and fused responses (R) of combined curves; each point results from two thresholds combined with a Boolean function (bf)

let $m \leftarrow$ number of distinct thresholds in T_a
let $n \leftarrow$ number of distinct thresholds in T_b
Allocate F an array of size $[2, m \times n]$
$BooleanFunctions \leftarrow$
$\{a \wedge b, \neg a \wedge b, a \wedge \neg b, \neg(a \wedge b), a \vee b, \neg a \vee b, a \vee \neg b, \neg(a \vee b), a \oplus b, a \equiv b\}$
Compute $ROCCH_{old}$ of the original curves
foreach $bf \in BooleanFunctions$ **do**
    for $i = 1, \ldots, m$ **do**
        $R_a \leftarrow (T_a \geq T_{a_i})$
        for $j = 1, \ldots, n$ **do**
            $R_b \leftarrow (T_b \geq T_{b_j})$
            $R_c \leftarrow bf(R_a, R_b)$
            Compute (tpr, fpr) using R_c and *labels*
            Push (tpr, fpr) onto F
    Compute $ROCCH_{new}$ of F
    Store thresholds and Boolean functions of vertices that exceeded the $ROCCH_{old}$: $s^*_{global} \leftarrow (T_{a_x}, T_{b_y}, bf)$
    Store the responses of these emerging vertices into R
    $ROCCH_{new} \leftarrow ROCCH_{old}$
Return $ROCCH_{new}, R, s^*_{global}$

of the 10 Boolean functions, $\mathrm{BC_{ALL}}$ combines the responses of each threshold from the first curve (R_{a_i}) with the responses of each threshold from the second (R_{b_i}). Responses of the fused thresholds are then mapped to points (fpr,tpr) in the ROC space. The thresholds of operation points that exceeded the original ROCCH of original curves are then stored along with their corresponding Boolean functions. The ROCCH is then updated to include the new emerging points. When the algorithm stops, the final ROCCH is the new MRROC in the Newman-Pearson sense. The outputs are the vertices of the final ROCCH, where each point is the results of two thresholds from the ROC curves fused with the corresponding Boolean function. These thresholds and Boolean functions form the elements of s^*_{global}, and are stored and applied during operations.

The $\mathrm{BC_{ALL}}$ technique requires no assumptions regarding the independence of detectors. It directly fuses the responses of each decision threshold, accounting for both independent and dependent cases. In fact, by applying all Boolean functions to combine the responses for each threshold, it implicitly accounts for the effects of correlation. In the worst-case scenario, when the responses of detectors provide no diversity of information, or when the shape of the ROC curve on design data differs significantly from that of test data, the $\mathrm{BC_{ALL}}$ is lower bounded by the MRROC of the original curves.

Exploiting all Boolean functions accommodates for the concavities in ROC curves. Indeed, AND and OR rules will not provide improvements for the inferior points that correspond to concavities (and make for improper ROC curves), or points that are close to the diagonal line in the ROC space. Other Boolean functions, for instance those that exploit negations of responses, may however emerge. $\mathrm{BC_{ALL}}$ can therefore be applied even when training and validation data are limited and heavily imbalanced.

Cumulative combination of multiple ROC curves involves selecting any pair of the detectors or ROC curves then combine the resulting responses with the third, then with the fourth and so on, until the last ROC curve (Tao and Veldhuis, 2008). As described in Algorithm 2, the thresholds (T_1 and T_2) of first two ROC curves are initially combined with $\mathrm{BC_{ALL}}$. Then, their combined responses (R_1) are directly input into line 8 of Algorithm 1 and combined with the thresholds of the third ROC curve (T_3). The time and memory complexity associated with the cumulative strategy can be considerably lower than that of for a pair-wise strategy due to the lower number of permutations. In additions, the pair-wise strategy requires combining all thresholds for each two curves, while combining the resulting responses with a new curve is less demanding since the number of selected responses is typically much lower than the number of thresholds.

Algorithm 2: $BCM_{ALL}([T_1,\ldots,T_K],labels)$: Cumulative BC of multiple ROC curves based on BC_{ALL}

Input: Thresholds of K ROC curves $[T_1,\ldots,T_K]$ (or their responses) and true *labels*
Output: ROCCH of combined curves; each point is the result of combination from several curves

1 $[ROCCH_{1:2},R_{1:2}] = BC_{ALL}(T_1,T_2,labels)$
2 **for** $k = 3,\ldots,K$ **do**
3 $\quad [ROCCH_{1:k},R_{1:k}] = BC_{ALL}(R_{1:k-1},T_k,labels)$
4 **Return** $ROCCH_{1:K},R_{1:K}$ and selected thresholds and corresponding Boolean functions

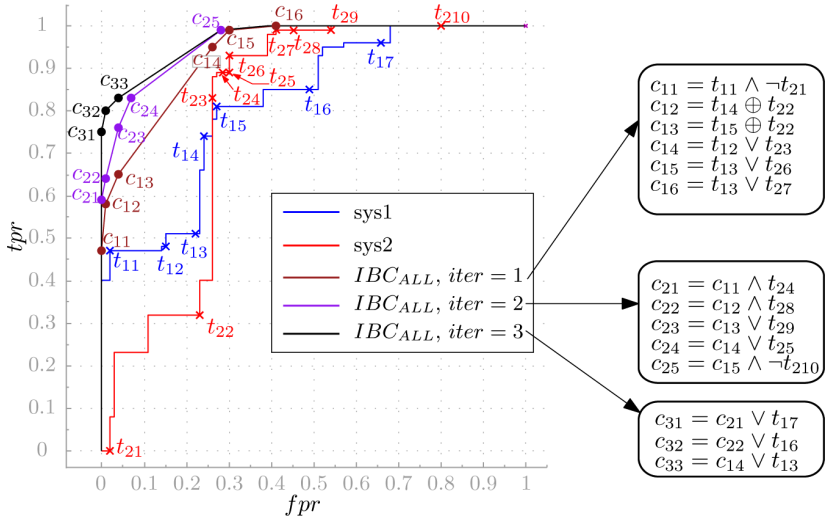
Algorithm 3: $IBC_{ALL}([T_1,\ldots,T_K],labels)$: Iterative Boolean Combination (IBC)

Input: Thresholds of K ROC curves $[T_1,\ldots,T_K]$ (or their responses) and true *labels*
Output: ROCCH of combined curves; each point is the result of a composite combination

1 $[ROCCH_{OLD},R_{OLD}] = BCM([T_1,T_2,\ldots,T_K],labels)$
2 **while** $(AUC(ROCCH_{NEW}) \geq AUC(ROCCH_{OLD}) + \varepsilon)$ *or* $(numberIterations \leq maxIter)$ **do**
3 $\quad [ROCCH_{NEW},R_{NEW}] = BC(R_{OLD},[T_1,T_2,\ldots,T_K],labels)$
4 **return** $ROCCH_{NEW},R_{NEW}$ and selected thresholds and corresponding Boolean functions

Further improvements in performance may be achieved by re-combining the output responses of combinations resulting from the $\mathrm{BC_{ALL}}$ (or $\mathrm{BCM_{ALL}}$) with those of the original ROC curves over several iterations. The Iterative Boolean Combination ($\mathrm{IBC_{ALL}}$) technique is presented in Algorithm 3, and maximises the AUC of K ROC curves by re-combining the previously selected thresholds and fusion functions with those of the original ROC curves over several iterations, until the overall ROCCH no longer improves. During the first iteration, the ROC curves of two or more detectors are combined using the $\mathrm{BC_{ALL}}$ or $\mathrm{BCM_{ALL}}$. This defines a potential direction for further improvements within the combination space. Then, the $\mathrm{IBC_{ALL}}$ proceeds in this direction by re-considering information from the original curves over several iterations. The iterative procedure accounts for potential combinations that may have been disregarded during the first iteration, but are useful when provided with limited and imbalanced training data. The iterative procedure stops when there are no further improvements to the AUC or when a maximum number of iterations are performed. The example in Figure 2 shows the impact of and iterative combination over all Boolean functions with IBC.

Figure 2 Illustration showing the thresholds and Boolean functions selected to combine the responses of two soft detectors, sys1 and sys2, after 3 iterations of the IBC_{ALL} technique. Both original ROC curves feature concavities that arise with complex pattern recognition problems and with limited design data. $\mathrm{IBC_{ALL}}$ improves the overall AUC performance based on the OR, AND, XOR and negation functions (see online version for colours)



Although sub-optimal, the $\mathrm{IBC}_{\mathrm{ALL}}$ algorithm overcomes the exponential growth in computational complexity associated with a brute-force strategy as in Barreno et al. (2008). Given a pair of detectors, C_a and C_b, having respectively n_a and n_b distinct thresholds on their ROC curves. During the design of the $\mathrm{IBC}_{\mathrm{ALL}}$ system, the worst-case time complexity (required for computing all 10 Boolean functions to combine thresholds) and memory complexity (required to store the temporary results (tpr,fpr) of each Boolean function) is $\mathcal{O}(n_a n_b)$. When the $\mathrm{BCM}_{\mathrm{ALL}}$ is applied to combine the response of several ROC curves of K detectors, the worst-case time can be roughly stated as K times that of the $\mathrm{BC}_{\mathrm{ALL}}$ algorithm. However, after combining the first two ROC curves, the number of emerging responses on the ROCCH, is typically very small with respect to the number of thresholds on each ROC curve. $\mathrm{BC}_{\mathrm{ALL}}$ is efficient in scenarios with limited and imbalanced data because the number of distinct thresholds is typically small.

Figure 3 An illustration of the ROCCHs obtained with IBC for 2 synthetic cases: (1) easy and (2) complex detection problems. To improve the overall performance, IBC combines decisions of two detectors, sys1 and sys2, by exploring all Boolean functions over several iterations (see online version for colours)

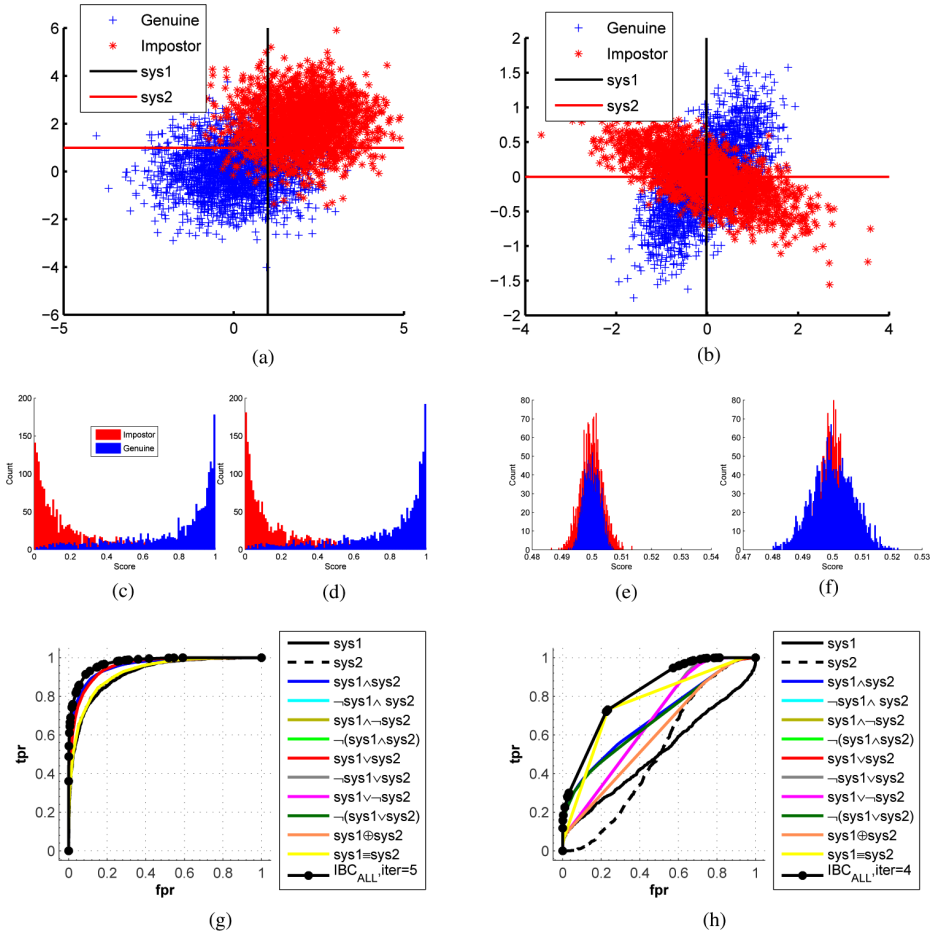


Figure 3 presents an application of IBC under two synthetic scenarios. In case 1 (Figure 3(a)), Genuine and Imposter data samples are represented with two moderately overlapping Gaussian distributions centred at (0,0) and (2,2), respectively in a 2D feature space. Both distributions are spherical, with a covariance matrix of [1 0; 0 1]. In case 2 (Figure 3(b)), Genuine data samples are represented with two Gaussian distributions (centred at (0.5, 0.3) with covariance matrix of [0.2 0.1; 0.1 0.2] and at (–0.5 –0.3) with covariance matrix of [0.2 0.1; 0.1 0.2]), while Imposter samples are represented with a Gaussian distribution centred at (0, 0) with covariance matrix [0.9 0.7; –0.1 0.1]. In both cases, data sets are formed by generating 2000 samples for Genuine classes and 2000 samples for Impostor classes.

In each case, IBC is used to combine responses of two Linear Discriminant Classifiers (LDCs), sys1 and sys2, where each classifier is trained independently. Respective decision bounds are shown to be vertical (sys1) and horizontal (sys2). The improvements achieved with IBC depend to some extent on the diversity of response by detectors, provided by using different data sets, feature sets, classification methods, etc., and is problem-dependent. Figure 3(c)–(f) show the Probability Distribution Functions (PDFs) of Genuine and Imposter classification scores produced by sys1 and sys2 on data samples. Scores correspond to normalised Euclidean distance values measured between Genuine and Imposter samples and LDC boundaries. For case 1, sys1 and sys2 provide well separated score distributions (represents an easy biometric detection problem), and IBC(sys1,sys2) mostly exploits AND and OR functions over 5 iterations to improve the ROCCH (see Figure 3(g)). In contrast, sys1 and sys2 provide very overlapping score distributions (represents a complex biometric detection problem) in case 2, and IBC(sys1,sys2) is shown to exploit many different functions over four iterations (see Figure 3(h)).

3 Experimental methodology

The main objective of experiments is to observe the performance and properties of IBC on real world biometric data. The performance is assessed for biometric authentication systems that perform decision-level fusion of responses from commercial iris-base biometric systems. The performance obtained using the unimodal biometric systems is compared to that of techniques that combine the responses of these systems with MRROC fusion (Scott et al., 1998) and with BC and IBC techniques. The impact on performance of increasing the population of individuals enrolled to biometric systems is also assessed in the ROC and DET spaces.

The experiments shown in this paper are conducted using binary decisions output from multiple different biometric systems for closed-set iris-based identification, as found in access control applications. Assume that a biometric sample is processed independently by four COTS systems, and that each one employs different pre-processing, feature extraction and classification algorithms. That is, during a prior enrolment process, an iris scanner produces a single sample per individual, and each unimodal biometric system uses that sample to design a user-specific iris model. So the individual presents himself or herself to the scanner, and produces a sample, but no other information. Then, during operations, when an unknown individual presents himself to the system and provides a sample, each system

produces classification scores, and thresholds are applied to produce a decision for each enrolled individual. These decisions correspond to iris scan samples from N individuals enrolled to each system. It is assumed that scores are not directly accessible for decision-level fusion, but decision thresholds may be adjusted to produce decisions.

Proof-of-concept simulations with fusion techniques are performed with two anonymised data sets collected by the Canada Border Services Agency (CBSA) for large-scale evaluation of state-of-the-art iris biometrics systems. The datasets, named G-100 and G-500, are formed with enrolled and passage images, corresponding to the same individuals. G-100 is the smaller subsets of G-500. In particular, a G-N dataset has N 'enrolled' images and 6N 'passage' images corresponding to the N enrolled travellers, where each enrolled passenger has exactly six passage images. Only right eye images are used.

Images in the datasets are only formed with images of the 'matched' or correctly recognised individuals. Each passage image has already been matched by the operational system to its corresponding enrolled image. The letter 'G' in the naming of the datasets refers to 'Genuine' to indicate the all images in the dataset come from Genuine transactions. The images are captured by a commercial iris acquisition system that applies some image quality check. However, the enrolled images are normally of better quality than passage images, since they are captured in a controlled environment at the time of enrolment, and under the guidance from an enrolling officer, while the passage data are captured in the airport with no guidance. The captured images are securely store using the system's proprietary format.

The 'Import' function is used to extract images from their original proprietary format into JPEG format, which results in image quality degradation. To mitigate the effect of such conversion on the evaluation results, all captured (enrolled and passage) anonymised iris images available in the operational database are imported to the JPEG format using the system's 'Import' function. Then, the compressed version of each image is compared to its original using the image quality function provided by the system, which can read both compressed and original images. If the image quality of both (compressed and original) versions of an image is the same, then the image is marked as 'not degraded', and may be used in experiments.

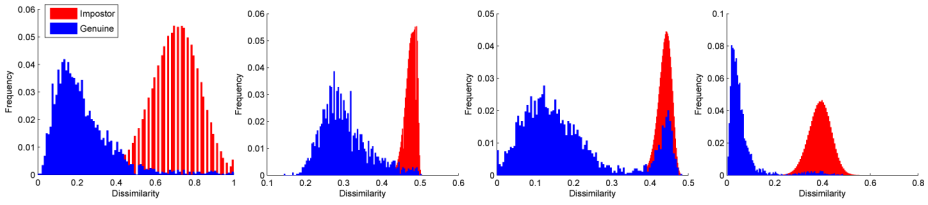
According to the CBSA testing protocol, for which the G-N datasets have been created, all $6N$ passage images are matched to all N enrolled images, resulting in $6N$ Genuine comparisons, and $6N(N-1)$ impostor comparisons. Therefore, both data sets contain six samples or passages images for each one of 100 or 500 enrolled individuals, respectively. The actual number of comparisons performed is often less than that due to a proportion of images being rejected by the system, due to the system's Failure of Acquire (FTA). Depending on the physical characteristics of the sensors, different systems failed to acquire different images. For an unbiased evaluation and combination of the systems, the FTA images have been filtered out. The result is the reduction of Genuine comparisons to 414 and Imposter comparisons $40,536$ for the G-100 data set, and the reduction of Genuine comparisons to 2000 and Imposter comparisons $975,077$ for the G-500 data set.

Genuine images are considered as non-target and impostor images as target. Passage images from each individual are matched against each enrolled image of those either 100 or 500 individuals. The number of Genuine and Impostor samples

Table 1 Number of genuine and impostor transactions and Failure to Acquire (FTA) iris images after presentation of G-100 and G-500 data to each system. For each system under test, the FTAs indicate the number of comparisons that are not performed due to FTA of either enrolled (FTA.E) or passage (FTA.P) images

Dataset	*System*	*1*	*2*	*3*	*4*	*Total*
G-100	Genuine	589	419	600	595	600
	Impostor	58,316	40,981	59,400	58,905	59,400
	FTA.E	11	181	0	5	
	FTA.P	1084	18,419	0	495	
	FTA	1095	18,600	0	500	
G-500	Genuine	2942	2049	2997	2963	3000
	Impostor	1,468,194	996,585	1,495,503	1,478,537	1,497,000
	FTA.E	58	951	3	37	
	FTA.P	28,806	500,415	1497	18,463	
	FTA	28,864	501,366	1500	18,500	

Figure 4 Normalised Probability Distribution Functions of genuine and imposter classification scores produced by each iris-based authentication system on the G-500 dataset. Note that in this case, the scores correspond to dissimilarity values (e.g., the Hamming distance) measured between passage and enrolment images. To present on common graphs, imposter and genuine distributions of each system were independently normalised according to respective transaction counts in Table 1 (see online version for colours)



(a) (b) (c) (d)

as well as the samples that triggered a Failure to Acquire (FTA) during enrolment (FTA.E) and passage (FTA.P) are shown in Table 1 for each system and each data set. The overall FTA is the sum of failures to acquire for enrolment and passage images, i.e., FTA = FTA.E + FTA.P. The histograms shown in Figure 4 illustrate the frequency distribution of Genuine and Imposter scores for the four individual COTS systems on G-500 data.

4 Simulation results and discussion

At first, the G-100 data set was used to compute the Boolean fusion functions for pairwise combinations of independent systems, and then for cumulative combination of all four systems according to the BC (with AND and OR) and IBC techniques (see Algorithm 3). Table 2 presents the partial AUC performance obtained by using the four individual systems, and with different combinations of these systems

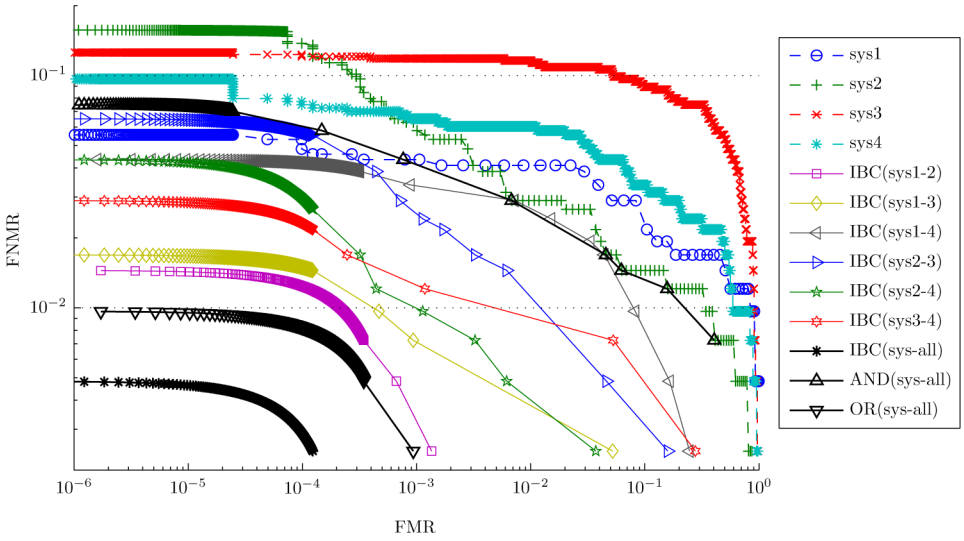
Table 2 Partial AUC performance, $AUC_{0.01}$, where the AUC is limited to a fpr between 0% and 1%, of individual systems (sys1 to sys4) and different combinations (using IBC, AND and OR). The order of pair-wise combinations with IBC(sys-all) is sys1-2-3-4

Systems		$AUC_{0.01}$	
		G-100	*G-500*
Individual	sys1	0.9585	0.9744
	sys2	0.9576	0.9488
	sys3	0.8825	0.9095
	sys4	0.9382	0.9712
Combination	IBC(sys1-2)	0.9659	0.9915
	IBC(sys1-3)	0.9600	0.9957
	IBC(sys1-4)	0.9595	0.9837
	IBC(sys2-3)	0.9597	0.9826
	IBC(sys2-4)	0.9610	0.9922
	IBC(sys3-4)	0.9407	0.9948
	IBC(sys-all)	0.9999	0.9983
	AND(sys-all)	0.9649	0.9661
	OR(sys-all)	0.9994	0.9969

on G-100 data. Figure 5 shows the DET curves for these same cases. These curves allow to observe the FNMR as a function of FMR in the range of $[10^{-6}, 1]$.

AUC is a global objective function that is generally used in the ROC space. The partial AUC values shown in Table 2 are a strong indicator of accuracy achieved with G-100 data during the BC design process, when BC functions are

Figure 5 The DET curves obtained with IBC Boolean Combination rules on the G-100 data (see online version for colours)



computed. The lower partial AUC results obtained with sys3 reflect the significant overlap between Imposter and Genuine samples in the score distribution (see Figure 4(c)). In the DET space, every pair-wise combination of individual systems improves the overall performance over the range of FMR values. Combining all four systems with IBC(sys-all) provides a considerably higher level of accuracy than any individual system or any pair-wise combination of systems. Results also show that IBC(sys-all) outperforms BC with OR(sys-all) and AND(sys-all).

To avoid excessive computational costs, the results shown in this paper reflect a uniform sub-sampling of only 2000 thresholds from each system prior to applying BC and IBC. Otherwise, with G-500 for instance (at about 1,000,000 samples per system), two curves would be combined with a resolution of 10^6 threshold points each, resulting in arrays F that require memory to store more than 10^{12} floating point values. Unfortunately, the number of samples and the increment of selected thresholds (resolution of the ROCCH) affect the shape of DET curves more than ROC curves.

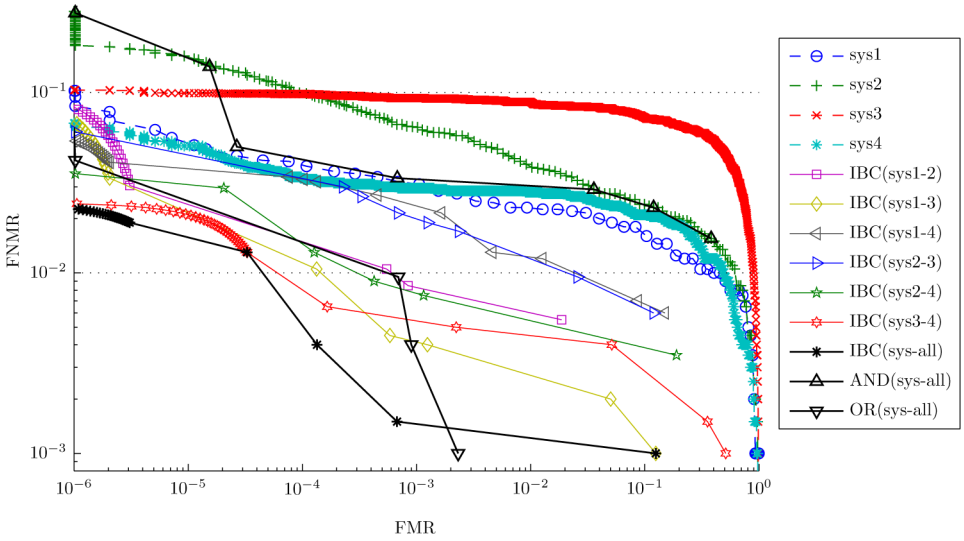
IBC design involves selecting thresholds and Boolean functions such that the AUC is optimised in the ROC space (see Algorithm 3). As the number of pair-wise combinations grows, the number of vertices (operation points) on the ROCCH and DET curves tends to decrease. In an extreme case, an ideal ROCCH (single point at (tpr,fpr) = (1,0) in ROC space) corresponds to a single point in the DET space. An IBC technique adapted for the DET space should however constrain its optimisation solutions to areas in the ROC space that correspond to lower FMR values. In fact, this lower part of the DET curve is the most important in biometric applications. Alternately, one can design fusion functions with more thresholds that are relevant for this area of the DET space.

The low cost approach applied in this paper consists in interpolating between ROCCH vertices with low fpr values. The DET curves presented in Figures 5 and 6 result from the use of interpolation to sample between two vertices of the empirical ROCCH: the vertice with the highest tpr at fpr = 0, and the next vertice with fpr > 0. This allows to generate an arbitrarily high number of realisable systems between these two vertices of a ROCCH (using the MRROC approach from Section 2.2.1), which translate to realisable systems in the DET space with an arbitrarily low FMR. This approach is possible even when the first vertice corresponds to the virtual vertice at (tpr,fpr) = (0,0). As an example, combining all four systems with IBC and interpolation – IBC(sys-all) – produces a Boolean fusion function that achieves an FNMR below 0.005 for an FMR of 10^{-6}.

Table 2 also presents the partial AUC performance obtained on G-500 data. Figure 5 shows the DET curves for these same cases in a range of FMR $\in [10^{-6}, 1]$. These curves are achieved by applying the BC and IBC combination rules obtained with G-100 (the design data). That is, BC functions produced on G-100 data are stored and then applied to the larger G-500 data set. Recall that for an unbiased combination and evaluation of systems, FTA images have been filtered out. The performance of each iris-based system is again compared to that of each pair-wise IBC combination of systems, and to that of a cumulative AND, OR and IBC combination of all systems.

When IBC is applied to combine multiple systems, the overall accuracy is seen to be fairly robust to variations of the number N of enrolled users, and thus to limited amounts of design data. As shown in Figure 6, among the pair-wise combinations,

Figure 6 The DET curves obtained by evaluating the IBC Boolean Combination rules (previously obtained using the G-100 data) on the G-500 data set (see online version for colours)



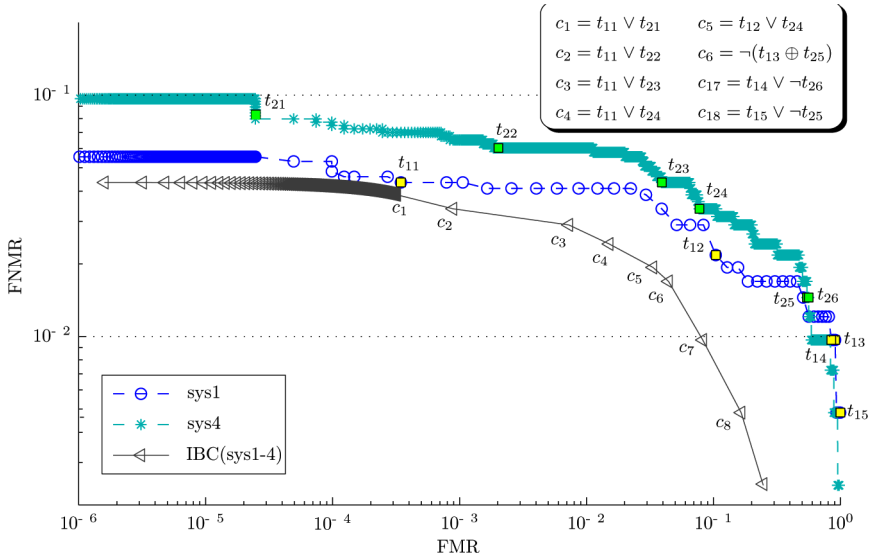
the responses from IBC(sys3-4) achieve the highest level of AUC accuracy in lower parts of the DET space. The cumulative combination of all four systems provides the highest level of AUC accuracy. Although it has been designed with limited and imbalanced design data, IBC is shown to exploiting the complimentary information provided by different systems to improve the overall performance.

In practice, combining all four systems may not be desirable for several practical, financial or other reasons. An alternative may consist of selecting the best pair-wise combination of individual systems. For instance, sys3-4 provides a sightly lower level of performance than with sys-all, while its performance is still considerably higher than with individual systems alone. Note that most of the combined systems rely mostly on a mixture of AND and OR functions (except sys1-4) because the corresponding ROC curves are smooth, proper and almost perfectly convex. The underlying score distributions (see Figure 4) are almost bi-normal. Figure 7 displays an illustrative example of the DET curve produced with IBC(sys1-4) on G-100 data. This curve is annotated with the thresholds and Boolean functions selected with IBC.

With BC and IBC, the order of pair-wise combinations has a significant impact on performance. Somewhat related is the availability of representative design data that corresponds to unique thresholds or operation points. Sys3 and sys4 are characterised by smooth and convex ROC curves. They provide good design data that contains a high density of unique operational points, while sys1 and sys2 produce many redundant points. By combining sys3 and sys4 first, BC and IBC limit the number of combinations related to redundant operation points.

IBC fusion appears to have an impact over all portions of the DET curve, although the lower portion has more significance for the biometric market. It should be emphasised that when plotting tradeoff DET or ROC curves, it is important to observe areas of prime interest – the area close to FMR = 0 for 'white-list' access-

Figure 7 Example of DET curves showing the thresholds and Boolean functions selected with IBC(sys1-4) (see online version for colours)



control applications, and the area close to FNMR = 0 for 'black-list' automated criminal screening applications. In the operational range that is suitable for white-list expedited border control in trusted traveller programmes, the FMR should be very low (e.g., FMR $\leq 10^{-5}$). From the DET curves, BC and IBC with interpolation allow to design accurate systems for very low FMR values. This remains to be verified for black-list applications, where the FNMR should be very low (e.g., FMR$\leq 10^{-4}$).

5 Conclusions

Decision-level fusion is considered in this paper to design accurate and reliable systems for biometric recognition. Among decision-level techniques, BC efficiently integrates the responses of multiple systems by optimising the combination of decision thresholds corresponding to operation points. In the absence of prior knowledge on a detection problem, the Iterative BC (IBC) technique is an efficient approach to implement a full iterative search over all Boolean functions. This general fusion technique improves performance even when detectors are conditionally-dependent and when their ROC curves have concavities. IBC produces a composite ROC convex hull over the entire ROC space, and each vertice on the facets of the convex hull activates different thresholds from different classifiers combined with Boolean functions. The systems inherit all desirable properties of BC in the ROC space. It covers the whole performance range of fpr and tpr, which allows for a flexible selection of the desired operating performance. As conditions change, such as prior probabilities and costs of errors, the composite ROC convex hull of combinations does not change; only the portion of interest, and the optimal operating point shifts to other vertices on the convex hull.

In this paper, the performance of IBC is examined for fusion of decisions in biometric authentication applications, in which the operational responses of several state-of-the-art COTS systems are combined. Experiments are conducted on real-world operational data collected by the CBSA, and comprised of responses generated by four iris-based systems with a populations of $N = 100$ and 500 enrolled individuals. Simulation results indicate that fusion with IBC can significantly improve the accuracy over any individual system alone, and over related BC techniques, especially when systems are designed using limited and imbalanced data collected during enrolment, and even when using highly accurate COTS systems. IBC performance is robust to variations of N and to limitations on the amount of data. Finally, IBC can effectively combine different biometric systems without accounting for the specific type of biometric scanner, data pre-processing, feature extraction and classification methods, which could negatively affect performance at the feature and score levels of fusion.

The topic of future research includes adapting IBC such that its thresholds and fusion functions are selected in the DET space to minimise low FMR values for white-list applications, or low FNMR values for black-list applications. At present, IBC with ROC space interpolation allows to design accurate systems for very low FMR values in the white-list case. In addition, increasing the number of Genuine samples would improve resolution of DET curves in lower FMR regions. In practical biometric applications, the underlying data distributions are highly imbalanced (the number of Imposter samples is significantly greater than Genuine ones) and may vary over time, and the misclassification costs cannot be specified exactly. Since the ROC curves are not influenced by asymmetries in Genuine and Impostor distributions, they can present an overly optimistic view of system accuracy. IBC should be adapted for other decision spaces that explicitly account for skewed data and misclassification costs.

The performance of iris recognition systems has been evaluated using a transaction-based analysis in the DET space, and, as described in Gorodnichy et al. (2011) and Gorodnichy (2011), FMRs and FNMRs allow for an order-1 analysis. However, this analysis presents a partial view of performance. Order-1 analysis does not, for instance, consider that a 1-to-N closed-set identification system intended for fully-automated access/border control applications may produce more than one match per transaction, indicating a lower confidence on decisions. The multi-order analysis presented in Gorodnichy (2011) provides a more comprehensive evaluation of performance. According to this analysis, a system with a worse DET curve may in fact be preferable to another system with a better DET curve. Furthermore, the performance of biometric systems may vary drastically from one person to the next, which is known as the 'Doddington zoo' effect (Tabassi, 2010). In subject-based analysis, the number of false matches and non-matches is assessed with different users in mind, rather than with the overall number of transactions (over the entire population of users).

Acknowledgements

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada.

Disclaimer

The results presented in this paper are intentionally made anonymous not to be associated with any production system or vendor product and are used solely for the tasks identified in this paper. In no way do the results presented in this paper imply recommendation or endorsement by the Canada Border Services Agency, nor do they imply that the products and equipment identified are necessarily the best available for the purpose.

References

Banfield, R., Hall, L., Bowyer, K. and Kegelmeyer, W. (2003) 'A new ensemble diversity measure applied to thinning ensembles', *Multiple Classifier Systems*, Vol. 2709, p.159.

Barreno, M., Cardenas, A. and Tygar, D. (2008) 'Optimal ROC for a combination of classifiers', *Advances in Neural Information Processing Systems (NIPS)*, January, p.20.

Bengio, S., Mariéthoz, S. (2007) 'Biometric person authentication is a multiple classifier problem', *Int'l Workshop on Multiple Classifier Systems*, Prague, Czech Republic.

Bergamini, C., Oliveira, L., Koerich, A. and Sabourin, R. (2009) 'Combining different biometric traits with one-class classification', *Signal Processing*, Vol. 89, pp.2117–2127.

Black, M.A., Craig, B.A. (2002) 'Estimating disease prevalence in the absence of a gold standard', *Statistics in Medicine*, Vol. 21, No. 18, pp.2653–2669.

Bouchaffra, D. and Amira, A. (2008) 'Structural hidden markov models for biometrics: fusion of face and fingerprint', *Pattern Recognition*, Vol. 41, No. 5, pp.852–867.

Bowyer, K.W., Hollingsworth, K. and Flynn, P.J. (2008) 'Image understanding for iris biometrics: a survey', *Computer Vision and Image Understanding*, Vol. 110, No. 2, pp.281–307.

Breiman, L. (1996) 'Bagging predictors', *Machine Learning*, Vol. 24, No. 2, August, pp.123–140.

Brown, G., Wyatt, J., Harris, R. and Yao, X. (2005) 'Diversity creation methods: a survey and categorisation', *Journal of Information Fusion*, Vol. 6, No. 1, pp.5–20.

Daugman, J. (2000) *Biometric Decision Landscapes*, Tech. Rep. UCAM-CL-TR-482, Universtity of Cambridge, UK.

Fawcett, T. (2004) *ROC Graphs: Notes and Practical Considerations for Researchers*, Tech. Rep. HPL-2003-4, HP Laboratories, Palo Alto, CA, USA.

Fawcett, T. (2006) 'An introduction to ROC analysis', *Pattern Recognition Lett.*, Vol. 27, No. 8, pp.861–874.

Fierrez-Aguilar, J., Garcia-Romero, D., Ortega-Garcia, J., Gonzalez-Rodriguez, J. (2005) 'Adapted user-dependent multimodal biometric authentication exploiting general information', *Pattern Recognition Letters*, Vol. 26, No. 16, pp.2628–2638.

Freund, Y. and Schapire, R.E. (1996) 'Experiments with a new boosting algorithm', *ICML 96*, pp.148–156.

Gorodnichy, D.O. (2011) 'Multi-order biometric score analysis framework and its application to designing and evaluating biometric systems for access and border control', *IEEE Workshop on Computational Intelligence in Biometrics and Identity Management*, 11–25 April, Paris, France, pp.44–53.

Gorodnichy, D.O., Dubrofsky, E., Hoshino, R., Khreich, W., Granger, E. and Sabourin, R. (2011) 'Exploring the upper bound performance limit of iris biometrics using score fusion and score calibration', *IEEE Workshop on Computational Intelligence in Biometrics and Identity Management*, 11–25 April, Paris, France.

Haker, S., Wells, W.M., Warfield, S.K., Talos, I-F., Bhagwat, J.G., Goldberg-Zimring, D., Mian, A., Ohno-Machado, L. and Zou, K.H. (2005) 'Combining classifiers using their receiver operating characteristics and maximum likelihood estimation', *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Vol. 3749, pp.506–514.

Ho, T.K. (1998) 'The random subspace method for constructing decision forests', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp.832–844.

Ho, T.K., Hull, J. and Srihari, S. (1994) 'Decision combination in multiple classifier systems', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 1, pp.66–75.

Jain, A., Nandakumar, K. and Ross, A. (2005) 'Score normalization in multimodal biometric systems', *Pattern Recognition*, Vol. 38, pp.2270–2285.

Jain, A., Ross, A. and Pankanti, S. (2006) 'Biometrics: a tool for information security', *IEEE Trans. on Information Forensics and Security*, Vol. 1, No. 2, pp.125–143.

Khreich, W., Granger, E., Miri, A. and Sabourin, R. (2010) 'Iterative boolean combination of classifiers in the roc space: an application to anomaly detection with hmms', *Pattern Recognition*, Vol. 43, No. 8, pp.2732–2752.

Khreich, W., Granger, E., Sabourin, R. and Miri, A. (2009) 'Combining hidden Markov models for anomaly detection', *International Conference on Communications (ICC)*, Dresden, Germany, pp.1–6.

Kittler, J., Hatef, M., Duin, R. and Matas, L. (1998) 'On combining classifiers', *IEEE Trans. on Pattern Analysis and Machine Intellignece*, Vol. 20, No. 3, pp.226–239.

Kittler, J., Poh, N., Fatukasi, O., Messer, K., Kryszczuk, K., Richiardi, J. and Drygajlo, A. (2006) 'Quality dependent fusion of intramodal and multimodal biometric experts', *Proc. of SPIE Vol. 6539, Biometric Technology for Human Identification IV*, pp.1–14.

Kuncheva, L.I. (2004) *Combining Pattern Classifiers: Methods and Algorithms*, Wiley-Interscience Publication, New Jersey, USA.

Kuncheva, L.I. and Whitaker, C.J. (2003) 'Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy', *Machine Learning*, Vol. 51, No. 2, pp.181–207.

Kung, S., Mak, M. and Lin, S. (2004) *Biometric Authentication: A Machine Learning Approach*, Prentice Hall, New Jersey, USA.

Nandakumar, K., Chen, Y., Dass, S. and Jain, A. (2008) 'Likelihood ratio-based biometric score fusion', *IEEE Trans. on Pattern Analysis and Machine Intellignece*, Vol. 30, No. 2, pp.342–347.

Neyman, J. and Pearson, E.S. (1933) 'On the problem of the most efficient tests of statistical hypotheses', *Royal Society of London Philosophical Transactions Series A*, Vol. 231, pp.289–337.

Oh, I-S. and Suen, C. (2002) 'A class-modular feedforward neural network for handwriting recognition', *Pattern Recognition*, Vol. 35, pp.229–244.

Oliveira, L., Justino, E., Sabourin, R. and Bortolozzi, F. (2008) 'Combining classifiers in the roc-space for off-line signature verification', *Journal of Universal Computer Science*, Vol. 14, No. 2, pp.237–251.

Oxley, M., Thorsen, S. and Schubert, C. (2007) 'A Boolean algebra of receiver operating characteristic curves', *10th International Conference on Information Fusion*, Québec City, Canada, pp.1–8.

Provost, F.J. and Fawcett, T. (2001) 'Robust classification for imprecise environments', *Machine Learning*, Vol. 42, No. 3, pp.203–231.

Roli, F., Fumera, G. and Kittler, J. (2002) 'Fixed and trained combiners for fusion of imbalanced pattern classifiers', *Information Fusion, 2002. Proceedings of the Fifth International Conference on*, Vol. 1, pp.278–284.

Ruta, D. and Gabrys, B. (2005) 'Classifier selection for majority voting', *Information Fusion*, Vol. 6, No. 1, March, pp.63–81.

Scott, M.J.J., Niranjan, M. and Prager, R.W. (1998) 'Realisable classifiers: improving operating performance on variable cost problems', in Lewis, P.H. and Nixon, M.S. (Eds.): *Proceedings of the Ninth British Machine Vision Conference*, September, University of Southampton, UK, Vol. 1, pp.304–315.

Shen, C. (2008) 'On the principles of believe the positive and believe the negative for diagnosis using two continuous tests', *Journal of Data Science*, Vol. 6, pp.189–205.

Snelick, R. (2005) 'Large-scale evaluation of multimodal biometric authentication state-of-the-art systems', *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 3, pp.450–455.

Tabassi, E. (2010) 'Image specific error rate: a biometric performance metric', *Proc. International Conference on Pattern Recognition*, pp.1124–1127.

Tao, Q. and Veldhuis, R. (2008) 'Threshold-optimized decision-level fusion and its application to biometrics', *Pattern Recognition*, Vol. 41, No. 5, pp.852–867.

Thomopoulos, S., Viswanathan, R. and Bougoulias, D. (1989) 'Optimal distributed decision fusion', *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 25, No. 5, pp.761–765.

Tulyakov, S., Jaeger, S., Govindaraju, V. and Doermann, D. (2008) 'Review of classifier combination methods', in Simone Marinai, H.F. (Ed.): *Studies in Computational Intelligence: Machine Learning in Document Analysis and Recognition*, Springer, pp.361–386.

Van Erp, M. and Schomaker, L. (2000) 'Variants of the borda count method for combining ranked classifier hypotheses', *Seventh International Workshop on Frontiers in Handwriting Recognition*, 11–13 September, Amsterdam, pp.443–452.

Verlinde, P., Chollet, G. and Achcrov, M. (2000) 'Multi-modal identity verification using expert fusion', *Information Fusion*, Vol. 1, No. 1, pp.17–33.

Walter, S.D. (2005) 'The partial area under the summary ROC curve', *Statistics in Medicine*, Vol. 24, No. 13, pp.2025–2040.

Zhang, T., Li, X. and Tao, D. (2008) 'Multimodal biometrics using geometry preserving projections', *Pattern Recognition*, Vol. 41, No. 5, pp.805—813.