Accepted in IEEE Symposium on Computational Intelligence for Security and Defence Applications (CISDA), 2012

A Comparison of Adaptive Matchers for Screening of Faces in Video Surveillance

Miguel De-la-Torre*†, Paulo V. W. Radtke*, Eric Granger*, Robert Sabourin*, Dmitry O. Gorodnichy‡
*École de technologie supérieure, Université du Québec, Montreal, Canada
miguel@livia.etsmtl.ca, eric.granger@etsmtl.ca, radtke@livia.etsmtl.ca, Robert.Sabourin@etsmtl.ca
†Centro Universitario de Los Valles, Universidad de Guadalajara, Ameca, México
‡Science and Engineering Directorate, Canada Border Services Agency, Ottawa, Canada
dmitry.gorodnichy@cbsa-asfc.gc.ca

***Abstract*—Video-based face screening is essentially a detection problem where faces captured in video sequences are matched against the facial models of individuals of interest. This problem is associated with several operational challenges, from lighting and pose changes, to natural aging of target individuals, and to the limited availability of reference samples from changing environments to design facial models. Some matchers proposed in literature may be employed to adapt facial models of individuals enrolled to the system in response to new reference samples. This paper reviews and compares the performance of these matchers, focusing on their ability for adapting to new data. An experimental methodology is proposed to assess their performance for video surveillance applications. This methodology is focused on transactional and subject-based performance, and considers the imbalance of positive and negative samples. Experiments are then performed with the Canegie Mellon University Face in Action video dataset, according to matching accuracy and resource requirements. Results indicate that ensemble-based matchers outperform traditional monolithic approaches, maintaining a higher level of accuracy over time when adapting to new reference samples.**

I. Introduction

Video surveillance networks are comprised of a growing number of digital IP cameras, providing massive quantities of data. It is very difficult for human operators to analyze all captured video sequences, even in moderately cluttered scenes. Video-based systems for face recognition (FR) may be used for the automated screening of faces captured in sequences against a restrained list of target individuals, providing an important function for decision support in enhanced surveillance and security systems. These systems store a facial model for each individual enrolled to the system, and a matcher compares facial regions to these models. The matcher produces a score for each comparison between a facial region and acquired during operations and a facial model, which is used on the decision making process. A biometric model (BM) consists of a set of one or more templates (genuine reference facial samples acquired during enrollment process) or the set of parameters of a neural or statistical classifier trained on reference samples.

Assuming that video streams are captured using one or more cameras, a FR system performs segmentation to locate and isolate facial regions of interest (ROIs) in each frame. Invariant and discriminant features of each ROI are then extracted and assembled into a feature pattern for matching against the facial model of target individuals. Local feature-based approaches apply a transformation on image pixels to extract specific features from facial regions. Commonly used local characteristics are features as eyes, ears, nose, and mouth. On the other hand, holistic approaches consider all pixels of the normalized ROI as features for FR, and data dimension corresponds to the number of pixels of the ROI. To avoid dealing with large feature patterns, holistic approaches (e.g., Eigenfaces or Fisherfaces) commonly use some techniques like Principal Component analysis (PCA) or Linear Discriminant Analysis (LDA) for dimension reduction.

Several challenges are present in FR for video surveillance applications. Imbalanced data is representative of the presence of considerable amount of negative samples. Biometric models for target individuals are also not representative, because they are designed using limited and incomplete data captured from uncontrolled environments. Facial captures are then subject to considerable variations due to limited control over operational conditions when acquiring images from unconstrained scenes (e.g., illumination, pose, facial expression, orientation and occlusion). Moreover, physiology of the individuals may change over time, either gradually (aging) or abruptly (illumination). New informations, such as input features and output classes, may suddenly emerge and previously acquired data may eventually become obsolete in dynamically changing environments.

In this paper, holistic based matchers available in the literature are compared for video-to-video face screening applications that are subject to changing environments. Matchers are distance based template matching, Open Set TCM-kNN (Transduction Confidence Machine- k-NN), Probabilistic Fuzzy ARTMAP (in batch and incremental modes), Learn++ and Ensemble of Detectors (EoD). These matchers are employed to adapt facial models over time, in response to newly acquired reference samples. The experimental methodology adapted for video surveillance applications relies on the Carnegie Mellon Faces in Action (FIA) video data [1] that mimics a passport checkpoint scenario. Performance is compared both in terms of transaction and subject-based evaluation. Finally, the methodology discusses imbalanced class distributions and accounts for the availability of abundant negative samples in the test environment.

The rest of the paper is organized as follows. Section II

DOI: 10.1109/CISDA.2012.6291529

presents a brief overview of FR in video, including specific challenges in video surveillance applications. In Section III different matching techniques for adaptive biometrics are discussed. The experimental methodology for surveillance applications (data set, protocol and metrics) is depicted in Section IV. Finally, simulation results are presented and discussed in Section V.

II. Face Recognition in Video Surveillance

Assume that 2D images are captured in one or more video cameras. FR in video involves several processing steps. First, the segmentation process isolates the ROIs corresponding to face appearing in each frame. Among a wide range of techniques in literature, appearance-based methods for image segmentation like the Viola-Jones algorithm, have been shown to efficiently detect facial ROIs in video streams.

Next, the tracking function follows the movement or expression of faces across video frames, while the classification function seeks to match input feature patterns to the face models of individuals enrolled to the system. Feature extraction module then extracts specific characteristics. Kalman filters or particle filters tracking features are typically the position in frames, speed, acceleration, and track number assigned to each ROI in the scene. On the other hand, classification relies on invariant and discriminant features extracted from ROIs: classification features are often image-based (e.g., LBP) or pattern recognition-based (e.g., PCA).

Input pattern $\mathbf{a}$ is then compared to the facial model of individual i stored within a biometric database, producing a similarity score $S_i(\mathbf{a})$. A decision module then uses an application specific threshold to produce a decision $d_i = 1$ if $S_i(\mathbf{a}) \geq \gamma$, otherwise $d_i = 0$. Biometric matching may be implemented using a statistical or neural network classifier trained on reference data. With neural network classifiers, for instance, the BM of individuals is defined by the neural architecture and the synaptic weights. Finally, over a sequence of video frames, the decision module may combine and accumulate the responses from the tracking and classification modules.

Several powerful techniques have been proposed for FR in static 2D images [2]. A common approach to recognize faces in video is to extend static image based techniques, exploiting only spatial information on face images obtained through segmentation on individual frames. The predominant techniques are appearance-based methods like Eigenfaces, and feature-based methods like Elastic Bunch Graph Matching [2].

FR systems for video may exploit spatio-temporal information on the appearance and motion of faces detected in a scene. The advantages of video FR include an increase in contextual knowledge and data in video [3]. For example, track-and-classify systems may combine spatial information with information on motion and appearance of faces in a scene [4]. Given a video sequence, the ROIs corresponding to an individual may be tracked, and the responses may be accumulated over time for improved performance. Regardless, the performance of these techniques may degrade considerably when applied in real-world video surveillance applications.

The collection and analysis of labeled biometric data from individuals is often difficult given that the presence of the individual is required, and older still images may not accurately represent his current physiology. Classifiers are designed during an a priori enrollment phase using sparse and unbalanced reference samples collected according to an unknown data distribution. BMs are often poor representatives of faces to be recognized during operations [5]. 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 capture conditions, to the large number of input features and individuals, and to limitations of cameras and signal processing techniques [6].

The performance of biometric systems may decline considerably because state-of-the-art neural and statistical classifiers employed for matching depend heavily on the availability of sufficient representative reference data and relevant prior knowledge, and such information is difficult to obtain in real applications. In addition, new information may emerge over time, and underlying data distributions may change gradually or abruptly in the classification environment. Performance may decline over time as BMs deviate from the actual data distribution [5], [6]. However the capacity to adapt in response to new reference data is not addressed in face recognition for video surveillance.

Video surveillance problems are addressed as an open-set or open-world problem, where individuals of interest are greatly outnumbered by other unknown individuals in a scene. During operations, the probability of seeing an individual of interest in scenes may be quite low. Li and Wechsler [7] proposed the Open Set TCM-kNN (Transduction Confidence Machine-k-NN) for surveillance applications, which considers training patterns from different classes to tune a global rejection threshold. Tax and Duin propose a multi-classifier composed of one class classifiers per person, in which posterior probabilities are normalized to apply a common rejection threshold across all people, but adapted to each distribution [8]. Ekenel et al. progressively combine confidence scores of *distance-to-model* and *distance-to-second-closest* schemes to estimate the identity of individuals entering to a door [9]. Kamgar and Parsi propose an approach based on the identification of the decision region(s) in the feature space of individual specific faces, by generating borderline images and projecting inside and outside the decision region. In their approach they use a dedicated classifier for each individual [10].

In other biometric applications like speaker recognition, the use of a “Universal Background Model” is widely used for better discrimination between target voice from all other sounds [11]. It is built by selecting samples of the background sound that characterizes a recording environment, and is used to discriminate between the individual (speaker) of interest and other sounds. In the same manner, the cohort model is a set of samples selected from non-target samples from already known voices to discriminate known individuals from other known speakers. These cohort and universal models constitute an important source of reference information to design matchers.

Data from individuals that are not in the cohort may improve the system's ability to detect individuals of interest, as well as to reject the unknown individuals. In video surveillance ROIs from several individuals may used in the construction of the Universal Model (UM) for system design, which guarantees the representation of unknown individuals during matcher design.

III. Adaptive Biometric Matchers

For accurate and timely screening of faces in video, it is important to efficiently adapt facial models over time in response to new training data from a changing pattern recognition environment. Adaptive biometric systems in literature traditionally incorporate newly-acquired reference samples to update the selection of a user's template from a gallery via clustering and editing techniques. These systems seek to improve representation of intra-class variations with a single template [12]. Others have performed on-line learning of genuine reference samples over time to update each user's single super template [13].

Biometric models maintained with self-adaptive or semi-supervised learning strategies are initially designed during enrollment using labeled training data, and then updated with highly confident unlabeled data obtained during operations. These strategies are, however, vulnerable to outliers, dispersion and overlap in class distributions. Highly confident data should be selected to minimize the probability of introducing imposter data into updated BMs.

In this paper, supervised learning strategies for adaptation in face based video surveillance are considered, and new data samples are assumed to be analyzed and labeled by an operator with expert knowledge of intra-class variations. If labeled data becomes available, for instance, over multiple re-enrollment sessions, or when operational videos from different cameras are analyzed off-line, they can allow an operator to gradually refine facial BMs.

In literature, some promising neural and statistical classifiers and multi-classifier systems have been proposed for supervised incremental learning of new data, and provide the means to maintain an accurate and up-to-date face model of individuals [14]. For example, the ARTMAP and Growing Self-Organizing families of neural network classifiers, have been designed with the inherent ability to perform incremental learning. In addition, some well-known pattern classifiers, such as the Support Vector Machine (SVM), the Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) neural networks have been adapted to perform incremental learning. FR in video surveillance corresponds to a series of detection problems, one per person of interest, and this section reviews classification systems for screening of faces appearing in video feeds.

A. *One-class classifiers*

Classification algorithms that are designed using only samples from the positive class are called *one-class classifiers*. For instance, template matching algorithms employ pixel intensity values of an ROI as templates, and typically match using the Euclidean distance to a single template (representing the whole face) [15]. A single template (reference sample from the individual of interest captured during enrollment) is commonly stored to perform matching. For improved performance, several templates per individual may be added incrementally to a gallery, affecting a trade-off between accuracy requirements and resources. The distance from the input pattern to the closest template is then used to compute a score. When new data becomes available, this approach updates the BMs by storing the new templates in the database.

B. *Two-class classifiers*

Matchers may take advantage of negative samples (templates from other unknown individuals) to refine decision boundaries between positive and negative templates. For instance, training two-class statistical or neural network classifiers on both positive and negative reference samples may allow to design more robust and compact matchers. A well known incremental learning classifier is the ARTMAP family of neural networks. In particular, Fuzzy ARTMAP integrate a Fuzzy ART model to process analog and binary valued inputs to the ARTMAP architecture. The probabilistic variant proposed by Lim and Harrison in [16], combines the Fuzzy ARTMAP learning to encode category prototypes and update centers of mass of estimated class distributions. In this way, the output prediction for an input pattern $\mathbf{a}$ for each category j is represented as a hyper-spherical Gaussian probability density function

$$p(\mathbf{a}|C_j) = g_j(\mathbf{a}) = \frac{1}{(2\pi)^{M/2}\sigma_j^M} e^{\left(-\frac{(\mathbf{a}-\mathbf{w}_j^{a-c})^T(\mathbf{a}-\mathbf{w}_j^{a-c})}{2\sigma_j^2}\right)}, \tag{1}$$

where the variance σ_j is the ratio of the squared minimum Euclidean distance between $\mathbf{w}_j^{a-c}$ and any other center M-dimensional pattern, to the value of an overlap parameter $r > 0$. Then, the probabilistic neural network is used for probability estimation of posterior probabilities using Equation 2.

$$\hat{P}(C_j|\mathbf{a}) = \frac{p(\mathbf{a}|C_j)P(C_j)}{\sum_{j=1}^{c} p(\mathbf{a}|C_j)P(C_j)} \tag{2}$$

where $P(C_j)$ is estimated based on the training set. When new data becomes available, PFAM learn it incrementally by adapting its weights and architecture, as well as the parameters of $p(\mathbf{a}|C_j)$ and $P(C_j)$.

C. *Multi-class classifiers with rejection option*

A multi-class classifier designed to address the open set problem in video face recognition is the Open Set TCM-kNN proposed by Li and Wechsler in [7]. This matcher takes advantage of transductive inference to generate a class prediction based on randomness deficiency. For an input pattern **a**, the outputs (p-values) associated to each class y are estimated as

$$p_y(\mathbf{a}) = \frac{f(\alpha_1) + f(\alpha_2) + ... + f(\alpha_l) + f(\alpha^y_{new})}{(l+1) f(\alpha^y_{new})} \quad (3)$$

where f is a monotonic non-decreasing function with $f(0) = 0$, e.g., $f(\alpha) = \alpha$. The measure of strangeness α for a pattern **a** against the k closest training samples is given by

$$\alpha(\mathbf{a}) = \frac{\sum_{j=1}^{k} d^y_j(\mathbf{a})}{\sum_{j=1}^{k} d^{\neg y}_j(\mathbf{a})} \quad (4)$$

where y is the prediction label from the sample **a**, $\neg y$ represents all labels different from y, $d_j(\mathbf{a})$ is the distance measure between samples j and **a**, and k is the parameter for the number of nearest neighbors. The rejection rule for unknown individuals is based on the peak-side-ratio (PSR) given by

$$PSR = \frac{p_{max} - p_{mean}}{p_{stdev}} \quad (5)$$

where p_{max} is the maximum p-value, and p_{mean} and p_{stdev} are the mean and standard deviation for the distribution of p-values for a sample **a**, without considering p_{max}. The *a priori* rejection threshold is given by $\Theta = PSR_{mean} + 3 \times PSR_{stdev}$, where PSR_{mean} and PSR_{stdev} are characteristic of the PSR distribution for presumed impostors. Samples are rejected if $PSR_{test} < \Theta$, otherwise are recognized as belonging to the class with maximum predicted value.

D. Ensembles of two-class classifiers

A well-known ensemble-based technique for incremental learning is Learn++, proposed by Polikar et al. in [17]. This technique is inspired in the AdaBoost algorithm, and allows for incremental learning by incorporating a new set of classifiers to the ensemble every time new data becomes available (see Fig. 1). The classical MLP was originally used as the base classifier, with its parameters adjusted to preserve resources and not necessarily produce a high accuracy. However different classification algorithms like PFAM or SVM may be used as base classifiers. The generation of the pool of classifiers each time a new dataset D_t becomes available, is performed using a bagging strategy, by training distinct instances of MLPs on bootstrap replicates of the training set. Selection criteria integrate a *fixed size* set of classifiers, in which every classifier produces an average error lower than random selection ($\epsilon < 1/2$), and newly added classifiers do not increase the overall classification error over random selection ($\epsilon_{global} < 1/2$).

Another alternative for adaptive ensembles is the modular architecture proposed in [18]. Fig. 2 presents this adaptive multi-classifier system (MCS) that allows for update facial models in response to new reference samples. It is composed of a long term memory (LTM), an ensemble of binary two-class classifiers or detectors (EoDs) $\mathbf{P}_i$ per individual, and a dynamic optimization module.

The EoD (PFAM) used to update the ensemble of classifiers in each module $\mathbf{P}_i, 1 \leq i \leq k$, works as follows. When a new data block D_t is available, a training subset D^t_t is randomly selected according to a uniform distribution, and the remaining data is stored in the LTM for validation. Data in D_t is used to generate a new pool of PFAMs. Three independent validation sets are maintained within the LTM, D^e_t to stop the training epochs of classifiers, D^f_t for fitness evaluation (PFAM parameter optimization) and D^c_t to estimate the fusion function and thresholds using Iterative Boolean Combination (IBC) [19]. A learning strategy is based in the dnPSO optimization algorithm [20]. It generates a diversified pool of PFAM classifiers, and the global best solution p_t is selected and added to the ensemble $\mathbf{P}_i$. The combination function for $\mathbf{P}_i$ is then updated using IBC and validated on D^c_t.

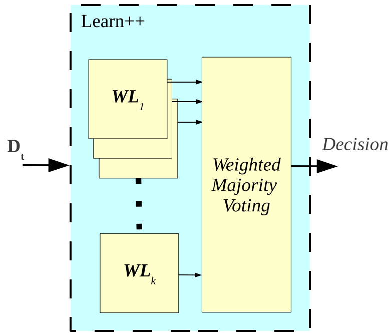

Fig. 1. Architecture of Learn++ system that incorporates a pool of weak classifiers every time a new data block D_t becomes available.

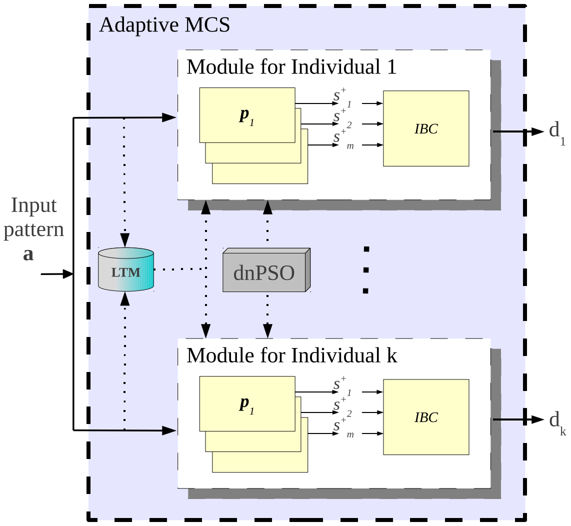

Fig. 2. Adaptive MCS for FR in video surveillance. Dotted arrows indicate pathways for enrollment/update with validation data.

The EoD approach uses IBC to combine previously-trained classifiers with those trained on new data. Given an ensemble of classifiers $\mathbf{P}_i = \{p_1, \ldots, p_t\}$ ranked according to AUC, IBC starts by combining all pairs of operating points (ROC space vertices) for two classifiers with the two highest AUC values. The convex hull of the newly generated operating points are successively combined with operating points of the remaining classifiers, one at a time, until all classifiers have been combined to provide an overall convex hull.

IV. Experimental Methodology for Surveillance Applications

In video screening applications, the fundamental task of FR systems is detecting the presence of an individual from a restrained group or cohort [11], in potentially dense and moving crowds. The set of individuals to be detected corresponds to the cohort of individuals populating a pre-established list of interest. Systems and technologies for FR in video surveillance should be evaluated in terms of their ability to accurately and efficiently detect an individual's face under various uncontrolled conditions.

A. Video dataset

The FIA database [1] has been used to evaluate the performance of the different approaches. The FIA database consists of 20 second videos of face data from 214 individuals mimicking a passport checking scenario. Grayscale ROIs are extracted from video sequences for training and testing. Indoor images are used from both focal lengths (4-mm and 8-mm), taken from three different horizontal angles $(-72.6^o, 0^o, 72.6^o)$. All images are resized to the highest possible resolution of the smallest face obtained after face detection with the well known Viola-Jones algorithm: 70×70 pixels. MSLBP (Multi Scale Local Binary Patterns) [21] is used as a feature extractor with three different block sizes (3×3, 5×5 and 9×9), along with the original pixel intensities. Resulting features are concatenated and vectorized, and PCA allows selecting 32 features with the highest eigenvalues.

Ten individuals of interest are randomly selected to form the cohort (labeled as 2, 3, 72, 82, 136, 140, 179, 188, 190 and 201). The classification module for individual i, e.g. $\mathbf{P}_i$, is trained using a balanced set of samples: 50% of positives samples from individual i, and the remaining 50% of negative samples are drawn uniformly from the cohort model (CM, other 9 individuals in the list of interest) and universal model (UM, samples from 88 random unknown individuals outside the list of interest). Other individuals are considered unknown individuals, and their samples appear only in the test set. Training samples for each module are randomly distributed in three blocks, D_F, D_L and D_R, each preserving the proportion of data from frontal, left and right cameras respectively. Each data block D_t, $t = \{F, L, R\}$, has a fixed size, where $|D_t^t| = 40$ (20 positive and 20 negative) and validation data sets D_t^e, D_t^f and D_t^c all have 10 samples per class. The test set D_{tst} contains a total 138,717 samples, from which $x, 611 \leq x \leq 1636$ samples are from the positive class. The difference in the number of test samples for each individual relies in the face detection process: the face of each individual is detected a different number times by the Viola-Jones algorithm. Negative class samples in D_{tst} are as follows: $12,869 - x$ samples from the cohort model, 26,179 samples from the universal facial model and 99,669 samples from individuals never seen by the system.

B. Experimental protocol

Given the limited positive samples, proof-of-concept simulations follow a 2×5-fold cross-validation process for 10 independent trials. After replication 5, the 5 folds are regenerated after a randomization of the sample order for each individual. The first step of a simulation scenario is the generation of the D_t dataset, which is used for system design. D_t is then divided into the following subsets, based on the 2×5 cross-validation methodology: D_t^t or training dataset used to represent facial models for different individuals, and D_t^{val} or validation dataset used to set system parameters, most notably the decision threshold.

After each design phase, D_{tst} is presented to the system under evaluation, and performance metrics are computed. Unlike D_t^t, the testing data set D_{tst} remains constant over all experiments. Variability in samples from both training and test sets are due to variations in capture conditions, e.g., different lighting and pose, aging, occlusion, camera angle, etc.

C. Transaction-based analysis

A *crisp* detector outputs a class label while a *soft* detector assigns scores or probabilities to the input samples, which can be converted to a crisp detector by setting a decision threshold on the scores. Given the responses of a crisp detector on a validation set, 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 over the total number of negative samples. Accuracy is commonly used to measure the frequency of correct binary decision, however it is prone to biased performance evaluations with imbalanced class distributions.

The receiver operating characteristic (ROC) curve is commonly used for evaluating the performance of detectors at different operating points, without committing to a single decision threshold. A ROC curve is a plot of tpr against fpr. A crisp detector produces a single data point in the ROC plane, while a soft detector produces a ROC curve by varying the decision thresholds. In practice, an ROC plot is a step-like function which approaches a true curve as the number of samples approaches infinity. For equal priors and cost of errors, the optimal decision threshold corresponds to the vertex that is closest to the upper-left corner of the ROC plane.

The area under the ROC curve (AUC) or the partial AUC (over a limited range of fpr values) is largely known as a robust scalar measure of detection accuracy over the entire range of true positive rate (tpr) and false positive rate (fpr). The AUC is equivalent to the probability that the classifier will rank a randomly chosen positive sample higher than a randomly chosen negative sample. 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. When the ROC curves cross, It is possible for a high-AUC classifier to perform worse in a specific region of ROC space than a low-AUC classifier. In such case, restraining the detection rate (tpr) for a fixed fpr gives a performance measure for the specific requirements in the application.

Receiver Operating Characteristic (ROC) and Detection Error Trade-off (DET) curves (which plot fpr versus fnr) are well-accepted graphical representations to express the performance of 1:1 classification. However there are others to evaluate detection quality at the transaction-level, such as ROC isometrics, Precision-Recall curves, Cost Curves, Lift Charts, etc.

D. Precision-recall curves and imbalanced data distributions

Open-set FR in video surveillance translates imbalanced settings, where the prior probability of the positive class (π_p) is significantly less than that of the negative class (π_n). Performance may also be measured as the proportion of the correctly predicted positive samples out of the total number of samples predicted to belong to a given individual. Otherwise, when processing highly imbalanced data, and the minority (positive) samples are of interest, a detector may outperform others by predicting a very large number of samples as minority, resulting in an increased tpr at the expense of an increased fpr. Accuracy is inadequate as a performance measure since it becomes biased towards the majority (negative) class. That is, as the skew ($\lambda = \pi_p/\pi_n$) increases, accuracy tends towards majority class performance, effectively ignoring the recognition capability with respect to the minority class [22].

In this situation, using an estimate of precision (in conjunction with recall) is more appropriate, as it remains sensitive to the performance on each class. In these applications recall only makes sense when combined with precision, as the prior class probabilities are unknown or highly variable. In these situations, end-users relate to precision-recall curves as they indicate how many true positives are likely to be found in a typical search.

The *Precision-Recall Operating Characteristic* (PROC) [22] space allows to represent detector performance graphically with data skew in mind. Performance measures are derived in a similar way to conventional ROC analysis, yet PROC curves rely on an inter-class measure, the precision between the positive and negative decisions, defined as:

$$precision = \frac{TP}{TP + FP}, \tag{6}$$

In general, the tpr (or $recall$) increases with the number of samples of the minority class , while the $precision$ decreases. Thus, an increase of the geometric mean indicates that the achieved increase in $recall$ is beneficial since it is not accompanied by a large decrease of $precision$. Another scalar metric that can be retrieved from the PROC space at a specific operating point is the F_β-measure (Eqn. 7).

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{7}$$

The F_β measure combines the precision and recall values, usually with $\beta = 1$.

In this paper, systems are evaluated by estimating the, $precision$, $recall$ (seen in precision-recall curves), F_1-measure, and compression. When the 10 trials of the experiment are completed, the average values and confidence interval of these estimates are computed.

E. Subject-based analysis

It has been investigated that performance of FR systems may vary drastically from one person to the next, which is known as the 'Doddington zoo' effect [23]. In subject-based analysis, the error rates are assessed with different types of individuals in mind, rather than with the overall number of transactions. An analysis of these individuals and their common properties can expose fundamental weaknesses in a biometric system, and allows to develop more robust systems.

It characterizes positive populations as being composed of *sheeps* and *goats*. According to this characterization, the sheep, for whom FR systems tend to perform well, dominate the population, whereas the goats, though in a minority, tend to determine the performance of the system through their disproportionate contribution to fnr. Goats are characterized by consistently low classification scores against themselves. In negative populations, some individuals – called *wolves* – are exceptionally successful at impersonating many different targets, others, called *lambs*, are easy to impersonate and thus seem unusually susceptible to many different impostors. Lambs, on average, tend to produce high match scores when being matched against by another user. Similarly, wolves receive high scores when matching against others. For both of these user groups, the match score distributions are significantly different than those of the general population.

V. Experimental Results

Concerning transaction-based analysis, the average performance of the six different matchers over 10 trials and 10 individuals is shown in Table I. The first column presents the blocks of data that have been learned by different matchers, going from D_F to D_L and D_R, with samples from by the frontal, left and right cameras respectively. The compression achieved by the classifiers is defined as the ratio between the number of training patterns and the prototypes stored by the algorithm. This measure ranges from 1 for approaches that store all the training set in the biometric database, to a maximum value of $|D_t^t|$. The detection performance is presented using recall, precision and the F_1 measure, ranging from 0 to 1, being 1 the best performance (perfect detection).

When learning the initial data, Eigenfaces (template matching), PFAM (batch and incremental) classifiers and EoD have no significant performance differences. However, as the system incrementally learns new data blocks, the EoD (PFAM) approach performs better than the other matchers. After incrementally learning the data blocks from left and right cameras, it seems to better capture the intra-class variability compared to all other approaches. The Eigenfaces approach initially provides a slightly better precision and recall performance, but the F_1 measure decreases as more data is learned. This

TABLE I
AVERAGE PERFORMANCE OF THE SYSTEM AT THE OPERATING POINT OF 5% fpr.

Block	**Matcher**	**Compression**	**Recall**	**Precision**	**F_1 measure**
D_F	Eigenfaces	1.000(±0.000)	**0.454(±0.034)**	**0.078(±0.007)**	**0.133(±0.011)**
	OS TCM k-NN	1.000(±0.000)	0.109(±0.044)	0.020(±0.009)	0.034(±0.016)
	PFAM$_{batch}$	10.741(±1.397)	0.447(±0.056)	0.076(±0.009)	0.129(±0.016)
	PFAM$_{inc}$	10.741(±1.397)	0.447(±0.056)	0.076(±0.009)	0.129(±0.016)
	Learn++ (PFAM)	**10.903(±1.373)**	0.281(±0.067)	0.049(±0.011)	0.083(±0.019)
	EoD (PFAM)	10.506(±1.298)	0.427(±0.058)	0.073(±0.010)	0.124(±0.017)
$D_F \to D_L$	Eigenfaces	1.000(±0.000)	0.357(±0.031)	0.062(±0.007)	0.105(±0.011)
	OS TCM k-NN	1.000(±0.000)	0.129(±0.050)	0.024(±0.011)	0.039(±0.018)
	PFAM$_{batch}$	4.652(±0.469)	0.394(±0.045)	0.068(±0.009)	0.116(±0.015)
	PFAM$_{inc}$	**7.114(±0.696)**	0.318(±0.063)	0.054(±0.011)	0.092(±0.018)
	Learn++(PFAM)	6.794(±0.725)	0.249(±0.065)	0.044(±0.011)	0.074(±0.019)
	EoD (PFAM)	6.292(±0.675)	**0.459(±0.054)**	**0.078(±0.009)**	**0.133(±0.016)**
$D_F \to D_L \to D_R$	Eigenfaces	1.000(±0.000)	0.357(±0.029)	0.061(±0.005)	0.103(±0.008)
	OS TCM k-NN	1.000(±0.000)	0.209(±0.120)	0.037(±0.022)	0.061(±0.036)
	PFAM$_{batch}$	4.034(±0.326)	0.358(±0.060)	0.060(±0.010)	0.102(±0.016)
	PFAM$_{inc}$	**6.246(±0.720)**	0.219(±0.049)	0.037(±0.008)	0.063(±0.014)
	Learn++ (PFAM)	5.647(±0.550)	0.208(±0.055)	0.038(±0.010)	0.064(±0.017)
	EoD (PFAM)	5.568(±0.570)	**0.477(±0.050)**	**0.081(±0.009)**	**0.138(±0.015)**

problem may be related to the sensitivity of the template matching algorithm to noisy data. The higher level of performance shown for modular systems with one-class and two-class classifiers over the multi-class Open Set TCM-kNN is directly related to the way the last approach establishes a single rejection threshold for all the individuals. On the opposite, the other approaches use positive and negative samples from the cohort and universal models to establish individual specific rejection threshold and classification parameters. Regarding the compression values, incremental learning techniques require less data to represent models when new data is learned (higher compression values). This implies in more efficient usage of computational resources, as smaller models are faster to test unknown samples and require less memory than large models.

A subject based analysis demonstrates that system average performance is not an indication of individual subjects performance for individuals in the cohort of this application (see Table II). Although Open Set TCM-kNN is a multi-class classifier, its performance has been issued per individual. The rejection threshold was applied to the test set before recognition, and a ROC curve for each individual was estimated by ranging the threshold over the outputs produced for that individual. The tpr performance is shown for each matcher as it evolves when a new block of training data is added: $D_F \to D_L \to D_R$. For example, it can be seen that individual 72 is hard to detect by most matchers and tpr values are lower than the average performance. This corresponds to a *goat* like individual in Dodington's zoo terminology. The opposite happens with individual 188, which achieves a higher detection accuracy for EoD (PFAM), making it a *sheep-like* individual for that matcher. It is also interesting to notice that individuals like 3 or 201 for instance, can present a sheep-like behavior for one classifier (EoD (PFAM)), and behave as a goat or a wolf for other (Eigenfaces or Open Set TCM-kNN). These results illustrate the difficulties faced in video-based face screening applications, and may provide important information concerning the fine tuning of the decision module that supports the human operator of such systems.

VI. CONCLUSION

In this paper the performance of adaptive matchers is compared for face recognition in video surveillance. An experimental methodology to asses the performance of the matchers for face screening applications in the context of a changing environment. The methodology employs a real video data set that incorporates changing condition. Comparison is performed in terms of transaction-based analysis, using precision-recall space metrics, and subject-based analysis to better understand system performance for different individuals, based on Doddington's zoo effect.

Experiments carried out with several adaptive matchers indicated that the ensemble of detectors (EoD) approach outperforms others on the FIA data. Transaction-based analysis performed with a 5% false alarm rate show that scalar precision-recall metrics for the EoD are significantly higher than others when the system is adapted to new data. With the compression achieved by incremental PFAM classifiers is higher, its accuracy is low. EoDs provide a good trade off between detection performance and resources. Subject-based analysis demonstrate that system performance per individual differs considerably from the average. This information can be used to fine tune the system's decision module to provide better support to human operators of these security systems.

TABLE II

DETECTION RATE (tpr) OF EACH INDIVIDUAL OF INTEREST FOR A fpr OF 5%, AFTER INCREMENTAL LEARNING OF BLOCKS $D_F \rightarrow D_L \rightarrow D_R$.

	2	**3**	**72**	**82**	**136**
• Eigenfaces	**0.49 ±0.01** → **0.45 ±0.01** → **0.35 ±0.01**	0.45 ±0.01 → 0.19 ±0.01 → 0.23 ±0.01	**0.40 ±0.01** → 0.23 ±0.00 → 0.24 ±0.00	0.32 ±0.00 → 0.36 ±0.01 → 0.51 ±0.02	**0.72 ±0.01** → 0.47 ±0.02 → 0.43 ±0.01
• Open Set TCM kNN	0.12 ±0.00 → 0.14 ±0.01 → 0.23 ±0.03	0.12 ±0.01 → 0.15 ±0.01 → 0.22 ±0.01	0.04 ±0.00 → 0.06 ±0.00 → 0.10 ±0.01	0.08 ±0.00 → 0.12 ±0.01 → 0.30 ±0.04	0.13 ±0.01 → 0.13 ±0.01 → 0.18 ±0.02
• PFAM$_{batch}$	0.40 ±0.03 → 0.41 ±0.02 → 0.34 ±0.03	0.62 ±0.03 → 0.35 ±0.04 → 0.33 ±0.06	0.27 ±0.03 → **0.31 ±0.05** → 0.26 ±0.03	**0.35 ±0.02** → **0.38 ±0.03** → 0.53 ±0.04	0.64 ±0.04 → 0.49 ±0.05 → 0.31 ±0.09
• PFAM$_{inc}$	0.40 ±0.03 → 0.24 ±0.05 → 0.20 ±0.04	0.62 ±0.03 → 0.16 ±0.06 → 0.08 ±0.01	0.27 ±0.03 → 0.20 ±0.05 → 0.18 ±0.05	**0.35 ±0.02** → 0.38 ±0.03 → 0.38 ±0.06	0.64 ±0.04 → 0.48 ±0.07 → 0.12 ±0.03
• Learn++ (PFAM)	0.16 ±0.03 → 0.15 ±0.03 → 0.17 ±0.03	0.34 ±0.08 → 0.34 ±0.08 → 0.17 ±0.06	0.18 ±0.05 → 0.18 ±0.05 → 0.18 ±0.05	0.14 ±0.03 → 0.12 ±0.03 → 0.10 ±0.03	0.42 ±0.08 → 0.42 ±0.08 → 0.37 ±0.07
• EoD (PFAM)	0.31 ±0.03 → 0.33 ±0.02 → **0.35 ±0.01**	**0.63 ±0.03** → **0.65 ±0.01** → **0.65 ±0.01**	0.23 ±0.03 → 0.27 ±0.03 → **0.28 ±0.03**	0.29 ±0.02 → 0.30 ±0.02 → 0.34 ±0.02	0.51 ±0.05 → **0.63 ±0.02** → **0.66 ±0.01**
	140	**179**	**188**	**190**	**201**
• Eigenfaces	0.37 ±0.01 → 0.36 ±0.01 → 0.36 ±0.01	0.37 ±0.00 → 0.43 ±0.01 → 0.34 ±0.00	0.43 ±0.01 → 0.43 ±0.01 → 0.47 ±0.00	**0.51 ±0.01** → 0.35 ±0.01 → 0.35 ±0.01	0.47 ±0.00 → 0.30 ±0.02 → 0.28 ±0.02
• Open Set TCM kNN	0.16 ±0.01 → 0.20 ±0.01 → 0.32 ±0.04	0.12 ±0.01 → 0.13 ±0.01 → 0.19 ±0.03	0.17 ±0.01 → 0.18 ±0.02 → 0.33 ±0.05	0.09 ±0.00 → 0.09 ±0.00 → 0.13 ±0.02	0.07 ±0.00 → 0.08 ±0.01 → 0.09 ±0.01
• PFAM$_{batch}$	0.35 ±0.05 → 0.40 ±0.06 → 0.20 ±0.02	0.41 ±0.01 → **0.46 ±0.02** → **0.46 ±0.03**	0.68 ±0.04 → 0.48 ±0.04 → 0.55 ±0.05	0.31 ±0.05 → **0.36 ±0.05** → 0.40 ±0.06	0.43 ±0.03 → 0.30 ±0.03 → 0.21 ±0.05
• PFAM$_{inc}$	0.35 ±0.05 → 0.27 ±0.05 → 0.21 ±0.04	0.41 ±0.01 → 0.40 ±0.04 → 0.23 ±0.05	0.68 ±0.04 → 0.41 ±0.09 → 0.33 ±0.07	0.31 ±0.05 → 0.26 ±0.06 → 0.23 ±0.04	0.43 ±0.03 → 0.37 ±0.04 → 0.23 ±0.03
• Learn++ (PFAM)	0.17 ±0.05 → 0.15 ±0.05 → 0.14 ±0.05	0.36 ±0.02 → 0.34 ±0.03 → 0.34 ±0.03	0.59 ±0.06 → 0.41 ±0.09 → 0.26 ±0.08	0.22 ±0.05 → 0.21 ±0.05 → 0.21 ±0.05	0.23 ±0.05 → 0.18 ±0.04 → 0.15 ±0.04
• EoD (PFAM)	**0.43 ±0.03** → 0.43 ±0.03 → **0.44 ±0.03**	**0.42 ±0.00** → 0.42 ±0.00 → 0.43 ±0.00	**0.73 ±0.02** → **0.73 ±0.02** → **0.73 ±0.02**	0.26 ±0.05 → **0.36 ±0.04** → **0.42 ±0.03**	**0.47 ±0.03** → **0.47 ±0.03** → **0.50 ±0.02**

ACKNOWLEDGMENT

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). This work was also supported by the Programa de Mejoramiento del Profesorado of the Secretaría de Educación Pública in México (PROMEP/103.5/094294).

REFERENCES

[1] R. Goh, L. Liu, X. Liu, and T. Chen, "The CMU Face In Action (FIA) Database," in *Analysis and Modelling of Faces and Gestures*, 2005, pp. 255–263.

[2] W. Zhao *et al.*, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, December 2003.

[3] F. Matta and J.-L. Dugelay, "Person recognition using facial video information: a state of the art," *Journal of Visual Languages and Computing*, vol. 20, no. 3, pp. 180–7, 2009.

[4] J. Connolly, E. Granger, and R. Sabourin, "An adaptive ensemble of fuzzy artmap neural networks for video-based face classification," *IEEE WCCI*, 2010.

[5] A. Rattani, "Adaptive biometric system based on template update procedures," Ph.D. dissertation, University of Cagliari, 2010.

[6] J. N. Pato and L. I. Millett, *Biometric Recognition: Challenges and Opportunities*, Whither Biometrics Committee, Ed., 2010.

[7] F. Li and H. Wechsler, "Open set face recognition using transduction," *IEEE Trans. on PAMI*, vol. 27, no. 11, pp. 1686 – 97, 2005.

[8] D. Tax and R. Duin, "Growing a multi-class classifier with a reject option," *Pattern Recognition*, vol. 29, no. 10, pp. 1565 – 70, 2008.

[9] H. K. Ekenel, J. Stallkamp, and R. Stiefelhagen, "A video-based door monitoring system using local appearance-based face models," *Comput. Vis. Image Underst.*, vol. 114, no. 5, pp. 596–608, May 2010.

[10] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi, "Toward development of a face recognition system for watchlist surveillance," *IEEE Trans. on PAMI*, vol. 33, no. 10, pp. 1925 – 37, 2011.

[11] A. Brew and P. Cunningham, "Combining cohort and ubm models in open set speaker detection," in *Proceedings on Multimedia Tools and Applications*, vol. 48, no. 1, Van Godewijckstraat 30, Dordrecht, 3311 GZ, Netherlands, 2010, pp. 141 – 159.

[12] F. Roli, L. Didaci, and G. L. Marcialis, *Adaptive biometric systems that can improve with use*. Springer, 2008, pp. 447–471.

[13] A. K. Jain and A. Ross, "Learning user-specific parameters in a multibietric system," *Int. Conf. on Im. Proc.*, September 2002.

[14] J.-F. Connolly, E. Granger, and R. Sabourin, "Supervised incremental learning with the fuzzy artmap neural network," *Artificial Neural Networks in Pattern Recognition. Third IAPR Workshop, ANNPR 2008*, pp. 66–77, 2008.

[15] R. Brunelli and T. Poggio, "Face recognition: features versus templates," *IEEE Trans. on PAMI*, vol. 15, no. 10, pp. 1042 – 52, 1993.

[16] C. P. Lim and R. F. Harrison, "Probabilistic fuzzy artmap: An autonomous neural network architecture for bayesian probability estimation," *Artificial Neural Networks*, pp. 148–153, June 1995.

[17] R. Polikar, L. Udpa, S. S. Udpa, and V. Honavar, "Learn++: An incremental learning algorithm for mlp networks," *IEEE Trans. on SMC*, vol. 31, no. 4, pp. 497–508, 2001.

[18] M. De-la Torre, E. Granger, P. V. W. Radtke, R. Sabourin, and D. O. Gorodnichy, "Incremental update of biometric models in face-based video surveillance," in *Proceedings of IJCNN*, 2012.

[19] W. Khreich, E. Granger, A. Miri, and R. Sabourin, "Iterative boolean combination of classifiers in the roc space: An application to anomaly detection with hmms," *Pat. Rec.*, vol. 43, no. 8, pp. 2732 – 52, 2010.

[20] A. Nickabadi, M. M. Ebadzadeh, and R. Safabakhsh, "Evaluating the performance of dnpso in dynamic environments," in *Proceedings on IEEE ICSMC*, Singapore, Singapore, 2008, pp. 2640 – 2645.

[21] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pat. Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971 – 87, 2002.

[22] T. C. W. Landgrebe *et al.*, "Precision-recall operating characteristic (p-roc) curves in imprecise environments," in *Proceedings of ICPR*, 2006, pp. 123 – 127.

[23] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "Sheep, goats, lambs and wolves: A statistical analysis of speaker performance," in *International conference on spoken language processing*, 1998, pp. 1351–1354.