IJDAR (2013) 16:83–103
DOI 10.1007/s10032-011-0180-6

ORIGINAL PAPER

Multi-feature extraction and selection in writer-independent off-line signature verification

Dominique Rivard · Eric Granger · Robert Sabourin

Received: 14 February 2011 / Revised: 7 October 2011 / Accepted: 5 November 2011 / Published online: 25 November 2011

Abstract Some of the fundamental problems faced in the design of signature verification (SV) systems include the potentially large number of input features and users, the limited number of reference signatures for training, the high intra-personal variability among signatures, and the lack of forgeries as counterexamples. In this paper, a new approach for feature selection is proposed for writer-independent (WI) off-line SV. First, one or more preexisting techniques are employed to extract features at different scales. *Multiple feature extraction* increases the diversity of information produced from signature images, allowing to produce signature representations that mitigate intra-personal variability. *Dichotomy transformation* is then applied in the resulting feature space to allow for WI classification. This alleviates the challenges of designing off-line SV systems with a limited number of reference signatures from a large number of users. Finally, *boosting feature selection* is used to design low-cost classifiers that automatically select relevant features while training. Using this global WI feature selection approach allows to explore and select from large feature sets based on knowledge of a population of users. Experiments performed with real-world SV data comprised of random, simple, and skilled forgeries indicate that the proposed approach provides a high level of performance when extended shadow code and directional probability density function features are extracted at multiple scales. Comparing simulation results to those of off-line SV systems found in literature confirms the viability of the new approach, even when few reference signatures are available. Moreover, it provides an efficient framework for designing a wide range of biometric systems from limited samples with few or no counterexamples, but where new training samples emerge during operations.

Keywords Biometrics · Handwriting recognition · Writer-independent signature verification · Feature extraction · Feature selection · Boosting · Decision tree classification · Incremental learning

D. Rivard · E. Granger (✉) · R. Sabourin
Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), École de technologie supérieure, 1100, rue Notre-Dame Ouest, Montreal, QC, H3C 1K3, Canada
e-mail: eric.granger@etsmtl.ca

D. Rivard
e-mail: rivard@livia.etsmtl.ca

R. Sabourin
e-mail: robert.sabourin@etsmtl.ca

1 Introduction

Biometrics has emerged from its extensive use in law enforcement and forensic sciences and is increasingly being adopted in a wide variety of civilian applications for enhanced security and privacy [1]. Biometric systems perform the recognition of individuals based on their physiological (i.e., face and fingerprint traits) and behavioral (i.e., voice print and handwritten signature) characteristics. Biometric traits are intrinsic to a person, and as such cannot be lost, stolen, or forgotten as with security tokens and secret knowledge [2]. Among the numerous biometric traits considered so far, handwritten signatures have long been established as one of the most widespread means for authenticating a person's identity by administrative and financial institutions. The procedure for acquisition of signature samples is familiar and noninvasive [3].

Biometric systems provide three recognition functions: identification, screening, and verification. Identification

seeks to establish a person's identity by matching his biometric sample against all user templates in the system database. Screening discreetly determines whether the biometric sample of an individual whose enrollment procedure is not typically well-defined matches any of the system's watch list of identities. Finally, verification authenticates the claimed identity of an individual by comparing his biometric sample to his template stored in the system database [4].

Given its behavioral nature, one of the main difficulties in handwritten signature verification is that an individual's signature can vary significantly from one sample to the next. In addition, a forger may attempt to reproduce signatures to bypass the system. Forgeries are usually divided into three types—random, simple, and simulated. A random forgery occurs when the forger does not know both the writer's name and the signature's morphology. It can also happen when a genuine signature presented to the system is mislabeled to another user. When the forger knows the writer's name but not the signature's morphology, the forger can only produce a simple forgery using a style of writing of his liking. A simulated forgery occurs when the forger has access to a sample to produce a reasonable imitation of the genuine signature.

Features extracted from handwritten signatures are broadly divided into two categories, static and dynamic, according to the acquisition method [5]. Static features are extracted by an off-line acquisition device after the writing process has been completed, while dynamic features are extracted by an online acquisition device during the writing process. By extension, automatic signature verification systems are either off-line or online. In either case, a neural or statistical classifier applied to signature verification is often trained using a limited number of training samples collected from a complex underlying distribution.

Two approaches have been proposed for off-line signature verification—writer-dependent (WD) and writer-independent (WI). The former approach models the signature of each individual from his samples, and a specialized classifier is trained for each writer. The WI approach uses a classifier to match each input questioned signature to one or more reference signatures, and a single classifier is trained for all writers [6]. Most automatic signature verification systems found in literature follow the WD approach. However, as in most biometric applications, the performance of these systems declines with large numbers of users and with limited number of reference signatures per user. For instance, in off-line verification of bank cheque signatures, the number of bank customers can easily reach the tens of thousands. In most cases, acquiring a sufficient number of reference signatures from each writer is not practical.

In contrast, the WI approach employs the dichotomy transformation to alleviate the difficulties of designing a verification system with a limited number of reference signatures from a large number of users. Off-line SV systems that follow the WI approach are designed in a space that is representative of the domain according to a population of writers (using some learning or development database) and hold several practical advantages. For one, this approach allows to exploit a system with only one signature per user. In addition, since input feature vectors are transformed into a distance space between signatures, the number of users is of little consequence. The signature of all writers is authenticated using a single two-class classifier. This classifier should be trained from a sufficient set of previously collected genuine signatures, although writers populating this learning set do not necessarily need to be enrolled to the system used during operations. The underlying hypothesis is that these dataset signatures are representative of the entire population of legitimate users enrolled to the verification system.

Several studies suggests that the accuracy and reliability of a biometric system can be improved by integrating the evidence obtained from multiple different sources of information [2]. Systems that employ multiple techniques to extract different feature types at multiple resolutions or scales may uncover diversified information from a given biometric sample. Since useful information may go undetected by using a single feature type and scale, these systems may improve the overall recognition rate [7]. In this respect, handwritten signature is a promising candidate since several powerful feature extraction techniques have been proposed in the literature [5,8]. Biometric sources of information are typically integrated at the sensor (raw biometric data), feature, score, and decision levels. Since the features extracted from sensor measurements contain richer information content about a biometric trait than scores, integration at the feature level should provide higher level of accuracy than at other levels.

This paper presents a novel approach for feature selection that is effective for the design of WI off-line signature verification systems. It is based on the combination of multiple feature extraction, dichotomy transformation, and boosting feature selection. Multiple feature extraction is adopted to extract several diverse handwritten signature representations from a signature image, using one or more preexisting feature extraction techniques at different levels of resolution or scales. Even though the approach applies with a wide range of feature extraction techniques (c.f., [5]), this paper considers that the representation and analysis of signature images are achieved by extracting features at multiple grid scales using two well-known grid-based techniques—extended shadow code (ESC) [9] and directional probability density functions (DPDF) [10]. While ESC extracts information about the spatial distribution of the signature, DPDF extracts information about the orientation of the strokes. These feature extraction techniques are seen as complementary, and once combined

into a single feature subset, they may provide a powerful multi-scale and spatio-directional representation of signatures.

In this paper, off-line signature verification is performed in a WI framework derived from a forensic document examination approach [11] and compared to the performance of state-of-the-art results using a database composed of 168 writers. Writer-independence is achieved by the verification system using the dissimilarity between each questioned signature and the reference signatures. Using the multiple feature extraction and dichotomy transformation provides signature representations with a large number of distance features and may therefore reduce the impact on performance caused by intra-personal variability. However, feature selection must be performed in the distance space to avoid the curse of dimensionality. Boosting feature selection is employed to efficiently select discriminant feature subsets from the potentially large number of features in the distance space, while training the classifier [12]. By virtue of the global WI (dissimilarity-based) approach, the proposed system selects features with knowledge of a population of users, which is difficult to achieve with a WD (feature-based) approach.

The specialized WI feature selection approach proposed in this paper allows learning a discriminant representation space from a corpus of signatures (i.e., development database) sampled from an independent population of writers that are not enrolled to the system evaluated during operations (or testing phases on some exploitation dataset). The approach proposed in this paper may be seen as an extension of the one presented in [13]. However, the approach proposed in this paper is based on the selection of representation spaces, not on the selection of classifiers, and the fusion of information is performed at the feature level instead of at the confidence score level. Moreover, the approach in this paper does not require selecting the best grid scale or set of grid scales. Boosting feature selection allows to perform low-cost feature learning, and several image zones of different sizes may be selected, with extraction techniques that are well adapted to the type of projection, type of stroke directions, etc. Finally, the dichotomy transformation employed with this approach allows to improve performance over time through incremental learning of new signature references and features during operations (exploitation phase), without having to retrain a system from the start on all cumulative training data.

The paper is organized as follows. Before presenting the multiple feature extraction and selection approach proposed in this paper (Sect. 3), Sect. 2 provides a survey of related WI systems that are suitable for off-line signature verification. In Sect. 4, the experimental methodology, including datasets, protocols, and performance metrics, are defined. In Sect. 5, simulation results are presented and discussed.

2 State-of-the-art in writer-independent signature verification

This section presents the state-of-the-art in WI systems for off-line signature verification. As with the standard WD case, WI systems take as input handwritten signatures and output verification results. However, as depicted by Fig. 1, when acquiring a new reference signature S_r (e.g., during enrollment), the corresponding feature vector $\mathbf{x}_r$ extracted from the signature image is stored for later use in the system's knowledge base. In verification mode, the image of a questioned signature S_q is presented to the system and its feature vector $\mathbf{x}_q$, along with the reference set $\{\mathbf{x}_r\}_1^R$ of signatures of the users enrolled to the knowledge base, are presented to the dichotomy transformation module. Then, the dichotomizer (two-class classifier) takes as input distance vectors $\{\mathbf{u}_r\}_1^R$ for each questioned signature S_q and produces the corresponding set of confidence scores $\{f(\mathbf{u}_r)\}_1^R$ that are combined to output a final decision $g(\mathbf{x}_q)$.

The rest of this section describes the dichotomy transformation [14], followed by its application to SV [11]. Then, an approach to WI signature verification based on dichotomizer ensembles [13] is reviewed.

2.1 Dichotomy transformation

A dichotomy transformation [14] allows to transform K-class pattern recognition problems where K is a *large* or *unspecified* value into a 2-class problem. In this context, the handwritten SV problem is formulated as follows. Given a reference signature and a questioned signature, the objective is to determine whether the two signatures were produced by the same writer. Formally, let $\mathbf{x}_q$ and $\mathbf{x}_r$ be two *feature vectors* from the feature domain labeled y_q and y_r, respectively, and let $\mathbf{u}_r$ be the *distance vector* in the distance domain resulting from the dichotomy transformation:

$$\mathbf{u}_r = \left|\mathbf{x}_q - \mathbf{x}_r\right| \tag{1}$$

where $|\cdot|$ is the absolute value. It is important to emphasize that each component of vector $\mathbf{u}_r$ equals the distance between the corresponding components of vectors $\mathbf{x}_q$ and $\mathbf{x}_r$, thus distance vector and feature vectors have the same dimensionality. In the distance domain, independently of the number of writers, there are only two classes: the *within* class $\omega_\oplus$ and the *between* class $\omega_\ominus$. The distance vector $\mathbf{u}_r$ is assigned the label v_r according to:

$$v_r = \begin{cases} \omega_\oplus & \text{if } y_q = y_r \\ \omega_\ominus & \text{otherwise.} \end{cases} \tag{2}$$

Intuitively, signatures from the same writer should be near one another in the feature space, thus clustering near the origin in distance space, whereas signatures from different writers should be distant from each other in the feature space

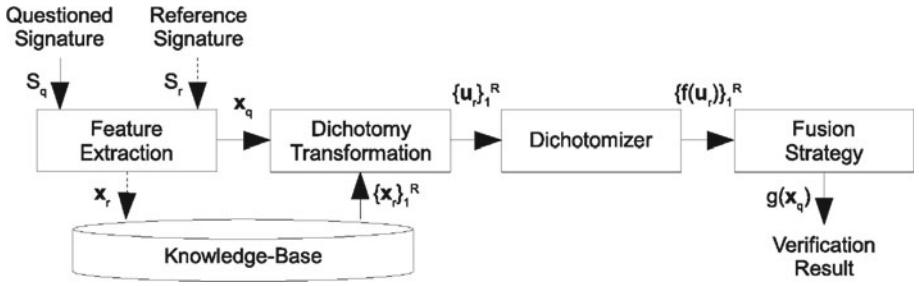

Fig. 1 A generic system for off-line WI signature verification. Enrollment process is indicated by *dotted arrows* while *solid arrows* illustrate the authentication process

and thus be scattered away from the origin in the distance space.

As for the number of distance vectors generated by the dichotomy transformation, if K writers provide a set of R references each, (1) generates up to $\binom{KR}{2}$ different distance vectors. Of these, $K\binom{R}{2}$ are of the *within* class and $\binom{K}{2}R^2$ are of the *between* class. Thus, using a small sample of references from each writer, the dichotomy transformation generates an appreciable quantity of samples in the distance domain.

Figure 2 presents an example to illustrate the dichotomy transformation. Suppose a set of three writers, $\{\omega_1, \omega_2, \omega_3\}$ and each writer provides three signatures. Some feature extraction technique produces a vector of two features $(x_1\, x_2)^{\mathrm{T}}$ from each signature. Figure 2a plots the feature vectors of signatures into the feature space. The dichotomy transformation calculates the distance between the features of each pair of signatures to form vectors $(u_1\, u_2)^{\mathrm{T}}$ in the distance space, as depicted in Fig. 2b.

The dichotomy transformation affects the geometry of distributions. In this example, multiple boundaries are needed to separate the three writers in the feature space as opposed to only one in the distance space. Also, the vectors in the distance space are always nonnegative since they consist of distances. Finally, the dichotomy transformation augments the number of samples in the distance space because they are made up of every pair of signatures.

To illustrate how the verification process is independent from the writer being verified, let $\mathbf{x}_q$, $\mathbf{x}_r$ be a questioned and a reference feature vectors, respectively, both from new writer ω_4. The dichotomy transformation (1) computes the distance vector $\mathbf{u}_r$ from $\mathbf{x}_q$ and $\mathbf{x}_r$. As it can be seen in the distance space (Fig. 2b), that distance vector $\mathbf{u}_r$ is located in the *within* region defined by the dichotomizer, which means that it authenticates both questioned and reference signatures as belonging to the same writer. On the other hand, the feature space boundaries (Fig. 2a) fail utterly by classifying one signature ($\mathbf{x}_q$) to writer ω_2 and the other ($\mathbf{x}_r$) to writer ω_3. In fact, it is impossible for the feature domain model to adequately classify the signatures as belonging to writer ω_4 since this writer did not contribute to the training set. Hence, the writer-independence is provided by the distance domain model.

One drawback of the dichotomy transformation is that perfectly clustered writers in feature domain may not be perfectly dichotomized in distance domain. In other words, the broader the spread of the feature distributions among the writers, the less the dichotomizer is able to detect real differences between similar signatures [14]. Thus, the performance of a dichotomizer is considerably affected by the choice of a feature set extracted from the handwritten signatures. Moreover, since the dichotomy transformation affects the spatial geometry of distributions, the best feature set may not be the same in feature domain as in the distance domain. The approach proposed in this paper (see Sect. 3) seeks to extract a large set of potential features and efficiently selects a small set of discriminant features in the distance domain.

2.2 Extension of dichotomy transformation for questioned document expert

The Questioned Document Expert's approach [11] is an extension to the dichotomy transformation that applies when users have more than one template stored in the knowledge base. The idea is to emulate the expert's approach, which consists of comparing the questioned signature input to the SV system to a set of genuine signatures. Each comparison leads to a partial decision from the expert, his/her final decision being based on all partial decisions. Intuitively, the more reference signatures that are available for comparison with the questioned signature, the more accurate the final decision will be.

Formally, the dichotomy transformation is applied between the questioned signature's feature vector $\mathbf{x}_q$ and the user's reference set $\{\mathbf{x}_r\}_1^R$ from the knowledge base and produces the set of distance vectors $\{\mathbf{u}_r\}_1^R$. The dichotomizer evaluates each distance vector individually and outputs a set of confidence values $\{f(\mathbf{u}_r)\}_1^R$ representing the partial decisions from the expert. The final decision of the system about the questioned signature is based on the fusion of all confidence values by a function $g(\cdot)$. The choice of the fusion function is dependent on the nature of the dichotomizer's output. For instance, if the output of the dichotomizer is a label, then the majority vote is an appropriate fusion strategy. On the other hand, if the output of the dichotomizer is a

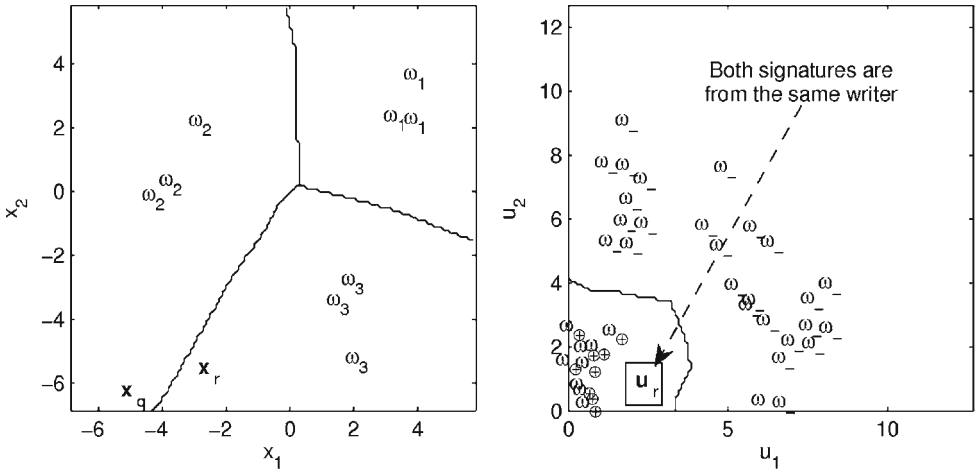

Fig. 2 Vectors from three different writers $\{\omega_1, \omega_2, \omega_3\}$ in feature space (*left*) projected into distance space (*right*) using the dichotomy transformation to form two classes $\{\omega_\oplus, \omega_\ominus\}$. Decision boundaries in both spaces are inferred by the nearest neighbor algorithm. The dichotomizer authenticates each questioned signature $\mathbf{x}_q$ with respect to reference signatures $\mathbf{x}_r$. The distance vector $\mathbf{u}_r$ resulting from a comparison is assigned to the *within* class, meaning both signatures belong to the same writer

probability, then a wider range of fusion strategies is available such as the sum, mean, median, max, and min functions, to name a few.

2.3 Ensemble of writer-independent dichotomizers

In [13], the original framework for WI off-line signature verification is improved by replacing the dichotomizer by an ensemble of dichotomizers. To achieve ensemble diversity, support vector machines (SVMs) are trained on a learning set of 40 writers using 16 different scales of the segmentation grid during feature extraction. The same four grid-based feature extraction techniques of [11] are used except for stroke curvature information, which is extracted based on cubic Bezier curves. Thus, a pool of 64 SVMs is overproduced, from which a genetic algorithm chooses a subset of base classifiers to form the final ensemble of dichotomizers. These dichotomizers are combined at the confidence score level using the sum rule fusion strategy.

Different objective functions are applied with the genetic algorithm and the authors conclude that maximizing the area under the receiver operating characteristic (ROC) curve, or AUC, is the most suitable. In all cases, the fitness of the objective function is evaluated on an independent validation set of 20 writers. The influence of the number of reference signatures in the validation set is evaluated by increasing their numbers from 3 to 15, repeating the ensemble optimization every time. The authors conclude that the authentication rate depends on a trade-off between the number of references and the intra-class variability of the reference set.

Overall, authentication using an ensemble of dichotomizers shows an improvement over 2-class classification by a single dichotomizer. However, this approach complicates the verification system, specifically in [13], ensembles count an average of 13 SVMs, using a total of 2,300 features and thus increasing the use of resources while reducing recognition speed. Moreover, the ensemble optimization process being stochastic in nature, each run may lead to a different ensemble, as demonstrated by their results. An improvement to the authentication rate is sought with the approach proposed in this paper (see Sect. 3). Results should improve by extracting different feature types at different scales and with classifier ensembles where the measurements are integrated at the confidence score level. However, ensembles are typically more effective to integrate different sources of information as early as possible in the verification system [2]. Integrating combined information at feature level should yield improvements since at this level information about the signatures is richer. The WI approach proposed in this paper is based on the selection of representation spaces, and the fusion of information occurs at the feature level. Moreover, this approach does not require selecting the best grid scale or set of grid scales. Boosting feature selection allows to perform low-cost feature learning, and several image zones of different sizes may be selected.

3 A framework for multi-feature extraction and selection

This section introduces a new approach for multi-feature extraction and selection that is efficient in WI off-line signature verification. This novel system uses an ensemble of dichotomizers to combine features across several scales and feature extraction techniques, leading to low cost and accurate SV. As depicted in Fig. 3, the proposed approach may be viewed in terms of the generic system for WI off-line SV shown in Fig.1.

However, as described in the following subsections, it uses (i) the multiple feature extraction at different scales and (ii) the boosting feature selection (BFS) technique for classification.

Multiple feature extraction can be applied to reference signatures S_r and questioned signatures S_q using one or more

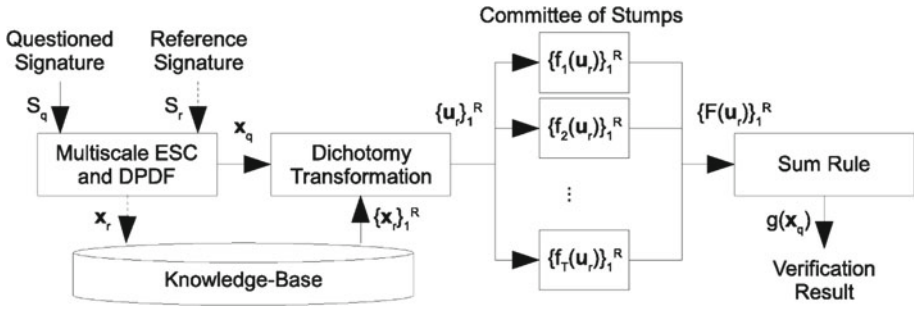

Fig. 3 Overview of a WI off-line SV system that applies a new approach for multi-feature extraction and selection. It is based on ESC and DPDF for multiple feature extraction, and the BFS algorithm to design a committee of stumps for classification

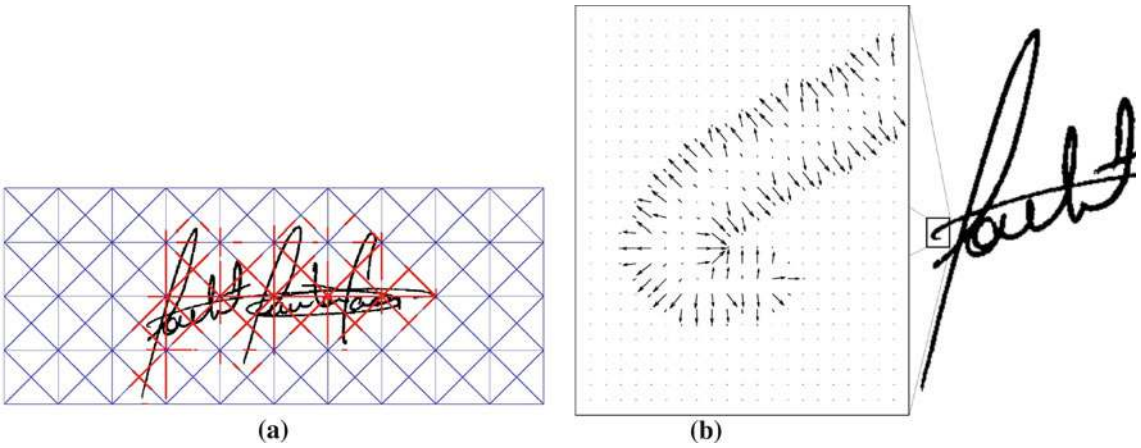

Fig. 4 **a** Example of the ESC technique applied to the extraction of features from a binary signature image. **b** Gradient of a handwritten signature binary image. *Arrows* indicate direction and magnitude of the gradient at each pixel location

preexisting feature extraction techniques, producing a large set of features. In this paper, the representation of signature images is achieved by using two grid-based techniques—ESC and DPDF. During a preliminary design phase, the BFS algorithm performs feature learning on a reference set of distance vectors and produces a committee of stumps, where each stump corresponds to a selected feature. During operations, the committee of stumps takes as input the distance vectors $\{\mathbf{u}_r\}_1^R$ for each questioned signature S_q. It produces the corresponding set of the committee's confidence scores, $\{F(\mathbf{u}_r)\}_1^R$, which are combined to output a final decision $g(\mathbf{x}_q)$.

3.1 Multiple feature extraction

For handwritten signatures, it is important for the feature extraction process to be text insensitive. In other words, the measurements taken on signature must not rely on the segmentation of specific letters, which can be a very difficult task especially if the signature is highly personalized [9]. A practical alternative is to partition the signatures using a virtual grid and to take local measurements in each of the grid cells. By varying the scale of the virtual grid, purely global to very local features are extracted. In the literature, grid-based approaches generally tend to find a grid scale suitable to their signature database. Here, the proposed approach is to extract features at multiple scales and let the classifier select the most suitable ones.

When a new signature is used for enrollment or operations, it is presented to the system as a gray-level image. From there, two preprocessing steps are necessary in order to prepare the signature for feature extraction. First, the signature is automatically segmented from its background using Otsu's threshold selection method from gray-level histograms [15]. According to questioned document experts, the proportion and orientation of handwritten signatures are intrinsic characteristics of the writer when guided by a form [16]. Consequently, the second preprocessing step corrects the binary signature images in translation by aligning their centroid with the center of the feature extraction grid.

The extended shadow code (ESC) [9,17] consists in the superposition of a bar mask array over the binary image of a handwritten signature as depicted by Fig. 4a. Each bar is assumed to be a light detector related to a spatially constrained area of the 2D signal. A shadow projection is defined as the simultaneous projection of each black pixel into its closest horizontal, vertical, and diagonal bars. A projected shadow turns on a set of bits distributed uniformly along the

bar. After all the pixels on a signature are projected, the number of *on* bits in each bar is counted and normalized to the range [0, 1] before features are extracted. Given a virtual grid composed of I rows by J columns, the cardinality of the ESC feature vector is equal to

$$\left|\mathbf{x}^{\mathrm{ESC}}\right| = 4IJ + I + J. \tag{3}$$

Directional probability density functions (DPDF) [10] have been used as a global shape factor for automatic offline handwritten signature verification. The rationale of this approach is that the stroke orientation of handwritten signatures is stable enough to properly discriminate writers. Thus, this technique extracts features based on the frequency distribution of the orientation of the gradient at the edge of the signature. Gradient features are used by other signature verification systems, for instance [18]. In this work, local DPDF are extracted from within each cell of a virtual grid placed over the handwritten signature image. This way, local information is extracted from different parts of the signature, consequently increasing its discriminating power. Moreover, the information extracted from the signature is complementary to that extracted using the ESC technique. While the ESC extracts information about the spatial distribution of the signature, DPDF extracts information about the orientation of the strokes. Since the same grid scale is used for both techniques, this leads to a powerful spatio-directional representation of handwritten signatures.

The gradient is computed from the binarized version of the signature image after it has been smoothed using a Gaussian low-pass filter to reduce the impact that residual noise can have on the two key derivatives used for gradient computation. Computing the gradient on a smoothed binary image has the definitive advantage that the intensity of the image is already normalized; consequently, there is no need of a threshold to detect the edges of the image. In fact, the background segmentation process has already managed to detect the edges of the signature and thus, the remaining task is to determine their orientation. This work uses Sobel operators to compute the gradient key derivatives.

Figure 4b illustrates the gradient of a handwritten signature image using arrows to indicate its direction and magnitude at each pixel location. Since the signature consists of a binary image, gradient is null in the background of the image and within the strokes of the signature where intensity is constant. Gradient is non-null along the edges of the signature and its direction varies perpendicularly to the contour of the signature.

In order to obtain a fixed number of features, gradient directions are quantized into an even Φ number of ranges. A greater Φ results into a more exact representation of the gradient of the signature, thus increasing between-writers discrimination. However, the more exact the representation, the more sensitive it is to intra-personal variance, thus lowering generalization capabilities. As a trade-off, this work uses $\Phi = 8$. It is important to realize that a quantized value $\phi \in \left\{1, 2, \ldots, \frac{\Phi}{2}\right\}$ indicates the same stroke orientation as value $\phi + \frac{\Phi}{2}$. Thus, after quantization, gradient magnitudes are summed according to each stroke orientation for every individual cell of the virtual grid, leading to a feature vector of cardinality

$$\left|\mathbf{x}^{\mathrm{GRD}}\right| = \frac{1}{2} IJ\Phi. \tag{4}$$

Note that both feature extraction techniques can be executed in parallel to reduce computation time.

3.2 Boosting feature selection

Boosting is a machine-learning procedure, which combines the performance of many weak classifiers into a powerful committee. The rationale behind boosting is that finding many moderately inaccurate rules of thumb using many simple classifiers can be easier than finding a single highly accurate prediction rule using a more elaborate learning algorithm. Boosting methods have proven to be very competitive in terms of generalization in a variety of applications [19]. The general idea of boosting is to form a committee of weak classifiers iteratively by adding one weak classifier at a time. At the beginning of the training procedure, a uniform weighting is assigned to the patterns of the training data set. Each time, a new classifier is added to the committee, the samples in the training data are re-weighted to reflect the performance of this weak classifier, assigning more importance to misclassified samples. Thus, the next weak classifier focuses on more difficult samples, and the procedure ends after a predefined number of weak classifiers have been trained.

The problem of feature selection is defined as follows: given a set of potential features, the objective is to select the best subset under some classification objectives. This procedure has three goals: (i) to reduce the cost of extracting features, (ii) to improve the classification accuracy, and (iii) to improve the reliability of the estimate of performance [20]. The boosting feature selection algorithm [12] (and further studied in [21]) explicitly incorporates feature selection into AdaBoost [22], the most commonly used variant of boosting. Boosting feature selection is performed by designing a weak classifier that selects the single most discriminant feature of a set of potential features and finds a threshold to separate the two classes to learn, effectively a decision stump. Consequently, features are selected in a greedy fashion according to the weighting *while* learning is conducted by the boosting algorithm. Given a very large set of features, the result is a committee built on the best subset of features representing the training data.

The next subsections describe the boosting algorithm and the weak classifier used in this work and are followed by a complexity study of the resulting committee.

3.2.1 Gentle AdaBoost

The problem of handwritten signature verification can have a significant class overlap, especially between genuine signatures and simulated forgeries, and as mentioned previously, the dichotomy transformation can exacerbate this phenomenon. Also, it has been observed by several authors that AdaBoost is not an optimal method on very noisy problems [23–25]. By design, Adaboost focuses on misclassified samples and this may results in fitting the noise.

Several boosting methods address the overfitting problem, mostly by adjusting the weighting scheme. For instance, MadaBoost [26] bounds the weight assigned to each sample by its initial probability, Gentle AdaBoost [27] takes adaptive Newton steps to update the weights more slowly, BrownBoost [28] uses a non-monotone weighting function decreasing the weight of samples far from the margin, AdaBoost$_\tau$ [24] and AdaBoost$^*_\nu$ [29] both use the concept of soft margin to regularize by allowing for misclassifications, SmoothBoost [25] constructs smooth distributions, which do not put too much weight on any single sample, and NadaBoost [30] prevents high weight values by thresholding.

Moreover, validation sets have long been used in machine learning to limit overfitting, and, as noted by the authors of Adaboost [31], a validation set could be used for early stopping. This work makes use of Gentle Adaboost and early stopping to address the significant class overlap problem. Early stopping is implemented using a holdout validation set. The early stopping criterion is based on the maximization of area under the receiver operating characteristics curve (AUC) on the holdout validation set.

Receiver operating characteristics (ROC) curves are graphs plotting the true positive rate of a classifier in function of its false positive rate. The points composing the curve are obtained by varying the decision threshold of the classifier (see Algorithm 1 of [32] for an efficient method for the generation of ROC points). ROC curves have an attractive property: they are insensitive to change in class distribution [32]. If the proportion of genuine signatures and forgeries changes between the design of a system and its exploitation, the ROC curves will not change. It is the case with signature verification applications, as the proportions of fraud for real applications are likely to vary in time and from place to place. The AUC of a ROC curve has an important statistical property: it is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [32]. As such, the AUC is invariant to the decision threshold optimized by Gentle AdaBoost, which is a significant advantage in the context of this work since the decision threshold is learned using random forgeries as counterexamples when in fact the committee is tested against random, simple, and simulated forgeries. Moreover, the decision threshold is a function of the priors and the classification cost, both of which are likely to vary in a signature verification application.

Figure 5 describes the Gentle AdaBoost algorithm with early stopping. Let $\mathcal{L} = \{\mathbf{u}_l, v_l, w_l\}_1^L$ be a learning set of L feature vectors $\mathbf{u} \in \mathbb{R}^D$ labeled to $v_l \in \{-1, 1\}$ and weighted by the distribution $w_l \in [0, 1]$, $\sum_{l=1}^{L} w_l = 1$. Similarly, let $\mathcal{H} = \{\mathbf{u}_h, v_h\}_1^H$ be a holdout validation set of H non-weighted samples. Let also $T_\mathcal{L}, T_\mathcal{H} \in \mathbb{N}$ be the maximum iteration stopping criterion and the early stopping criterion, respectively. The first half of the algorithm implements the four steps of the Gentle AdaBoost algorithm: (i) train a new decision stump $f_t(\mathbf{u})$ based on $\mathcal{L}$, (ii) add $f_t(\mathbf{u})$ to the committee $F(\mathbf{u})$, (iii) update the weights w_l according to the response of $f_t(\mathbf{u}_l)$, and (iv) renormalize the weights to ensure a distribution. The second half implements a holdout validation scheme using the AUC for criterion. The AUC is computed using Algorithm 2 described in [32]. If the algorithm reaches $T_\mathcal{H}$ iterations without increase in the AUC, it early stops. The Gentle AdaBoost with early stopping procedure outputs the committee $F(\mathbf{u})$ composed of T decision stumps.

3.2.2 Decision stumps

Decision trees classify a pattern through a sequence of questions [33]. Each question tests a single feature of the data and is represented by a tree node, the first question being the root of the tree and each possible decision spanning a branch to new node (i.e., the next question) and so on until a terminal node, called a leaf, is reached and the pattern is classified. Each decision outcome is called a split since it effectively splits the data into subsets, binary decisions being referred to as single splits and higher number of decisions as multi-splits.

Decision stumps are one-level, single split trees [34]. A decision stump $f_t(\mathbf{u})$ is composed of four parameters: d_t the dimension to split, τ_t the splitting threshold in that dimension, and ρ_t^{left} and ρ_t^{right} the weighted means of the response for the left and right leaves, respectively. Figure 6 illustrates the decision stump learning algorithm. The first step in the algorithm is to compute $W_\oplus$, $W_\ominus$ the positive, and negative weight totals, respectively. Then, the algorithm independently searches each problem dimension to find the best split point (d_t, τ_t). For a given dimension, the samples are first sorted by increasing feature values. Then, the algorithm computes $w_\oplus$, $w_\ominus$ the positive, and negative weight cumulative distribution functions, respectively. Based on the CDFs and the weight totals, the splitting threshold is selected to minimize the probability that a training sample would be

Inputs: Learning set $\mathcal{L} = \{\mathbf{u}_l, v_l, w_l\}_1^L$, hold-out validation set $\mathcal{H} = \{\mathbf{u}_h, v_h\}_1^H$, early stopping criterion $T_{\mathcal{H}}$, maximum iteration stopping criterion $T_{\mathcal{L}}$.
Output: Classifier $\text{sign}\left[F\left(\mathbf{u}\right)\right] = \text{sign}\left[\sum_{t=1}^{T} f_t\left(\mathbf{u}\right)\right]$.
Initialize: $w_l = 1/L$, $F = \emptyset$, and local variable $A_{\text{max}} = -\infty$.
for $t = 1$ to $T_{\mathcal{L}}$ **do**
  /* Gentle AdaBoost algorithm */
  Train $f_t\left(\mathbf{u}\right)$ using $\mathcal{L}$.
  Update $F\left(\mathbf{u}\right) \leftarrow F\left(\mathbf{u}\right) + f_t\left(\mathbf{u}\right)$.
  Update $w_l \leftarrow w_l \exp\left(-v_l f_t\left(\mathbf{u}_l\right)\right)$.
  Renormalize $w_l \leftarrow w_l / w_{\text{tot}}$, where $w_{\text{tot}} = \sum_l w_l$.
  /* Early stopping check */
  if $\text{AUC}\left(\mathcal{H}, F\right) > A_{\text{max}}$ **then**
    Update $A_{\text{max}} = \text{AUC}\left(\mathcal{H}, F\right)$.
    Update $T = t$.
    Reset $counter = 0$.
  else
    Increment $counter \leftarrow counter + 1$.
    if $counter = T_{\mathcal{H}}$ **then**
      Exit by early stopping.
    end if
  end if
end for

Fig. 5 Gentle AdaBoost algorithm with early stopping

Inputs: Pre-sorted weighted learning set $\mathcal{L} = \{\mathbf{u}_l, v_l, w_l\}_1^L$.
Output: Decision stump $f_t(\mathbf{u})$ with parameters d_t, τ_t, ρ_t^{left} and ρ_t^{right}.
Initialize: ρ_t^{left}, $\rho_t^{right} \leftarrow 0$, and local variables $\varepsilon_{tot} \leftarrow \infty$, $W_{\oplus}$, $W_{\ominus}$, W_{left}, $W_{right} \leftarrow 0$.
/* Compute total positive and negative weights */
$W_{\oplus} = \sum_l w_l$ where $v_l = 1$
$W_{\ominus} = 1 - W_{\oplus}$
/* Find the best split point for this learning set */
for $d = 1$ to D **do**
  Sort $\mathcal{L}$ by increasing u_d values
  $w_{\oplus} \leftarrow 0$, $w_{\ominus} \leftarrow 0$
  for $l = 1$ to $L - 1$ **do**
    if v_l is a positive sample **then**
      $w_{\oplus} \leftarrow w_{\oplus} + w_l$
    else
      $w_{\ominus} \leftarrow w_{\ominus} + w_l$
    end if
    if $u_{d,l} \neq u_{d,l+1}$ **then**
      $\varepsilon_{left} \leftarrow 1 - \max\{w_{\oplus}, w_{\ominus}\}$
      $\varepsilon_{right} \leftarrow 1 - \max\{W_{\oplus} - w_{\oplus}, W_{\ominus} - w_{\ominus}\}$
      if $\varepsilon_{left} + \varepsilon_{right} < \varepsilon_{tot}$ **then**
        $\varepsilon_{tot} \leftarrow \varepsilon_{left} + \varepsilon_{right}$
        $d_t \leftarrow d$
        $\tau_t \leftarrow \frac{u_{d,l} + u_{d,l+1}}{2}$.
      end if
    end if
  end for
end for
/* Compute weighted means of the response in the left and right leaves of the stump */
$\rho_t^{left} = \sum_l w_l \cdot v_l / \sum_l w_l$ where $u_{l,d_t} < \tau_t$
$\rho_t^{right} = \sum_l w_l \cdot v_l / \sum_l w_l$ where $u_{l,d_t} \geq \tau_t$
end

Fig. 6 Decision stump learning algorithm

misclassified. Once the split point is optimized, the algorithm computes the weighted means of the response for both leaves.

When presented a sample $\mathbf{u}$ to classify, the decision stump $f_t(\mathbf{u})$ thresholds the feature d_t of $\mathbf{u}$ at τ_t and assigns the sample to the corresponding leave. Formally:

$$f_t(\mathbf{u}) = \begin{cases} \rho_t^{\text{left}} & \text{if } u_{d_t} < \tau_t \\ \rho_t^{\text{right}} & \text{otherwise.} \end{cases} \tag{5}$$

Decision stumps typically have high bias and low variance. However, boosting algorithms are capable of reducing both bias and variance, hence the increase in performance from committee of stumps [27]. Moreover, when boosting implements a re-weighting strategy (as opposed to re-sampling), like it is the case for Gentle AdaBoost, decision stumps cause boosting to become deterministic in the sense that multiple runs on the same learning set will result in identical committees. Also, the order of the presentation of the samples and of the features of the learning set do not affect the resulting committees. Finally, following (5), a decision stump classifies a pattern using a single feature. This means that boosted decision stumps will greedily select informative features *while* building the committee, ignoring redundant and irrelevant features. It is worth noting that the committee may learn several stumps based on the same feature, each with a different decision threshold and response.

"Appendix A" presents a detailed analysis of the time complexity for feature learning of a learning dataset with BFS and for classification using a committee of decision stumps. During operations, the total worst-case time required to classify a distance vector using the committee of stumps is $\mathcal{O}(T)$, making for very fast classifications. When training with larger databases ($L \gg T$) as considered in this research, the total average case time required to learn with Gentle AdaBoost and early stopping, including the time for quicksort, training decision stumps, and computing the area under the ROC curve is $\mathcal{O}(DLT)$, resulting in a fast learning algorithm that scales linearly.

4 Experimental methodology

The objective of this experimental protocol is to assess and compare the performance of the new approach proposed in Sect. 3 for feature selection for WI off-line SV. WI verification implies that there is only one classifier for all writers. Therefore, the protocol employs two disjoint sets of writers for system design (during development phase) and system testing (during exploitation phase). The underlying hypothesis is that the set of writers used for training is representative of the set of writers encountered during exploitation.

In most real-world applications, few genuine signatures are available for each writer, and random forgeries are typically the only type of counterexample available for designing a SV system. Consequently, only random forgeries were included in the training database, although the system is tested against random, simple, and simulated forgeries.

The following subsections describe the signature database, feature sets, and protocols used to evaluate performance. In fact, there is a generic protocol that consists in using an increasing number of reference signatures per writer, and this protocol is evaluated under different settings of features and their combination. These settings allow to evaluate performance for (1) single-scale representations, (2) information fusion at the feature extraction level, (3) information fusion at the confidence score level, and (4) fast incremental learning.

4.1 Signature database

The signature database used in this work is composed of 168 writers divided into a 108 writer development database $\mathcal{D}$ and a 60 writer exploitation database $\mathcal{E}$. As described in [13], the signatures were provided by 168 under-graduated students in four different sessions, ten samples at a time, once a week during one month, for a total of 40 genuine signatures per writer. The signatures were collected on a white A4 sheet of paper with no overlap and then scanned in gray level with 300 dpi. Regarding forgeries, ten people with no experience in producing forgeries were selected as forgers, to produce one simple and one simulated forgery for the 60 first writers. Simple forgeries were produced by supplying only the name of the writer to the forger. Simulated forgeries were produced by showing the forger four genuine signatures of the writer.

To build a WI classifier, distance vectors must be computed from the feature vectors using the dichotomy transformation as explained in Sect. 2.1. The learning set $\mathcal{L}$ and holdout validation set $\mathcal{H}$ are both generated from the development database $\mathcal{D}$. To do so, the 40 genuine signatures of each writer in $\mathcal{D}$ are partitioned into subsets of 30 and 10 signatures denoted $\mathcal{D}_1$ and $\mathcal{D}_2$, respectively. The signatures populating $\mathcal{D}_1$ are selected randomly.

The learning set $\mathcal{L}$ is generated using exclusively the genuine signatures of subset $\mathcal{D}_1$. The *within* class samples are computed using all genuine signatures from every writer, giving $108 \cdot \frac{30 \cdot 29}{2} = 46,980$ distance vectors. To generate an equivalent number of counterexamples, the dichotomy transformation is applied, for each writer, to 29 genuine signatures used as signatures of references against 15 random forgeries selected from the genuine signatures of 15 other writers. The result is $108 \cdot 29 \cdot 15 = 46,980$ *between* class distance vectors. Thus, the learning set is defined as $\mathcal{L} = \{\mathbf{u}_l, v_l\}_{l=1}^{93,960}$ with equiprobable priors.

The holdout validation set $\mathcal{H}$ is generated using the genuine signatures of subset $\mathcal{D}_1$ as signatures of reference against the genuine signatures of subset $\mathcal{D}_2$. Since each writer has 30 references signatures in $\mathcal{D}_1$ and 10 genuine signatures in $\mathcal{D}_2$,

the number of *within* class samples is equal to $108 \cdot 30 \cdot 10 = 32,400$. To generate an equivalent number of counterexamples, for each writer, 10 random forgeries are selected from 10 different writers in $\mathcal{D}_2$. The random forgeries are compared to the 30 references from $\mathcal{D}_1$, giving $108 \cdot 30 \cdot 10 = 32,400$ *between* class distance vectors. Thus, the holdout validation set is defined as $\mathcal{H} = \{\mathbf{u}_h, v_h\}_{h=1}^{64,800}$ with equiprobable priors.

To perform a WI evaluation of the system, both reference $\mathcal{R}$ and questioned $\mathcal{Q}$ sets are generated from the exploitation database $\mathcal{E}$ whose writers are unknown to the verification system. The reference set $\mathcal{R}$ is composed of 30 randomly selected genuine signatures from each writer of the exploitation database $\mathcal{E}$. Thus, the reference set is defined as $\mathcal{R} = \{\mathbf{x}_r, y_r\}_{r=1}^{1,800}$. The questioned set $\mathcal{Q}$ is composed of the 10 remaining genuine signatures and simple and simulated forgeries from each writer plus 10 random forgeries selected from the genuine signatures of 10 different writers. Thus, the questioned set is defined as $\mathcal{Q} = \left\{\mathbf{x}_q, y_q\right\}_{q=1}^{2,400}$.

4.2 Feature sets

ESC feature vectors and DPDF vectors are extracted from both signature databases $\mathcal{D}$ and $\mathcal{E}$. Resolution depends on the size of the extraction grid; the smaller the grid cells, the higher the resolution. The highest resolution used in this work is a cell of 20×20 pixels. Since the width of a stroke measures an average of 10 pixels, higher resolutions would results mostly in saturated cells and empty cells. On the other hand, the lowest resolution is limited by the size of the image. Since the signature images are 400 pixels high by 1,000 pixels wide, the lowest resolution consists of a single cell of that size. Let $\mathcal{I} = \{1, 2, 5, 10, 20\}$ be a set of 5 horizontal scales defined by their number of grid rows and $\mathcal{J} = \{1, 3, 6, 12, 25, 50\}$ be a set of 6 vertical scales defined by their number of grid columns. The Cartesian product $\mathcal{I} \times \mathcal{J}$ results in the 30 single scales used in this work. Finally, multiple features are achieved by combining feature sets from every scale, for a total of 15,457 and 14,744 features for ESC and gradient histogram, respectively, and a grand total of 30,201 features when both techniques are combined.

4.3 Single-scale representations

In order to independently characterize both ESC and DPDF feature extraction techniques, single-scale committees are designed from the development database $\mathcal{D}$ according to the following protocol. For each available feature set, the Gentle AdaBoost algorithm with early stopping (see Fig. 5) builds a committee of decision stumps on the learning set $\mathcal{L}$ using the holdout validation set $\mathcal{H}$ to prevent overfitting. The early stopping criterion $T_{\mathcal{H}} = 100$ and the maximum iteration stopping criterion $T_{\mathcal{L}} = 100,000$.

The performance of the WI committees is evaluated with the exploitation database $\mathcal{E}$, using reference signatures from new writers in the set $\mathcal{R}$ to authenticate the questioned signatures of set $\mathcal{Q}$. To measure the impact of the cardinality of the reference set, reference subsets containing $1, 3, \ldots, 15$ randomly selected signatures are used for authentication. To simulate the effect of enrolling new signatures over time, previously selected reference signatures are kept and new signatures are added to increase the size of the references subsets. This procedure is repeated 100 times for variance estimation.

ROC analysis is used to compare the performance of the committees on the questioned set. Additionally, the committees are evaluated using their error rates on the questioned set. To do so, the decision threshold minimizing the zero-one loss is used as it permits a characterization of the questioned set and also a comparison with previous systems.

4.4 Multi-feature information fusion at the feature extraction level

Biometric sources of information are typically integrated at the sensor (raw biometric data), feature, score, and decision levels. Contrary to other WI systems [13], which implement information fusion at the confidence score level, the proposed system implements information fusion at the feature extraction level. Since the features extracted from sensor measurements contain richer information content about a biometric trait than scores, integration at the feature level may provide higher level of accuracy. Three multi-feature committees are compared: (i) using a multi-feature feature set only from the ESC technique, (ii) using a multi-feature feature set only from the DPDF technique, and (iii) using a multiple feature set from both ESC and DPDF techniques. Multiple feature sets are achieved by concatenating appropriate single-scale feature sets. To permit a straightforward result comparison, training and evaluation protocols are the same as for single-scale feature set, as described previously.

4.5 Multi-feature information fusion at the confidence score level

In order to compare the approach proposed in this paper to the overproduce and choose approach used in [13], the latter is shown here using the same feature extraction techniques and database partitions used in this paper. Since fusion is performed at the confidence score level, committees designed from single-scale feature sets are combined into an ensemble of committees by summing their individual confidence levels. Ensembles of committees are optimized using a genetic algorithm based on bit representation, one-point crossover, bit-flip mutation, and roulette wheel selection with elitism. Parameters are set as in [13], with a population = 100,

for replication 1 to 100 **do**
  Initialize $\mathcal{H} = \{\emptyset\}$
  Initialize $F_{\text{eoc}} = 0$
  Initialize $A_{\text{max}} = -\infty$
  for $p = 1$ to 60 **do**
    Acquire new handwritten signature representation datasets $\mathcal{H}_p, \mathcal{L}_p, \mathcal{Q}_p, \mathcal{R}_p$ from feature extraction expert
    Build committee of stumps $F_p \leftarrow GentleAdaBoost(\mathcal{L}_p, \mathcal{H}_p)$
    Add $F_{\text{eoc}} \leftarrow F_{\text{eoc}} + F_p$
    Add $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{H}_p$
    if $AUC(\mathcal{H}, F_{\text{eoc}}) > A_{\text{max}}$ **then**
      Update $A_{\text{max}} \leftarrow AUC(\mathcal{H}, F_{\text{eoc}})$
      Add $\mathcal{Q} \leftarrow \mathcal{Q} \cup \mathcal{Q}_p$
      Add $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_p$
      Evaluate F_{eoc} on $\mathcal{Q}$ using $1, 3, \ldots, 15$ references from $\mathcal{R}$
    else
      Dismiss $\mathcal{H} \leftarrow \mathcal{H} \setminus \mathcal{H}_p$
      Dismiss $F_{\text{eoc}} \leftarrow F_{\text{eoc}} - F_p$
    end if
  end for
end for

Fig. 7 Forward selection of representations protocol

number of generations = 300, probability of crossover = 0.7, and probability of mutation = 0.03. The chromosomes are composed of 60 bits, that is, one bit per single-scale committee and for a given chromosome. Bits with a value of "1" indicate the selected committees. The fitness function is the maximization of the AUC on the holdout validation set $\mathcal{H}$. Once an ensemble of committees is optimized, it is evaluated on the questioned set using the evaluation protocol described in Sect. 4.3. This procedure is also repeated 100 times for variance estimation, thus allows direct comparison with the other approaches explored in this work.

4.6 Fast incremental learning

In order to demonstrate the modularity of the proposed system, a third experimental protocol implements a fast incremental learning of handwritten signature representations. For instance, suppose that domain experts extract new representations from the design database. As new signature representations become available, they are learned incrementally by the verification system in order to increase its recognition rate.

A greedy incremental learning scheme is implemented to demonstrate the modularity of the proposed system. Single-scale signature representations are presented to the system one at a time. For each representation, a committee of stumps is built and combined with previous committees to form an ensemble of committees. Then, the AUC of the ensemble of committees is evaluated on the holdout validation set $\mathcal{H}$ and the newly added committee is kept if it improves the AUC of the ensemble or dismissed otherwise. The 60 single-scale representations are presented in random order and this procedure is replicated 100 times to evaluate the variance of the learning scheme.

Figure 7 details this experimental protocol. Given a signature representation p, the learning set $\mathcal{L}_p$ and holdout validation set $\mathcal{H}_p$ are both generated from the design database $\mathcal{D}$ and the reference set $\mathcal{R}_p$ and questioned set $\mathcal{Q}_p$ are generated from the exploitation database $\mathcal{E}$ whose writers are unknown to the verification system. Committees of stumps are the result of the Gentle AdaBoost algorithm with early stopping, described at Fig. 5, and they are evaluated according to the evaluation protocol described in Sect. 4.3.

Finally, Fig. 8 describes the experimental protocol designed to evaluate the impact of the quantity of signature representations on the boosted feature selection algorithm with early stopping. Let $\mathcal{L}$, $\mathcal{H}$, $\mathcal{Q}$, $\mathcal{R}$ be the learning, holdout, questioned, and references sets, respectively. They are all initialized as empty sets. Then, the 60 signature representations are randomly selected one at a time and added to the sets. When the sets contain 1, 5, 10, 15, 20, 25, 30, 40, 50, and 60 representations, the committee of stumps F_p is built from sets $\mathcal{L}$ and $\mathcal{H}$ using the boosted feature selection algorithm with early stopping and then tested on set $\mathcal{Q}$ using 1, 3, 5, 7, 9, 11, 13, 15 references from set $\mathcal{R}$. This protocol is repeated 10 times to show replicability.

5 Results and discussion

5.1 Single-scale representations

First, the results from committees built on single-scale feature sets are presented. Table 1 presents the committee size

for replication 1 to 10 **do**
  Initialize $\mathcal{L} = \mathcal{H} = \mathcal{Q} = \mathcal{R} = \{\emptyset\}$
  for $p = 1$ to 60 **do**
    Acquire new handwritten signature representation datasets $\mathcal{H}_p, \mathcal{L}_p, \mathcal{Q}_p, \mathcal{R}_p$ from feature extraction expert
    Add $\mathcal{L} \leftarrow \mathcal{L} \cup \mathcal{L}_p$
    Add $\mathcal{H} \leftarrow \mathcal{H} \cup \mathcal{H}_p$
    Add $\mathcal{Q} \leftarrow \mathcal{Q} \cup \mathcal{Q}_p$
    Add $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_p$
    if $p \in \{1, 5, 10, 15, 20, 25, 30, 40, 50, 60\}$ **then**
      Build committee of stumps $F_p \leftarrow GentleAdaBoost(\mathcal{L}, \mathcal{H})$
      Evaluate F_p on $\mathcal{Q}$ using $1, 3, \ldots, 15$ references from $\mathcal{R}$
    end if
  end for
end for

Fig. 8 Protocol to evaluate the impact of the quantity of signature representations on the boosted feature selection algorithm with early stopping

Table 1 Committee size and ratio of selected features for single-scale ESC representation committees

Rows	Col.					
	50	25	12	6	3	1
20	1,028/0.16	802/0.22	303/0.21	717/0.44	468/0.54	234/0.64
10	623/0.20	784/0.34	607/0.43	471/0.54	259/0.62	186/0.86
5	425/0.26	956/0.53	536/0.57	561/0.74	161/0.74	118/1.00
2	784/0.69	810/0.82	539/0.87	300/0.96	232/1.00	287/1.00
1	484/0.78	493/0.96	344/0.98	271/1.00	110/1.00	232/1.00

Table 2 Error rate for the committee of ESC features at scale 2×3 (%)

Type	Cardinality of the reference set							
	1	3	5	7	9	11	13	15
Genuine S.	21.99 (4.14)	16.75 (3.81)	15.40 (3.05)	14.39 (2.46)	14.29 (2.16)	14.07 (2.61)	13.59 (2.19)	13.26 (2.02)
Random F.	0.77 (0.39)	0.45 (0.19)	0.39 (0.18)	0.37 (0.15)	0.34 (0.16)	0.33 (0.16)	0.32 (0.12)	0.35 (0.11)
Simple F.	1.45 (0.68)	0.81 (0.37)	0.61 (0.31)	0.56 (0.27)	0.48 (0.24)	0.43 (0.23)	0.42 (0.18)	0.41 (0.17)
Simulated F.	20.42 (3.32)	18.75 (2.83)	17.92 (2.40)	18.01 (2.14)	17.54 (1.78)	17.25 (2.11)	17.31 (1.74)	17.29 (1.63)
Overall	11.16 (0.75)	9.19 (0.49)	8.58 (0.44)	8.33 (0.36)	8.16 (0.34)	8.02 (0.30)	7.91 (0.26)	7.83 (0.24)

and ratio of features selected for single scale with the ESC representations. Since no committee has reached $T_{\mathcal{L}}$ iterations, early stopping has occurred for every scale. The lower selection rate obtained at higher resolutions indicates the presence of redundant and irrelevant features. Best overall error rates from ESC representations are obtained at scale 2×3 (see Table 2). Results are presented as the mean error rate over 100 replications, along with one standard deviation (in parenthesis).

Table 3 presents the committee size and ratio of selected feature from single-scale DPDF representations. Again, no committee has reached $T_{\mathcal{L}}$ iterations, and early stopping has occurred for every scale. When compared to ESC representation committees, the DPDF representation committees are usually larger and have a higher ratio of selected features that indicate that DPDF representations generally contain less redundant and irrelevant information. Best overall error rates from DPDF representations are obtained at scale 20×6 (see Table 4). Compared to best ESC representation committee, the DPDF representation committee provides a lower error rate.

5.2 Information fusion at feature level

This subsection presents results from committees built on multi-scale ESC representation, committees built on

Table 3 Number of terms for single-scale DPDF representation committees

Rows	Col.					
	50	25	12	6	3	1
20	909/0.15	930/0.25	796/0.33	638/0.45	778/0.64	589/0.96
10	1,358/0.28	761/0.32	836/0.45	598/0.56	628/0.76	361/1.00
5	1,018/0.38	500/0.41	963/0.62	730/0.77	258/0.78	326/1.00
2	1,288/0.86	939/0.94	983/1.00	1,056/1.00	632/1.00	334/1.00
1	681/0.94	557/1.00	705/1.00	746/1.00	285/1.00	360/1.00

Table 4 Error rate for the committee of DPDF features at scale 20×6 (%)

Type	Cardinality of the reference set							
	1	3	5	7	9	11	13	15
Genuine S.	20.69 (3.56)	16.47 (3.07)	14.65 (2.76)	14.96 (2.64)	15.12 (2.29)	14.93 (2.67)	15.07 (2.79)	15.33 (2.84)
Random F.	0.87 (0.44)	0.69 (0.29)	0.60 (0.25)	0.53 (0.21)	0.49 (0.18)	0.48 (0.17)	0.44 (0.13)	0.43 (0.14)
Simple F.	1.48 (0.56)	0.97 (0.31)	1.00 (0.25)	0.90 (0.21)	0.88 (0.18)	0.88 (0.18)	0.87 (0.19)	0.84 (0.18)
Simulated F.	16.72 (2.62)	14.83 (2.56)	15.33 (2.23)	14.44 (2.12)	13.90 (1.84)	13.94 (2.27)	13.74 (2.47)	13.52 (2.50)
Overall	9.94 (0.59)	8.24 (0.40)	7.90 (0.36)	7.71 (0.30)	7.60 (0.26)	7.56 (0.24)	7.53 (0.22)	7.53 (0.21)

Table 5 Selected features rate for the multi-scale ESC representation committee

Rows	Col.					
	50	25	12	6	3	1
20	0.03	0.03	0.06	0.06	0.11	0.25
10	0.03	0.03	0.06	0.07	0.08	0.25
5	0.06	0.05	0.09	0.12	0.16	0.27
2	0.05	0.11	0.20	0.25	0.34	0.45
1	0.12	0.13	0.20	0.32	0.12	0.50
Overall selected features rate: 0.05						

multi-scale DPDF representation, and committees built on a multi-feature representation (both multi-scale ESC and DPDF representations).

The committee built on multi-scale ESC representation is composed of 1,095 terms. Table 5 details the individual ratio of selected features at each scale. Features from every scale have been selected for a total of 818 features out of 15,457, resulting in an overall selected features rate of approximately 5%. Table 6 presents the mean error rates over 100 replications with one standard deviation for the multi-scale ESC representation committee. Using the multi-scale approach leads to lower overall error rates compared to single-scale ESC representation. This is explained by the greater quantity of features available to build committees.

The committee built on DPDF multi-scale representation is composed of 1,288 terms. Table 7 details the individual ratio of selected features for each scale. Features from every scale have been selected for a total of 888 features out of 14,744, resulting in an overall selected features rate of approximately 6%; a result similar to the one obtained with multi-scale ESC representation.

Table 8 presents the mean error rates (in %) of the 100 replications with one standard deviation for the multi-scale

Table 6 Error rate for the committee of ESC multi-scale representation (%)

Type	Cardinality of the reference set							
	1	3	5	7	9	11	13	15
Genuine S.	19.37 (3.66)	14.41 (2.87)	13.67 (2.68)	13.34 (2.49)	12.81 (2.48)	12.46 (2.20)	12.29 (2.00)	12.19 (2.13)
Random F.	0.14 (0.11)	0.07 (0.10)	0.05 (0.09)	0.05 (0.09)	0.03 (0.07)	0.02 (0.06)	0.01 (0.05)	0.02 (0.05)
Simple F.	0.39 (0.22)	0.24 (0.11)	0.18 (0.06)	0.18 (0.05)	0.18 (0.04)	0.17 (0.03)	0.17 (0.02)	0.17 (0.00)
Simulated F.	16.17 (2.68)	15.41 (2.45)	14.82 (2.48)	14.53 (2.22)	14.82 (2.33)	14.91 (2.12)	14.97 (1.92)	14.93 (1.90)
Overall	9.02 (0.61)	7.53 (0.41)	7.18 (0.33)	7.02 (0.28)	6.96 (0.25)	6.89 (0.19)	6.86 (0.18)	6.83 (0.18)

Table 7 Selected features rate for the multi-scale DPDF representation committee

Rows	Col.					
	50	25	12	6	3	1
20	0.04	0.04	0.06	0.08	0.11	0.41
10	0.03	0.04	0.04	0.06	0.12	0.30
5	0.06	0.07	0.09	0.11	0.22	0.35
2	0.14	0.11	0.19	0.25	0.21	0.38
1	0.18	0.18	0.31	0.54	0.25	0.75
Overall selected features rate: 0.06						

DPDF representation committee. Similarly to the ESC multi-scale representation, the multi-scale approach leads to lower error rates compared to the single-scale DPDF representation. However, the multi-scale DPDF representation committee provides lower error rates than the multi-scale ESC representation committee.

The committee built a multi-feature ESC+DPDF representation (multi-scale representations with ESC and DPDF) is composed of 679 terms. Table 9 details the features selection rate for each scale. Features from every scales have been selected for a total of 555 features out of 30,201 resulting in an overall selected features rate of less than 2%. By providing more diversified information to BFS results in a lighter committee using less features.

Table 10 presents the mean error rates over the 100 replications for the committee built on ESC+DPDF multi-feature representations. The error rates are lower than those obtained from the committees based only on one of the two multi-scale representations. This confirms that ESC+DPDF multi-feature representation are complementary and that together, they provide greater diversity to BFS.

5.3 Information fusion at the confidence score level

This section presents the results obtained from the overproduce and choose approach. A mean of 19.81 committees are selected per replication. Each committee uses a mean of 1736.93 ESC features and 2004.12 DPDF features for a total of 3741.05 features. Table 11 details the individual ratio of selected features at each representation. Of the 30 resolutions, 4 are systematically selected for both types of representation and 11 are systematically discarded. Interestingly, the remaining fifteen (that is, half of the resolutions available) are equally shared between both types of representation, indicating the complementarity of the two feature extraction techniques.

Table 12 presents the mean error rates over 100 replications for the ensemble of committees. The overall error rates are lower than those obtained from the committees based only on one of the two multi-scale representations, but higher than those obtained from the committee based on both multi-scale representations.

5.4 Fast incremental learning

This section presents results from the forward incremental learning scheme. Figure 9a presents the mean error rate as a function of both the number of references per writer and the number of representations presented to the system. The actual number of selected representations is indicated on Fig. 9b using the mean number of features as a function of the number of representations that has been presented to the system.

The mean error rate decreases monotonically according to both the number of representations and the number of references. In both cases, there seems to be a limit to the improvement provided by adding new references and new representations since the improvement lessens as more references or representation are added. However, the figure clearly shows that adding new representations leads to a greater impact on accuracy than by adding new reference signatures.

Figure 10a presents the mean error rate as a function of both the number of references per writer and the number of representations used by the verification system. The mean error rate decreases monotonically for both the number of representations and the number of references. In both cases, there is a limit to improvements provided by adding new references and representations. However, the figure clearly shows that adding new representations has a greater impact on accuracy than adding new references.

Table 8 Error rate for the committee of DPDF multi-scale representation (%)

Type	Cardinality of the reference set							
	1	3	5	7	9	11	13	15
Genuine S.	15.68 (2.77)	11.44 (2.86)	10.61 (2.57)	10.23 (2.59)	10.23 (2.33)	10.21 (2.10)	9.94 (2.06)	9.88 (1.97)
Random F.	0.21 (0.14)	0.15 (0.11)	0.15 (0.11)	0.14 (0.11)	0.12 (0.11)	0.11 (0.10)	0.11 (0.09)	0.12 (0.10)
Simple F.	0.87 (0.40)	0.67 (0.27)	0.67 (0.21)	0.69 (0.15)	0.68 (0.16)	0.69 (0.14)	0.70 (0.14)	0.70 (0.14)
Simulated F.	15.40 (2.33)	14.03 (2.70)	13.68 (2.43)	13.45 (2.36)	13.14 (2.26)	13.00 (1.95)	13.05 (1.99)	13.05 (1.94)
Overall	8.04 (0.49)	6.57 (0.35)	6.28 (0.26)	6.13 (0.23)	6.05 (0.19)	6.00 (0.17)	5.95 (0.17)	5.94 (0.15)

Table 9 Selected features rate for the multi-feature ESC+DPDF representation committee

Rows	Col.					
	50	25	12	6	3	1
20	0.01/0.02	0.01/0.01	0.01/0.01	0.01/0.02	0.02/0.02	0.01/0.02
10	0.01/0.02	0.00/0.01	0.00/0.03	0.02/0.01	0.00/0.05	0.00/0.06
5	0.02/0.03	0.02/0.07	0.02/0.02	0.02/0.02	0.01/0.03	0.00/0.04
2	0.01/0.03	0.06/0.04	0.01/0.04	0.00/0.07	0.00/0.07	0.00/0.00
1	0.02/0.08	0.02/0.04	0.00/0.00	0.00/0.00	0.00/0.00	0.00/0.00
Overall selected features rate: 0.01/0.02						

Table 10 Error rate for the multi-scale ESC+DPDF representation committee (%)

Type	Cardinality of the reference set							
	1	3	5	7	9	11	13	15
Genuine S.	13.53 (3.01)	10.66 (2.41)	9.83 (2.25)	9.75 (2.18)	9.36 (2.12)	9.69 (1.75)	9.80 (2.19)	9.77 (1.93)
Random F.	0.12 (0.10)	0.06 (0.08)	0.04 (0.07)	0.03 (0.06)	0.03 (0.07)	0.02 (0.06)	0.03 (0.06)	0.02 (0.05)
Simple F.	0.43 (0.22)	0.33 (0.09)	0.32 (0.06)	0.33 (0.04)	0.32 (0.04)	0.33 (0.03)	0.32 (0.05)	0.32 (0.05)
Simulated F.	14.95 (2.54)	12.52 (2.14)	11.87 (2.20)	11.36 (2.01)	11.55 (2.07)	11.11 (1.79)	10.77 (2.00)	10.65 (1.92)
Overall	7.26 (0.58)	5.89 (0.38)	5.52 (0.29)	5.37 (0.23)	5.32 (0.25)	5.29 (0.22)	5.23 (0.18)	5.19 (0.18)

Table 11 Selection rate of ESC+DPDF representation committees

Rows	Col.					
	50	25	12	6	3	1
20	0.00/0.00	0.00/0.00	0.00/1.00	0.94/1.00	0.82/0.97	1.00/1.00
10	0.00/0.46	1.00/0.00	0.48/0.01	0.16/0.01	0.00/0.53	0.99/0.98
5	0.29/0.99	1.00/0.03	1.00/0.01	0.84/0.00	0.09/0.00	0.00/0.85
2	0.01/0.60	0.00/0.00	0.00/0.96	0.00/0.01	0.00/0.00	0.00/0.00
1	0.98/0.00	0.70/0.00	0.10/0.00	0.00/0.00	0.00/0.00	0.00/0.00
Overall ESC+DPDF representation selection rate: 0.35/0.31						

Table 12 Error rate for the ensemble of committees (%)

Type	Cardinality of the reference set							
	1	3	5	7	9	11	13	15
Genuine S.	14.36 (3.00)	12.29 (2.55)	11.32 (2.47)	11.42 (2.48)	11.49 (2.01)	11.39 (1.88)	11.38 (1.65)	11.00 (1.98)
Random F.	0.02 (0.05)	0.00 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
Simple F.	0.35 (0.19)	0.23 (0.13)	0.21 (0.11)	0.20 (0.10)	0.18 (0.04)	0.17 (0.03)	0.17 (0.03)	0.19 (0.08)
Simulated F.	14.24 (2.49)	12.11 (2.38)	11.98 (2.23)	11.49 (2.26)	11.16 (1.90)	11.03 (1.80)	10.90 (1.67)	11.15 (1.92)
Overall	7.24 (0.50)	6.16 (0.37)	5.88 (0.30)	5.78 (0.26)	5.71 (0.24)	5.65 (0.20)	5.61 (0.17)	5.59 (0.15)

Figure 10b compares the mean error rates when applying information fusion at the feature level compared to information fusion at the confidence score level. Committees resulting from information fusion at feature level obtain a lower error rate, which is explained by the richer information conveyed by features rather then scores.

5.5 Discussion

Table 13 summarizes the results presented in this paper, bold indicating lowest error rates among the compared SV systems. Results indicate that the proposed multi-feature ESC+DPDF approach provides a high level of performance,

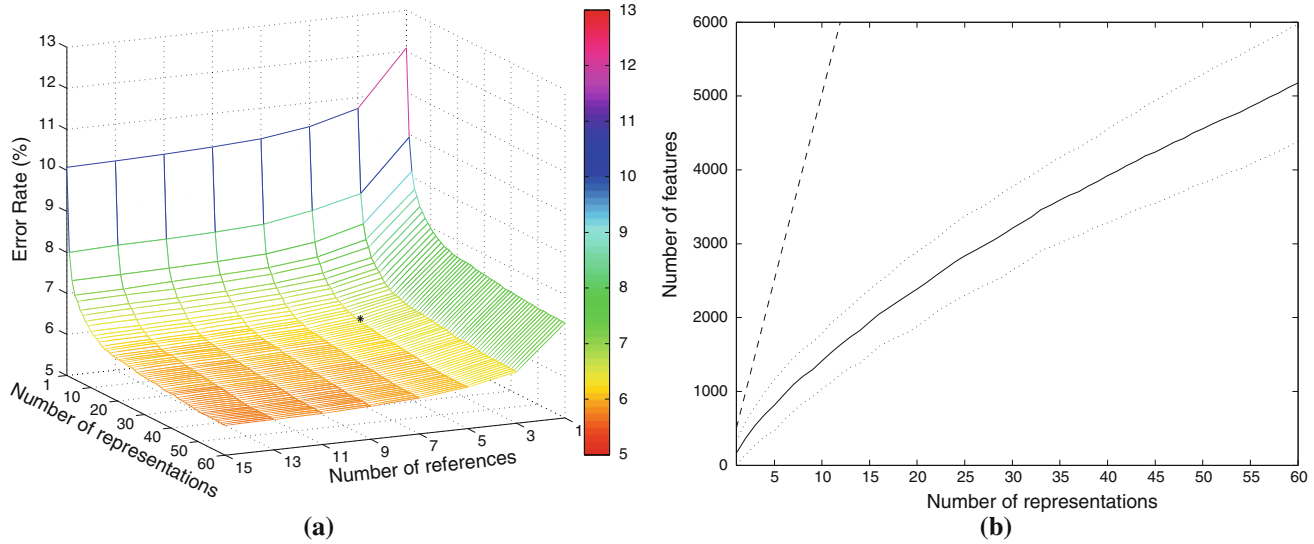

Fig. 9 **a** Mean error rate in function of both the number of references per writer and the number of signature representations *presented* to the system. **b** Mean (*solid line*) and standard deviation (*dotted lines*) of the number of features *used* by the system in function of the number of representations that has been presented. The *dashed line* represents the number of feature that would be used by the system, should it select all representations presented to it. The *dashed line* continues outside the graphic to reach 30,201 features when the 60 representations have been presented

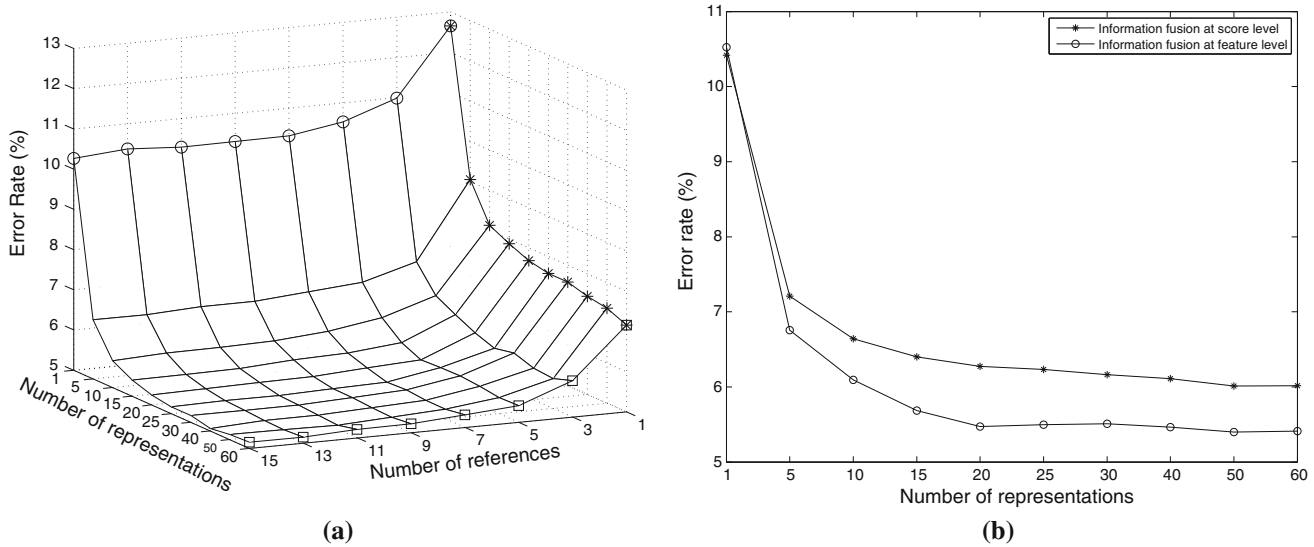

Fig. 10 The impact of the quantity of signature representations on the boosting feature selection algorithm with early stopping. **a** Mean error rate in function of both the number of references per writer and the number of signature representations *presented* to the system. **b** Mean error rate in function of number of signature representations using 5 references

yet is faster than the optimal overproduce and choose approach [13]. This is true despite the fact that the proposed approach uses features extracted with two simple techniques (ESC and DPDF), while the optimal overproduce and choose approach uses more sophisticated graphometric features adapted to signature traces. Compared to [13], the improvements are due to a combination of larger overall feature spaces provided by multi-feature extraction and of boosting feature selection. With the WI feature selection approach proposed in this paper, many samples are generated in a large feature space, and feature learning with BFS allows for personalized feature selection, focusing on the more relevant features. Regarding both best single-scale representations, DPDF are significantly more discriminant than

Table 13 Error rates comparison with other systems (%)

Approach	Number features	Cardinality of reference set							
		1	3	5	7	9	11	13	15
Single scale ESC 2 × 3	29	11.16	9.19	8.58	8.33	8.16	8.02	7.91	7.83
Single scale DPDF 20 × 6	216	9.94	8.24	7.90	7.71	7.60	7.56	7.53	7.53
Multiscale ESC	818	9.02	7.53	7.18	7.02	6.96	6.89	6.86	6.83
Multiscale DPDF	888	8.04	6.57	6.28	6.13	6.05	6.00	5.95	5.94
Multi-feature ESC+DPDF	555	7.26	**5.89**	**5.52**	**5.37**	**5.32**	**5.29**	**5.23**	**5.19**
All representations	10,137	7.54	6.59	6.34	6.27	6.24	6.20	6.20	6.17
Forward incremental selection	5,178	7.31	6.25	6.01	5.87	5.81	5.76	5.74	5.73
Overproduce & choose (this work)	3,741	**7.24**	6.16	5.88	5.78	5.71	5.65	5.61	5.59
Overproduce & choose [13]	2,300	–	7.86	7.32	6.32	7.04	7.19	6.73	6.48
MLP [11]	-	–	–	8.02	–	–	–	–	–

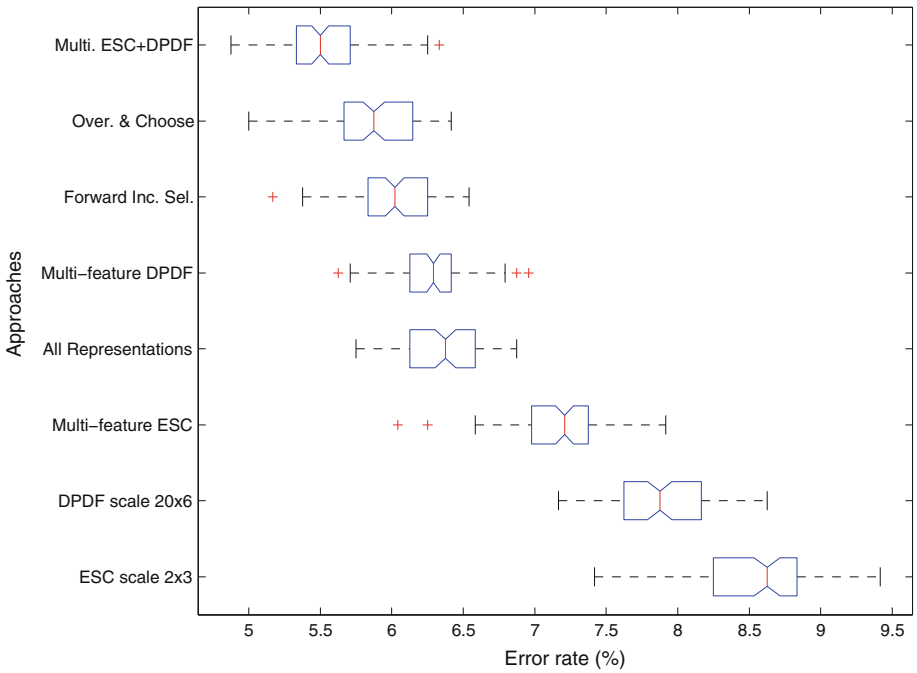

Fig. 11 Comparison of the error rates of the different approaches presented in this work using boxes and whiskers. Error rates are obtained with reference sets of 5 signatures

ESC representations. However, an interesting fact is that ESC performs better at low resolution while DPDF provides better performance at higher resolution. In this respect, the two feature extraction techniques are complementary.

Figure 11 presents a notched box and whisker plot of the error rates of the different approaches explored in this work. The notches represent a robust estimate of the uncertainty about the medians for box-to-box comparison. Boxes whose notches do not overlap indicate that the medians of the two groups differ at the 5% significance level. All error rates significantly differ except for the multi-scale DPDF and the approach combining all 60 independent committees, whose notches overlap.

The proposed solution is highly modular in the sense that each new representation can generate an independent classifier, which in turn can be integrated to the classification module for increased performance. For instance, if all independent committees built from every representations extracted in this work are combined to form ensemble of 60 committees, this "All Representations" (see Table 13) system provides highest level of performance of any system built on a single-scale representation.

The ensemble of committees can also be optimized using a genetic algorithm to further improve performance by filtering out redundant and irrelevant representations. The drawback of this approach is that the optimization process must be

repeated each time a new representation is available. In this case, an incremental learning strategy is more appropriate. Results show that even forward incremental selection, arguably the simplest incremental selection scheme, provides a viable means for filtering representations and increase performance. As shown in Fig. 9a, there is more to be gained from extracting new representations than by sampling new references. Consequently, such a verification system can run with a few signatures of reference if it is composed of adequate representations.

The combination of independent committees into ensemble of committees result in the fusion of information at the confidence score level. When committees are built across multiple representations, the information fusion occurs at the feature level and results small committees with better generalization performance. Both multi-scale ESC and multi-scale DPDF committees outrank their single representation counterparts and use only 5 and 6% of all available features, respectively. This result is even more convincing for a multi-scale committee built across all 60 representations; the committee uses even less features (2%) and provides a lower error rate.

6 Conclusion

This paper presents a practical solution to some of the fundamental problems encountered in the design of off-line signature verification (SV)—the large number of users and features, the limited number of reference signatures, the high intra-personal variability of the signatures, and the lack of forgeries as counterexamples. A new approach for feature selection is proposed for cost-effective design of writer-independent (WI) off-line SV systems. It combines multiple feature extraction, dichotomy transformation, and boosting feature selection (BFS). Computer simulations performed on real-world signature data (comprised of random, simple, and skilled forgeries) indicate that this approach provides enhanced performance when extended shadow code and directional probability density function features are extracted at different scales.

The multi-feature extraction and selection approach proposed in this paper involve dichotomy transformation to mitigate the effects of designing a system with many users and a limited number of reference signatures. The global WI approach allows to explore and select from a large set of features by incorporating prior knowledge of a population of users. Experimental results show the writer-independence by training and testing the system on two disjoint sets of writers and allows for signature verification from only a single reference signature per writer. Results further demonstrate the viability of using random forgeries to train a classifier in the distance space of the dichotomy transformation, thus addressing the lack of skilled forgeries.

The high intra-personal variability of handwritten signatures is dealt with by extracting a large diversified set of features, using one or more preexisting techniques (such as ESC and DPDF) at different scales. Simulation results have shown that these two complementary feature extraction techniques provide a powerful multi-scale and spatio-directional representation of signature images. Given the large number of features, BFS allows to select features *while* learning. Originally proposed for traditional feature vectors, results also indicate the effectiveness of BFS with distance vectors resulting from the dichotomy transformation. Further, this approach result in low cost, efficient classifiers that are suitable for real-time applications.

Another significant advantage of the proposed framework resides in the modularity of its classification architecture. Using the properties inherited from the WI approach, new samples and signature representations can be added to the system during operations. A single classifier may be built off-line, using all available signature representations, thereby fusing information at the feature level. In contrast, one classifier may be built per representation and then grouped into an ensemble of classifiers, thereby fusing information at the confidence score level. The former approach yields the more efficient classifiers, yet requires all representations to be available during the design phase, while the latter allows to design the system using a single representation and then updates it incrementally when new representations become available. There is no need to retrain the WI classifier from the start using all cumulative references signatures. Results indicate that after starting with a single representation and a single reference signature, the accuracy of the system improves the most by adding representations rather than references. The proposed framework is therefore suitable to applications where few reference samples are available.

One issue of off-line signature verification research is the availability of large-scale data sets. Future research will include assessing performance over a wider range of feature extraction techniques, resolutions, and data sets. Other feature extraction techniques and scales would be considered to increase information diversity, and thus system accuracy. An analysis should uncover the impact of different feature types and scales on performance. Future work will also focus on adapting the classification function dynamically to the specific writer or signature for authentication, and thus combining the advantages of both WI and WD approaches. Finally, significant improvement in learning time is expected from distributed computing.

Acknowledgment This research has been supported by the Natural Sciences and Engineering Research Council of Canada and the Fonds de recherche sur la nature et les technologies.

Appendix

A complexity analysis for BFS

Suppose a committee composed of T decision stumps is built from a two-class D-dimensional problem with training and validation datasets of $L = |\mathcal{L}|$ and $H = |\mathcal{H}|$ patterns, respectively. Let t_1 be the time taken to perform an addition, substraction, or comparison, and t_2 the time for a multiplication, division, or exponentiation. For the purpose of this analysis, suppose that later operations are an order of magnitude greater than the former such as $t_2 = 10t_1$.

During testing, a decision stump classifies (5) an input distance vector regardless of the number of features and thus has a constant time complexity of t_1. A committee of stumps repeats this operation T times and then sums the $T - 1$ responses from the stumps. Thus, the total worst-case time required to classify an input vector using the committee $t_{\text{test}}^{\text{wc}}$ (normalized by t_1) is $\frac{t_{\text{test}}^{\text{wc}}}{t_1} = (2T - 1)$. The corresponding growing rate, valid when $T \gg 1$, is $\mathcal{O}(T)$, making for very fast classification during operations. By comparison, RBF-SVM classification time complexity scales linearly with the number of features and support vectors $\mathcal{O}(DN_s)$ [35], where N_s is the number of support vectors. For noisy problems such as WI signature verification, the set N_s increases dramatically and causes a major slowdown for SVM during operation. On the other hand, the boosting approach does not suffer this inconvenience. Moreover, when working with high-dimensional databases such as in this work, $D \gg T$, which makes the boosting approach an attractive alternative to SVMs.

During training, the total worst-case time required to learn with Gentle AdaBoost and early stopping includes the time for quicksort, training decision stumps, and computing the area under the ROC curve. Using quicksort algorithm [36], the worst-case time required to sort the values of one feature is defined as:

$$t_{\text{sort}}^{\text{wc}} = \left(\frac{L^2}{2} + \frac{L}{2}\right) t_1 \tag{6}$$

Once values are sorted, training a decision stump (see Fig. 6) has a worst-case time of

$$\begin{aligned} t_{\text{stump}}^{\text{wc}} &= (13D(L-1) + 5L)\, t_1 + (D(L-1) + L + 2)\, t_2 \\ &= (23DL - 23D + 15L + 20)\, t_1 \end{aligned} \tag{7}$$

The algorithm to compute the area under an ROC curve [32] has a worst-case time of

$$\begin{aligned} t_{\text{AUC}}^{\text{wc}} &= t_{\text{sort}}^{\text{wc}} + (5H + 3)t_1 + (2H + 4)t_2 \\ &= \left(\frac{H^2}{2} + \frac{51H}{2} + 43\right) t_1 \end{aligned} \tag{8}$$

Thus, the total worse-case time required to perform the Gentle AdaBoost with early stopping in a normalized format (see Fig. 5) is expressed by:

$$\begin{aligned} t_{\text{gab}}^{\text{wc}} &= Dt_{\text{sort}} + Tt_{\text{stump}} + Tt_{\text{AUC}} + (TL + 2L + T)t_1 + 4Lt_2 \\ \frac{t_{\text{gab}}^{\text{wc}}}{t_1} &= \frac{DL^2}{2} + \frac{H^2T}{2} + \frac{DL}{2} + \frac{51HT}{2} + 23DLT - 23DT \\ &\quad + 16LT + 42L + 64T \end{aligned} \tag{9}$$

and the corresponding growth rate, when $D, L, T \gg 1$, is:

$$\mathcal{O}(DL^2 + TH^2 + DTL). \tag{10}$$

By comparison, the time complexity of computing a radial basis function kernel matrix for a support vector machine scales to $\mathcal{O}(DL^2)$ [35], not including the time spent on parameters selection.

It is worth noting that quicksort does much better in the average case with $t_{\text{sort}}^{\text{ave}} \approx L \log L$. Thus, the average case time complexity of the Gentle AdaBoost with early stopping is $\mathcal{O}(DLT)$. In computer simulations[1] , the authors have observed the average case analysis to be more representative of the reality than the worst-case analysis. Considering this, when working with large databases such as in this work, $L \gg T$, which makes the boosting approach a method of choice since $\mathcal{O}(DLT)$ grows significantly slower than $\mathcal{O}(DL^2)$ making the Gentle AdaBoost with early stopping a fast learning algorithm that scales linearly. Our simulations with SVMs were executed on the same machine as for BFS, a dual-core Opteron 875 running at 2.2 GHz with 32 GB of memory, using the small and medium single-scale resolution datasets presented in this work at Sect. 4. By comparing execution durations, the SVM approach increased the time complexity by two and three orders of magnitude for learning and testing, respectively, thus confirming the theoretical complexity analysis presented herein. Knowing that the BFS approach took 2 days for learning and classifies around 4,800 samples per second on the largest of the multi-feature datasets, by projection, the SMV approach would require over 6 months for learning and the resulting SVM classifier would only classify up to 5 samples per second.

References

1. Prabhakar, S., Kittler, J., Maltoni, D., O'Gorman, L., Tan, T.: Introduction to the special issue on biometrics: progress and directions. IEEE Tran. Pattern Anal. Mach. Intell. **29**(4), 513–516 (2007)
2. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Trans. Circuits Syst. Video Technol. **14**(1), 4–20 (2004)
3. Fairhurst, M.C.: Signature verification revisited: promoting practical exploitation of biometric technology. Electron. Commun. Eng. J. **9**(6), 273–280 (1997)

[1] The results supporting this claim are not provided in this paper to limit the manuscript's length.

4. Jain, A.K., Ross, A., Pankanti, S.: Biometrics: a tool for information security. IEEE Trans. Inf. Forensics Secur. **1**(2), 125–143 (2006)
5. Batista, L., Rivard, D., Sabourin, R., Granger, E., Maupin, P.: State of the art in off-line signature verification. In: Verma, B., Blumenstein, M. (eds.) Pattern Recognition Technologies and Applications: Recent Advances, pp. 39–62 (2008)
6. Srihari, S.N., Xu, A., Kalera, M.K.: Learning strategies and classification methods for off-line signature verification. In: Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004), pp. 161–166 (2004)
7. Fabregas, J., Faundez-Zanuy, M.: Biometric dispersion matcher. Pattern Recognit. **41**(11), 3412–3426 (2008)
8. Impedovo, D., Pirlo, G.: Automatic signature verification: the state of the art. IEEE Trans. Syst. Man. Cybern. C Appl. Rev. **38**(5), 609–635 (2008)
9. Sabourin, R., Genest, G.: An extended-shadow-code based approach for off-line signature verification. i. evaluation of the bar mask definition. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference B: Computer Vision & Image Processing, vol. 2, pp. 450–453 (1994)
10. Drouhard, J., Sabourin, R., Godbout, M.: A neural network approach to off-line signature verification using directional pdf. Pattern Recognit. **29**(3), 415–424 (1996)
11. Santos, C., Justino, E.J.R., Bortolozzi, F., Sabourin, R.: An off-line signature verification method based on the questioned document expert's approach and a neural network classifier. In: Proceedings of the Ninth International Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004), pp. 498–502 (2004)
12. Tieu, K., Viola, P.: Boosting image retrieval. Int. J. Comput. Vis. **56**(1), 17–36 (2004)
13. Bertolini, D., Oliveira, L.S., Justino, E., Sabourin, R.: Reducing forgeries in writer-independent off-line signature verification through ensemble of classifiers. Pattern Recognit. **43**(1), 387–396 (2010)
14. Cha, S.-H., Srihari, S.N.: Writer identification: statistical analysis and dichotomizer. Advances in Pattern Recognition, Lecture Notes in Computer Science, vol. 1876/2000, pp. 123–132. Springer, Berlin (2000). doi:10.1007/3-540-44522-6_13
15. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. **9**(1), 62–66 (1979)
16. Harrison, W.R.: Suspect Documents, Their Scientific Examination. Nelso-Hall publishers, Chicago (1981)
17. Sabourin, R., Genest, G.: An extended-shadow-code based approach for off-line signature verification. ii. evaluation of several multi-classifier combination strategies. In: Proceedings of the Third International Conference on Document Analysis and Recognition, vol. 1, pp. 197–201 (1995)
18. Kalera, M.K., Srihari, S., Xu, A.: Offline signature verification and identification using distance statistics. Int. J. Pattern Recogn. Artif. Intell. **18**(7), 1339–1360 (2004)
19. Schapire, R.E.: The Boosting Approach to Machine Learning: An Overview, chap. 8. Springer, New York (2003)
20. Kudo, M.: Comparison of algorithms that select features for pattern classifiers. Pattern Recognit. **33**(1), 25–41 (2000)
21. Redpath, D.B., Lebart, K.: Observations on boosting feature selection. In: 6th International Workshop on Multiple Classifier Systems (MCS 2005), pp. 32–41 (2005)
22. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: Proceedings of Thirteenth International Conference on Machine Learning, Eurpean Coordinating Committee for Artificial Intelligence; Italian Association for Artificial Intelligence. Morgan Kaufmann Publishers, San Francisco (1996)
23. Dietterich, T.G.: Ensemble methods in machine learning. In: First International Workshop on Multiple Classifier Systems, pp. 1–15. Springer, Berlin (2000)
24. Rätsch, G., Onoda, T., Müller, K.R.: Soft margins for adaboost. Machine Learn. **42**(3), 287–320 (2001)
25. Servedio, R.A.: Smooth boosting and learning with malicious noise. J. Mach. Learn. Res. **4**, 633–648 (2003)
26. Domingo, C., Watanabe, O.: Madaboost: a modification of adaboost. In: Research Reports on Mathematical and Computing Sciences Series C (Computer Science), no. C-138, pp. 1–26. (1999)
27. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). Ann. Stat. **28**(2), 337–407 (2000)
28. Freund, Y.: An adaptive version of the boost by majority algorithm. Mach. Learn. **43**(3), 293–318 (2001)
29. Rätsch, G., Warmuth, M.K.: Efficient margin maximizing with boosting. J. Mach. Learn. Res. **6**, 2131–2152 (2005)
30. Nakamura, M., Nomiya, H., Uehara, K.: Improvement of boosting algorithm by modifying the weighting rule. Ann. Math. Artif. Intell. **41**(1), 95–109 (2004)
31. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
32. Fawcett, T.: An introduction to roc analysis. Pattern Recognit. Lett. **27**(8), 861–874 (2006)
33. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, New York (2001)
34. Iba, W., Langley, P.: Induction of one-level decision trees. In: Proceedings of the Ninth International Machine Learning Conference. Morgan Kaufmann, San Mateo (1992)
35. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Min. Knowl. Discov. **2**(2), 121–167 (1998)
36. Hoare, C.A.R.: In: Jones, C.B. (eds.) Essays in computing science. Prentice-Hall, Upper Saddle River (1989)