Detector Ensembles for Face Recognition in Video Surveillance

Christophe Pagano, Eric Granger, Robert Sabourin and Dmitry O. Gorodnichy

Abstract— **Biometric systems for recognizing faces in video streams have become relevant in a growing number of private and public sector applications, among them screening for individuals of interest in dense and moving crowds. In practice, the performance of these systems typically declines because they encounter a variety of uncontrolled conditions that change during operations, and they are designed a priori using limited data and knowledge of underlying data distributions. This paper presents multi-classifier system that can achieve a high level of performance in real-world video surveillance applications. This system assigns an ensemble of detectors (2-class classifiers) per individual, where base detectors are co-jointly trained using population-based evolutionary optimization. During enrolment of an individual, an aggregative Dynamic Niching Particle Swarm Optimization (DNPSO)-based training strategy generates a diversified homogenous pool of ARTMAP neural network classifiers using reference data samples. Classifiers associated with local optima of the aggregative DNPSO are directly selected and efficiently combined in the Receiver Operating Characteristic (ROC) space. Performance is assessed in terms of both accuracy and resource requirements on facial regions extracted from video streams of the Face in Action database. A comparison between a standard global and modular classification architectures is provided in this paper. Simulation results indicate that recognizing an individual using the aforementioned ensemble of detectors provides a scalable architecture that maintains a significantly higher level of accuracy and robustness as the number of individuals grows.**

I. Introduction

Biometric recognition of individuals based on their behavioral or physiological traits provides a powerful alternative to traditional authentication schemes (e.g., passwords and identification cards) that are presently applied in many security and surveillance systems. Despite the emergence of many commercial applications, the public sector (government, military, law enforcement, etc.) remains the principal user of biometric technologies for enhanced security. Biometric systems for automated recognition of faces in video streams are relevant in several different real-world applications, ranging from to open-set video surveillance or screening to white-list access-control. Surveillance applications differ from closed-set identification in that the sampling process is performed covertly, and it seeks to determine if a given biometric sample corresponds to an individual of interest enrolled to a restrained black-list.

C. Pagano, E. Granger and R. Sabourin are with the Laboratoire d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, 1100 Notre-Dame Ouest, Montreal, QC, Canada, H3C 1K3, email: cpagano@livia.etsmtl.ca, eric.granger@etsmtl.ca, robert.sabourin@etsmtl.ca.

D. O. Gorodnichy is with the Video Surveillance and Biometrics Section, Science and Engineering Directorate, Canada Border Services Agency, 14 Colonnade Dr., Ottawa, ON, Canada, K2E 6T7, email: dmitry.gorodnichy@cbsa-asfc.gc.ca

Video-surveillance networks are usually comprised of a growing number of IP cameras. In this context, the ability to automatically recognize and track individuals of interest in crowded airports or other public places, and across a network of surveillance cameras, may provide enormous benefits in terms of enhanced screening and situation analysis. However, it is presently difficult to perform fully-automatic recognition of individuals under surveillance using commercial systems For instance, public security organizations are largely limited to human recognition abilities due to the low quality and resolution of facial images typically captured by these cameras. State-of-the-art systems applied to video-based face recognition often perform poorly in practice because they face complex environments that change during operations. In addition, their facial models are designed during a preliminary enrolment process, using limited data and knowledge of individuals. Automated systems should however operate reliably under a wide variety of uncontrolled conditions, and be fast and scalable to several individuals and high definition IP cameras.

Face recognition in video surveillance should be modeled in terms of independent user-specific detection problems. Each one is implemented using one or more one- or two-class pattern classifiers (i.e., detectors), with thresholds applied to their output scores [1], [2]. The advantages of a class-modular architecture include the ease with which biometric models of individual may be added, updated and removed from the systems, and the possibility of specializing feature subsets and decision thresholds to each specific individual. Moreover, using N user-specific detectors has been shown to outperform global N-class classifiers in applications where the training data is sparse w.r.t. the complexity of underlying class distributions, and to the number of features and classes [3].

Some authors have argued that biometric recognition is in essence a multi-classifier problem, and that biometric systems should co-jointly solve several classification tasks in order to achieve state-of-the-art performance [4]. The combination of a diversified pool of classifiers, has been shown improve the overall system accuracy [5]. User-specific ensembles are justified for face recognition in video surveillance by the limited amount of reference samples to design biometric models, and the considerable level of uncertainty of facial models with respect to the complexity of unconstrained video scenes. Creating diversity among base classifiers through *representation space traversal* is considered to be an efficient way to exploit limited data to design accurate heterogeneous classifier ensembles [6]. It allows to explore hyper-parameter space to train classifiers of the same type, on the same data, but with different learning dynamics.

In this paper, a modular multi-classifier system (MCS) is proposed for accurate recognition of individuals in video-to-video surveillance applications. This system assigns an ensemble of binary two-class classifiers (or ensemble of detectors, EoDs) per individual, where base detectors are co-jointly trained using population-based evolutionary optimization. During the enrolment of an individual, an aggregative Dynamic Niching Particle Swarm Optimization (aggregative DNPSO)-based training strategy is employed to generate a diversified pool of ARTMAP neural network classifiers [7] using a mixture of the user's reference samples versus universal and cohort model samples. The Probabilistic Fuzzy ARTMAP classifier [8] is considered for fast and efficient matching of facial regions isolated in video streams against the biometric model of individuals enrolled to the system. As each particle corresponds to an ARTMAP classifier in the hyper-parameter space, the aggregative DNPSO algorithm generates a diversified homogenous pool of classifiers per individual. Classifiers associated with local optima from the aggregative DNPSO optimization are directly selected, according to both accuracy and diversity. Multi-Objective (MO) optimization is proposed to consider the classifier's complexity (or size) as an additional objective, in order to favour lightweight solutions. The iterative Boolean combination technique [9] is employed for decision-level fusion of selected classifiers in the Receiver Operating Characteristic (ROC) space. The accuracy and resource requirements of the proposed approach is assessed using facial regions extracted from real-world video surveillance streams of the Face in Action database [10].

The rest of this paper is structured as follows. The next section briefly reviews the area of face recognition, along with classification architectures for video surveillance applications. In Section 3, the new MCS is proposed video-based face recognition. Section 4 reviews the experimental methodology (data set, protocol and performance measures) used for performance evaluation. Finally, simulation results obtained on real-world video streams are presented and discussed in Section 5.

II. Face Recognition in Video Surveillance

The problem addressed in this paper is the design of accurate and robust systems to recognize faces captured in video feeds across a network of digital surveillance cameras. These systems are considered for video surveillance applications, where individuals of interest must be detected within dense and moving crowds, as found at major events and airports. Face recognition is assumed to be embedded as software executing inside some human-centric decision support system for intelligent video surveillance. During enrollment, an analyst may gradually design and adapt (during operations) facial models of interest over time, based on knowledge and data samples emerging from the particular scene or other sources, e.g., watch lists.

In still-to-video applications, facial models used for classification are designed using one or more Regions of Interest (ROIs) from reference still images, e.g., watch list photographs. Video-to-video applications differ in that facial models are designed using ROIs isolated in reference video streams. In video-to-video applications, facial models used for classification may be initially designed by capturing one or more reference video streams. In this context, an analyst may decide to enroll individuals of interest in some video stream, and then recognize and track their activities over multiple video feeds (from various cameras).

A. A generic system:

Figure 1 presents a generic biometric systems for automated recognition of faces in video. Each camera captures streams of 2D images or frames, where each one provides the system with a particular view of individuals populating the scene. First, the system performs segmentation to isolate ROIs corresponding to the faces in a frame, from which invariant and discriminant features are extracted and selected for classification (matching) and motion tracking functions. For classification, some features are assembled into an input pattern, $\mathbf{a}$, that corresponds to a spatial vector or an ordered sequence of measurements.

During enrolment, one or more reference patterns $\mathbf{a}$ are captured for an individual, and employed to design his user-specific facial model that is stored in a biometric database. Recognition is typically implemented using a template matcher or using a neural or statistical classifier trained a priori to map the input pattern space to one of N predefined classes, each one corresponding to an individual enrolled to the system. Although each facial model may consist of a set of one or more templates (reference captures) for template matching, this paper assumes that a model consists of a statistical representation of reference captures for neural or statistical pattern classification.

During operations, input patterns $\mathbf{a}$ are matched against the model of individuals enrolled to the system. The resulting classification score $S_i(\mathbf{a})$ indicates the likelihood that pattern $\mathbf{a}$ corresponds to individual i, for $i = 1, 2, ..., N$, and is compared against decision threshold, γ_i, to provide an application-specific decision. In surveillance applications, the system outputs a list of all possible identities. To reduce ambiguities during the decision process, some features are also assembled into an input pattern $\mathbf{b}$ for tracking of an individual's motion or appearance over successive ROIs.

Systems for face recognition in video encounter several challenges in practice. In particular, the biometric models are often poor representatives of faces to be recognized during operations [11]. They are typically designed during an a priori enrolment phase, using sparse and unbalanced reference samples collected according to an unknown probability distribution of classes over the input feature space, $p(\mathbf{a})$. The underlying data distribution corresponding to individuals enrolled to the system is complex mainly due to inter- and intra-class variability, to changes that occur during operations, to variations in the interaction between sensor and individual, to the large number of input features and individuals, and to limitations of cameras and signal processing techniques used for segmentation, scaling, filtering, feature extraction and

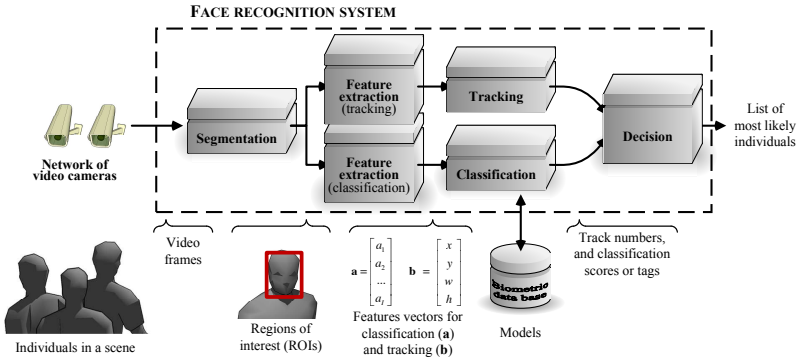


Fig. 1. A generic system for video-based face recognition.

selection, and classification [12]. The performance of face recognition systems may decline considerably because state-of-the-art neural and statistical classifiers depend heavily on the availability of representative reference data of users and possibly a universal model. In addition, $p(\mathbf{a})$ may change gradually or abruptly in the classification environment. All these factors contribute to a growing divergence between the facial model of an individual and its underlying class distribution.

B. State-of-the-art in video surveillance:

A common approach to recognizing faces in video consists in only exploiting spatial information, and applying extensions of image-based (still-to-still) techniques on high quality ROIs isolated during segmentation. Several powerful techniques have been proposed to recognize frontal views of faces in 2D still images under controlled operational conditions. The predominant techniques are holistic or appearance-based methods like Eigenfaces, and local or feature-based methods like Elastic Bunch Graph Matching [13].

However, face recognition in video surveillance remains a difficult task since the faces captured in video frames are typically lower quality and generally smaller than still images. Furthermore, faces acquired from uncooperative individuals in unconstrained scenes may vary considerably due to limited control over operational conditions (e.g., illumination, pose, facial expression, orientation and occlusion), and due to changes in an individual's physiology (e.g., aging) [14], [15]. Limited reference samples has lead to much research on semi-supervised and supervised incremental learning in adaptive biometric systems [16]. Matching ROIs to facial models for a large number of frames from network of cameras also increases the computational burden. Despite these challenges of video-based face recognition, it is possible to exploit spatio-temporal information extracted from video sequences to improve performance[1](see Figure 1). Weak evidence in individual frames can be integrated over long streams, potentially leading to more accurate recognition. For example, track-and-classify approaches combine information from the motion and appearance faces in a scene to reduce ambiguity (e.g., partial occlusion) [14], [15], [17].

Beyond adaptation and spatio-temporal approaches, accurate and robust classification architectures have also been proposed for accurate open-set face recognition. In video surveillance, open-set or open-world face recognition operates under the assumption that most faces encountered during operations do not correspond to an individual of interest – the system only seeks to detect the presence of a restrained group of individuals [18].

The Transduction Confidence Machine-*k* Nearest Neighbour (kNN) is proposed for open-set face recognition using a multi-class architecture and a rejection option for individuals not enrolled to the system [18]. Kamgar-Parsi et al. [19] proposed an approach for watch list applications where the decision regions of target individual are identified in the face space using projection techniques. Both a kNN approach and a set of Gaussian mixtures are used by Stallkamp et al., [20] for real-time video-based face identification. Finally, in [2] open-set face recognition is modeled in terms of user-specific detectors, each one implemented using an Support Vector Machine with thresholds applied to output scores.

III. An Ensemble-Based Classification System

Figure 2 presents a modular multi-classifier system (MCS) proposed for face recognition in video surveillance application. It is composed of a long-term memory (LTM), an ensemble of binary 2-class classifiers or detectors (EoDs) per individual, and a dynamic multi-objective optimization module. Once an input $\mathbf{a}$ is provided by applying one or more state-of-the-art feature extraction techniques, this MCS performs feature selection, classification and decision functions depicted in Figure 1.

During an enrolment phases, when reference data is acquired for an individual, an aggregative DNPSO-based training strategy generates an accurate and diversified pool of base detectors, and selects the subset of detectors for EoD fusion that corresponds to the local optimization optima. Detectors

[1]Figure 1 is an example of a system that combines spatial and temporal computations into separate processing streams that cooperate for enhanced detection of individuals of interest. The general track-and-classify strategy has been shown to provide a high level of performance in video-based face recognition [14], attempting to group successive ROIs into tracks based on input, and fusing the responses via evidence accumulation. This component of Figure 2 is outside the scope of this paper.

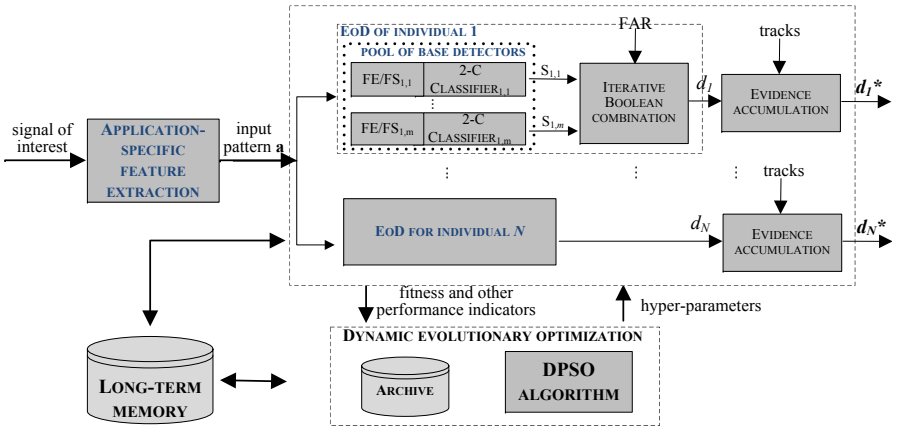


Fig. 2. Architecture of multi-classifier system for face recognition in video surveillance.

may be trained to process different input feature subsets selected from input **a**. The iterative Boolean combination (IBC) technique performs decision-level fusion of selected detectors in the ROC space.

This paper focuses on the validation of the classification module of the system in Figure 2, which uses an EoD per individual. This architecture is compared to standard global and class-modular ones, which are presented in the following sub-section. The rest of this section provides additional details and justifications of the proposed system's EoD training strategy and fusion process.

A. *Classification architectures:*

The Probabilistic Fuzzy ARTMAP (PFAM) [8] is a versatile classifier that may provide a high level of accuracy with moderate time and memory complexity. It is promising for biometric matching (of input feature patterns against facial models) due to its ability to perform fast, stable, online, unsupervised or supervised, and incremental learning from limited amount of training data.

Three specialized architectures are considered and compared for classification of ROIs, and provide a decision based on their output scores. First, the *global or monolithic architecture* shown in Figure 3(a) is composed of a single N-class PFAM classifier, trained to detect the presence of N individuals of interest. Training is performed using a balanced proportion of genuine references samples from each individual enrolled to the system. Training this classifier with an additional class (or alternately setting a reject option) associated with samples from an Universal Model (UM) allows to flag ROIs that do not resemble any of the individuals enrolled to the system. During operations, the PFAM network generates N scores $S_i(a)(i = 1,2,...,N)$, per input feature vector **a**, where each one is compared to a user-specific threshold g_i, in order to produce the final decision d_i, assessing the presence in the scene of individual i. Although the training process is common for all individuals, this architecture provides decision boundaries that grow in complexity with the number of individuals, and the classifier must be retrained or trained incrementally to enrol new individuals. In addition, training with a representative UM class in such architecture may not be feasible, as it should cover an unlimited amount of unknown individuals and variations [2].

In video surveillance applications, the classification task can be be modeled in terms of independent user-specific detection problems, each one implemented using one or more 1- or 2-class pattern classifiers [1], [2]. In the 1-class case, training is performed on representative genuine samples, and in the 2-class case, it is performed using genuine and imposter samples. Accordingly, the *class-modular architecture* shown in Figure 3(b) is composed of a 2-class PFAM classifier per individual, where each classifier is trained using a balanced proportion of genuine references samples from that individual, against a random selection of representative imposter data from UM and Cohort Models (CM, other individuals of interest). The output decision d_i is based on the positive or genuine classification score $S_i(a)$ for classifier i. As mentioned in [1], such architectures have the advantage of facilitating the addition or removal of individuals without retraining the entire system. In addition, the reduction of a global recognition problem into several problems with simpler decision boundaries tends to improve the system's classification performance [2]. Simpler user-specific 2-class classifiers balances system complexity and allows to select specialized feature subsets.

Finally, the *class-modular architecture with EoDs* shown in Figure 3(c), and chosen for the general system in Figure 2, is composed of an ensemble of 2-class PFAM classifiers per individual, where each classifier is trained using a balanced proportion of genuine references samples from the individual, against a random selection of data from the UM and CM. During operations, fusion of classifier decisions is performed in the ROC space, using the Incremental Boolean Combina-

tion (IBC) of ROC curves [9]. This architecture combines a diversified pool of classifiers to address poorly-defined recognition problems that occur when limited and variable reference samples are employed for systems design.

B. Aggregative DNPSO training strategy – generation and selection:

Theoretical and empirical evidence has shown that by generating a diversified pool of classifiers, each one contributes different decision boundaries and commits different errors, and their strategic combination can improve the overall system accuracy [5]. The use of an ensemble is justified by the limited reference data, and the considerable level of uncertainty of facial models with respect to the complexity of unconstrained video scenes. Although not addressed in this paper, EoDs are also well adapted to handle various types of changes to $p(\mathbf{a})$ that occur in real-world environments [21]. Ensemble diversity may be achieved by generating different training sets for the classifiers through, for examples, bootstrapping, boosting and random subspaces. In this paper, heterogeneous ensembles of structurally diverse classifiers are produced using the a population-based evolutionary optimization technique.

The dynamic evolutionary optimization module shown in Figure 2 is implemented with a DPSO-based training strategy proposed in [6], where each particle of a swarm is defined by a base detector in the hyper-parameter (optimization) space. Originally developed for static mono-objective optimization, PSO has been adapted for dynamic and multi-objective optimization. In previous research, the authors have shown empirically that an increase of diversity among particles in the hyper-parameter space is correlated with an increase in diversity among detectors in the input space. Selecting the subset of detectors for EoD fusion can therefore be reduced from a costly search among all base detectors in the input space, to a direct selection of sub-swarm representatives from the archive (i.e., the local optima from optimization).

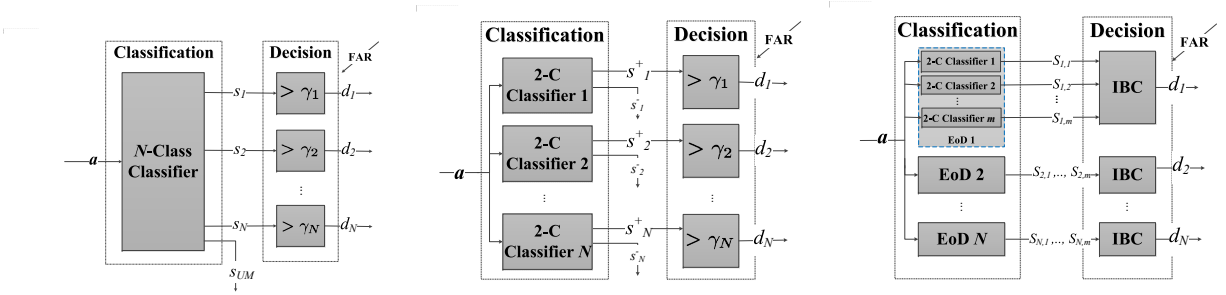
During an enrolment process, an aggregative DNPSO-based training strategy is used to generate an accurate, lightweight and diversified pool of base 2-class PFAM classifiers for the corresponding individual in response to reference genuine and imposter data acquired from the environment. This strategy allows for co-joint optimization of parameters and architectures of identical base detectors, using the same sparse data, but with different learning dynamics (set using hyper-parameters), area under the ROC curve (AUC) is maximized. A heterogeneous EoD is then formed by selecting and then fusing a subset of detectors corresponding to non-dominated solutions stored inside the archive.

C. Iterative Boolean combination – fusion:

Since each user-specific detector is implemented using a pool 2-class classifiers, performance may be characterized in the ROC space [22] (see description in Section IV). Each PFAM detector assigns scores to the input samples, which can be converted to a crisp detector by thresholding the scores. A ROC curve is obtained by varying the threshold that discriminates between genuine and impostor classification scores. These scores are converted into a compact set of operational points, which indirectly convey information about the score distributions.

Given two or more crisp or soft detectors (PFAM networks), their decisions may be combined according to selected thresholds and Boolean functions. While common decision-level techniques combine responses directly on a fixed threshold, ROC-based combination of soft detectors involve sweeping the entire range of true positive and false positive rates, allowing for a flexible selection of the desired operating performance. A change of conditions, such prior class probabilities or costs of errors, lead to a shift in the optimal operating point on the composite convex hull, but the overall convex hull does not change. Fusion in the ROC space is not influenced by asymmetries in genuine and impostor distributions, and normalization of scores is not required because ROC curves are invariant to monotonic transformation of thresholds.

Recently, the authors have proposed an iterative BC (IBC) technique [9] for efficient fusion of multiple ROC curves using all Boolean functions, without any prior assumptions. It has a time complexity that is linear w.r.t. the number of detectors, and has been shown to outperform reference



(a) monolithic architecture with UM (b) modular architecture (c) modular architecture with EoDs

Fig. 3. Architectures for classification and decision process of the proposed system.

techniques. In the framework of Figure 2, IBC was employed for decision-level fusion of selected detectors, to design heterogeneous EoDs. It cumulatively combines base detectors from a DPSO archive that provides the greatest AUC accuracy over some validation data.

IV. Experimental Methodology

A. Video database

The Carnegie Mellon University Faces In Action (FIA) face database [10] has been used to evaluate the detection performance of the classification architectures on real video-based data. It is composed by 20-second videos of face data from 221 participants, mimicking a passport checking scenario, in both indoor and outdoor scenario. Videos have been captured from three different angles, with two different focal length for each, at a resolution of 640x480 pixels at 30 images per second. Data have been captured in three sessions, with at least a month between each one. On the first session, 221 participants were present, 180 of whom returned for the second session, and 153 for the third.

The simulations have been performed using pictures captured by the two frontal cameras, in the first two indoor sessions – the first one for training and the one for testing. Among the all individual, 45 have been selected to populate the watch list because they are present in every session, with at least 150 ROIs for training and 300 ROIs for testing. This guarantees up to 15 samples per fold when performing 10 folds cross-validation, and thus the possibility to experiment with different amounts of training samples. Among the remaining 176 classes, 88 have been randomly chosen to build the UM for training, which guarantees the presence of another unknown 88 individuals for testing.

Segmentation is performed using the OpenCV implementation of the Viola-Jones algorithm, and ROIs are normalized to 70x70 pixels. Features are extracted with the Multi-Bloc Local Binary Pattern (LBP) [23] algorithm features for block sizes of 3x3, 5x5 and 9x9 pixels, concatenated with the grayscale pixel intensity values and reduced to 32 features using Principal Component Analysis.

B. Simulation Protocol

The PSO-based training strategy relies on the Dynamic Niching PSO (DNPSO) algorithm, with the following parameters[6]: 60 particles per swarm; max of 30 iterations; neighborhoods of 6 particles; max of 40 sub-swarms; max of 5 particles per sub-swam; early stopping if the best solution ever encountered remains fixed for 5 iterations. The bounds for PFAM parameter are: $0 \leq \bar{\rho} \leq 0.9$; $0 \leq \alpha \leq 1$; $0 \leq \beta \leq 1$; $0 \leq \varepsilon \leq 1$; $0.0001 \leq r \leq 200$. In order to control the classifier complexity, multi-objective (MO) optimization is performed through the aggregation of a performance and network compression measure, as proposed in the Aggregated PSO [24].

The simulation scenario follows the 2x5 folds cross-validation process for 10 replications, with the randomization of the samples order at the 5th replication. The first step of a simulation scenario is the generation of the *dbLearn* dataset, which is used to perform training and optimization of the PFAM networks. *dbLearn* remains unchanged for the two sets of five replications, in order to combine the results on a coherent basis, especially when random samples from UM are picked. With the global architecture, *dbLearn* is composed by reference samples of the N classes of the watchlist for the training session. *nSamples* are then computed as the average amount of samples for each individual, and *nSamples* samples are selected at random from UM. With the two modular architectures, each individual i under surveillance has a dedicated PFAM network or EoD as well as a specific learning dataset $dbLearn_i$. It is composed of reference samples of class i, as well as the same amount of samples randomly selected from the UM dataset, and the same amount of samples randomly selected from the other classes in the watchlist (Cohort Model).

Dataset *dbLearn* is divided into the following subsets, based on the 2x5 cross-validation methodology:

- *dbTrain*: the training dataset used to estimate the parameters of PFAM networks.
- *dbVal1*: the first validation dataset, used to validate the number of training epochs.
- *dbVal2*: the second validation dataset, to compute classifier fitness during optimization.
- *dbVal3* (only with EoDs): the third validation dataset, used to determine the IBC parameters for fusion of ensembles.

With modular architecture (with or without EoD), this separation is made for every class i, thus generating the dedicated subsets $dbTrain_i$, $dbVal1_i$, $dbVal2_i$ and $dbVal3_i$.

DNPSO is then used to co-jointly optimize the hyper-parameters for a PFAM network. The fitness computation follows the same 2 step process for all architectures:

1) Presentation of the training dataset *dbTrain* to the PFAM network, and evaluation of its performance with *dbVal1*, as well as the number of categories in the PFAM network N_a. This step is repeated for several epochs (presentations of the training dataset to the classifier) until the performance starts to converge or decrease (the stopping criterion is no increase for two consecutive epochs), to avoid over-training.
2) Evaluation of the fitness function with *dbVal2*.

For the global architecture, when the stopping criterion is met, the best particle *bestPFAM* found during the optimization is selected as the optimal solution to be used for final testing. For the modular architecture, the same process is followed for each class i, using the datasets $dbTrain_i$, $dbVal1_i$ and $dbVal2_i$. A best solution $bestPFAM_i$ is then chosen for each class. Finally, for the modular architecture with EoDs, the best solution $bestPFAM_i$ is obtained for each class in the same way as with the modular architecture, with the addition of the m local bests generated by the optimization process. These classifiers are combined through IBC, using $dbVal3_i$ to determine the fusion parameters.

Finally, selected classifiers or ensembles are evaluated using the testing dataset. When the 10 replications are

completed, the mean and the standard deviation of the results are computed using a Student distribution and a confidence interval of 10%.

C. Performance measures

Given the responses of a detector for a set of test samples, the true positive rate (tpr) is the proportion of positives correctly classified over the total number of positive samples. The false positive rate (fpr) is the proportion of negatives incorrectly classified (as positives) over the total number of negative samples. A ROC curve is a parametric curve in which the tpr is plotted against the fpr. In practice, an empirical ROC curve is obtained by connecting the observed (tpr, fpr) pairs of a soft detector at each threshold.

The area under the ROC curve (AUC) or the partial AUC has been largely suggested as a robust scalar summary of 1- or 2-class classifier performance. The AUC assesses ranking in terms of class separation – the fraction of positive–negative pairs that are ranked correctly. For instance, with an $AUC = 1$, all positives are ranked higher than negatives indicating a perfect discrimination between classes. A random classifier has an $AUC = 0.5$, and both classes are ranked at random. Accuracy of a classifier is assessed in terms of pAUC for a false alarm rate of 0% to 10%. The complexity of PFAM networks is measured as compression – the ratio of the number of samples in the training dataset, n, over the number of categories in the PFAM network N_a.

Based on the same principle used with the aggregative PSO [24], the fitness function for the aggregative DNPSO-based evolution is a weighted sum of the pAUC accuracy and compression, n/N_a:

$$pAUC * 100 + \frac{N_a}{n} \tag{1}$$

This choice of weights allows to favor compact PFAM networks in case of near equality of pAUC accuracy.

V. Simulation Results and Discussion

This paper focuses on comparing the performance of the global and modular classification architectures presented in Figure 3. Table I presents the average pAUC accuracy achieved for each individual randomly selected (among the 45 available candidates) to form a watch list of $N = 10$ individuals. Results were produced for a reduced training subset generated from ROIs of the first session. For the global architecture, an average of 18.5 genuine ROIs per person for training and 64.6 ROIs for validation, and the same amount of UM samples. For both modular architectures, the genuine ROIs have been completed by the same amount of UM and CM samples, leading to an average of 370 imposter ROIs for training, and 1232 for validation. Note that the greater amount of UM samples with the modular architectures is a result of the separate training dataset for each individual.

Results in Table I indicate that the modular architecture yields a significant increase in detection accuracy w.r.t. the global architecture, and is followed most closely by the modular architecture with EoDs. For example, the detection performance for the individual 2 highlights the advantage of using EoDs when facing a complex problem, were limited data is available to design facial models. In this case, the individual wears a hat in the testing dataset, leading to increased intra-class variation. Individuals 58 and 151 are also interesting cases, as the use of modular architectures significantly reduces the variability of detection accuracy, and thus provides a more robust detection performance.

Table II shows the impact on performance of doubling the number of individuals in the cohort. (Considering the space constraints of this documents, only average overall performances are presented in this table.) A significant decline in the pAUC accuracy is observed for global architecture when the number of individual of interest increases from 10 to 20. This decline is not observed with either one of the modular architectures. While the modular architecture with EoDs provides the better overall accuracy, the differences between modular architectures remain within confidence interval limits. This may be explained by the fact that decision boundaries in a global architecture tend to become more complex with the number of classes. As with 10 individuals, this decline is even more visible with the complex case of the individual 2, where the average pAUC for the global, modular and modular with EoDs are about 0.27, 0.31 and 0.40, respectively. This stability in the performance of the modular architectures is a promising result, considering the complexity of real-life surveillance scenario where the number of individuals of interest may be greater than 20.

For the modular architectures, compression is an average for each user-specific detector. Therefore, the modular architecture provides several simpler 2-class classifiers, but the overall system complexity is greater than with the global one. This observation is even more acute when using the modular with EoDs architecture. Even if the modular architectures outperform the global one in terms of pAUC accuracy, memory consumption and processing may limit practical application.

VI. Conclusions

This paper presents an accurate and robust system for face recognition in video surveillance, based on ARTMAP neural classifiers and DPSO. The modular multi-classifier system proposed in this paper is a viable approach for designing advanced face recognition technologies, and yields better detection performances than other standard global of class-modular architectures when working with real video-based data. Although proposed with video-based face recognition in mind, it could be adapted for a wide range of real-world security and surveillance applications.

In addition, although not addressed in this paper, the proposed MCS is suitable for adaptive biometrics. A human-centric video surveillance system may for instance acquire new video streams from the environment or other sources corresponding to individuals of interest after it has originally been deployed for operations. With supervised incremental learning and semi-supervised learning strategies, facial models that are initially designed during enrolment using labelled

Classification Architecture	Individuals									
	2	23	58	106	147	151	176	188	190	209
Global	0.37± 0.038	0.58± 0.095	0.68± 0.12	0.94± 0.036	0.42± 0.13	0.71± 0.11	0.73± 0.05	0.81± 0.076	0.53± 0.065	0.9 ± 0.068
Modular	0.35± 0.04	0.64± 0.15	0.85± 0.04	0.84± 0.075	0.69± 0.13	0.85± 0.035	0.61± 0.054	0.84± 0.06	0.66± 0.096	0.92± 0.054
Modular w. EoDs	0.45± 0.036	0.72± 0.094	0.89± 0.035	0.9 ± 0.069	0.82± 0.11	0.91± 0.043	0.75± 0.054	0.88± 0.049	0.7 ± 0.062	0.97± 0.024

TABLE I

AVERAGE PAUC ACCURACY FOR 10 INDIVIDUAL IN INTEREST.

Classification Architecture	Performance from 10 $\rightarrow$ 20 individuals	
	pAUC	Compression
Global	$0.67 \pm 0.043 \rightarrow 0.66 \pm 0.038$	$7.7 \pm 2 \rightarrow 5.2 \pm 1.5$
Modular	$0.70 \pm 0.031 \rightarrow 0.74 \pm 0.039$	$13 \pm 1.3 \rightarrow 11 \pm 1.2$
Modular w EoDs	$0.80 \pm 0.029 \rightarrow 0.81 \pm 0.025$	$1.4 \pm 0.2 \rightarrow 1.3 \pm 0.096$

TABLE II

AVERAGE OVERALL PERFORMANCE (PAUC AND COMPRESSION) FOR 10 AND 20 INDIVIDUAL IN INTEREST.

training data may be updated with emerging reference samples (labelled by an operator with expert knowledge of intra-class variations) or highly confident unlabelled data [6], [11], [25], respectively. Future work should then involve evaluation of the proposed MCS' ability to maintain accurate biometric models when new reference data become available, allowing an operator to gradually build and refine facial models over time.

ACKNOWLEDGEMENT

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada, and the Defence Research and Development Canada Centre for Security Science Public Security Technical Program (PTSP 03-0401BIOM).

REFERENCES

[1] D. M. J. Tax and R. P. W. Duin, "Growing a multi-class classifier with a reject option", Pattern Recognition Letters, 29(10):1565?70, 2008.

[2] Ekenel, H. K., et al., 'Open-Set FR-Based Visitor Interface System', LNCS 5815, 43-52, 2009.

[3] Oh, I.-S., and Suen, C.Y., 'A Class-Modular FF NN for Handwriting Recognition,' Pattern Recognition, 35, 229-244, 2002.

[4] Bengio, S., and Marithoz, S., 'Biometric Person Authentication IS A Multiple Classifier Problem,' Int'l Workshop on MCS, Prague, Czech Republic, 2007.

[5] Rokach, L., 'Ensemble-Based Classifiers,' Artificial Intelligence Review, 33:1, 1-39, 2010.

[6] Connolly, J.F., Granger, E. and Sabourin, R., "An Adaptive Classification System for Video-Based Face Recognition", Information Sciences, In Press, march 2010.

[7] Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., and Rosen, D. B., "Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps," *IEEE Trans. on Neural Networks*, **3:5**, 698-713, 1992.

[8] C. P. Lim and R. F. Harrison, "Probabilistic fuzzy artmap: an autonomous neural network architecture for bayesian probability estimation", *In Proceedings of 4th International Conference on Artificial Neural Networks*, 148?53, 1995.

[9] Khreich, W., Granger, E., Miri, A. and Sabourin, R., "Iterative Boolean combination of classifiers in the ROC space: An application to anomaly detection with HMMs", Pattern Recognition, 43, 2732-2752, 2010.

[10] Goh, R., Liu, L., Liu, X. and Chen, T., 'The CMU Face In Action Database,' Proc. of IEEE ICCV Workshop on Analysis and Modeling of Faces and Gestures, Beijing, China, 255-263, 2005.

[11] Rattani, A., Adaptive Biometric System Based on , Dept. of E and E Eng., University of Cagliari, PhD Thesis, 2010.

[12] Pato, J.N., et al., Biometric Recognition: Challenges and Opportunities, Whither Biometrics Committee, National Research Council of the NSA, National Academies Press, 2010.

[13] Zhao, W., et al., 'Face Rec.: A Literature Survey,' ACM Computing Surveys, 35:4, 399-458, 2003.

[14] Matta, F., and Dugelay J.-L., 'Person recognition using facial video information: A state of the art,' J. of Visual Languages and Computing, 20, 180-187, 2009.

[15] Zhou, S. K., et al., Unconstrained Face Recognition, Springer, 2006.

[16] Roli, F., et al., 'Adaptive Biometric Systems that Can Improve with Use,' Advances in Biometrics: Sensors, Springer-Verlag, 3, 447-471, 2008.

[17] Barry, M., and Granger, E., "Face recognition in video using a what-and-where fusion neural network", *In International Joint Conference on Neural Networks*, 2255?60, 2007.

[18] Li, F., et al., 'Open-Set Face Recognition Using Transduction', TPAMI, 27:11,1686-1697, 2005.

[19] Kamgar-Parsi, B., et al., 'Toward Development of FR System for Watchlist Surveillance', TPAMI, 33:10, 1925-1937, 2011.

[20] J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen, " Video-based face recognition on real-world data", *In 2007 11th IEEE ICCV*, 229?36, 2007.

[21] Kuncheva, L.I., Combining Pattern Classifiers: Methods and Algorithms, Wiley, 2004.

[22] Fawcett, T., "An introduction to roc analysis", Pattern Recognition Letters, 27(8):861?874, 2006.

[23] Ahonen, T. and Hadid, A. and Pietikainen, M., "Face description with local binary patterns: application to face recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, n. 42, 2037-2041, 2006.

[24] K. E. Parsopoulos and M. N. Vrahatis." Particle swarm optimization method in multiobjective problems", *In Applied Computing 2002: Proceeedings of the 2002 ACM Symposium on Applied Computing*, 603-607, 2002.

[25] Poh, N., et al., 'Challenges and Research Directions for ,' Int'l Biometric Conf., 753-764, 2009.