Pattern Recognition 69 (2017) 61–81

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Dynamic ensembles of exemplar-SVMs for still-to-video face recognition

Saman Bashbaghi[a,*], Eric Granger[a], Robert Sabourin[a], Guillaume-Alexandre Bilodeau[b]

[a] Laboratoire d'imagerie de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montréal, Canada
[b] LITIV, Polytechnique Montréal, Montréal, Canada

ARTICLE INFO

Article history:
Received 28 September 2016
Revised 12 February 2017
Accepted 12 April 2017
Available online 13 April 2017

Keywords:
Video surveillance
Watch-list screening
Face recognition
Single sample per person
Multi-classifier system
Random subspace methods
Domain adaptation
Dynamic classifier selection

ABSTRACT

Face recognition (FR) plays an important role in video surveillance by allowing to accurately recognize individuals of interest over a distributed network of cameras. Systems for still-to-video FR are exposed to challenging operational environments. The appearance of faces changes when captured under unconstrained conditions due to variations in pose, scale, illumination, occlusion, blur, etc. Moreover, the facial models used for matching may not be robust to intra-class variations because they are typically designed a priori with one reference facial still per person. Indeed, faces captured during enrollment (using still cameras) may differ considerably from those captured during operations (using surveillance cameras). In this paper, an efficient multi-classifier system (MCS) is proposed for accurate still-to-video FR based on multiple face representations and domain adaptation (DA). An individual-specific ensemble of exemplar-SVM (e-SVM) classifiers is thereby designed to improve robustness to intra-class variations. During enrollment of a target individual, an ensemble is used to model the single reference still, where multiple face descriptors and random feature subspaces allow to generate a diverse pool of patch-wise classifiers. To adapt these ensembles to the operational domains, e-SVMs are trained using labeled face patches extracted from the reference still versus patches extracted from cohort and other non-target stills mixed with unlabeled patches extracted from the corresponding face trajectories captured with surveillance cameras. During operations, the most competent classifiers per given probe face are dynamically selected and weighted based on the internal criteria determined in the feature space of e-SVMs. This paper also investigates the impact of using different training schemes for DA, as well as, the validation set of non-target faces extracted from stills and video trajectories of unknown individuals in the operational domain. The performance of the proposed system was validated using videos from the COX-S2V and Chokepoint datasets. Results indicate that the proposed system can surpass state-of-the-art accuracy, yet with a significantly lower computational complexity. Indeed, dynamic selection and weighting allow to combine only the most relevant classifiers for each input probe.

1. Introduction

Face analysis and recognition are widely used in applications of law enforcement, forensics, e-learning, biometric authentication, health monitoring and surveillance. In decision support systems for video surveillance, recognizing the faces of target individuals is increasingly employed to enhance security in public places, such as airports, subways, shopping malls, etc [1]. These systems must accurately detect the presence of the individuals of interest across a distributed network of video cameras based on their corresponding facial models. Still-to-video FR systems capture faces appearing in videos, and then match them against facial models generated based on high-quality target face stills [2]. Spatio-temporal recognition and multi-view analysis are typically exploited to enhance performance in such applications [3].

In still-to-video FR, facial models are designed using one or more target facial regions of interest (ROIs) isolated in reference still images for template matching, or for determining a set of classifier parameters [4]. Still-to-video FR systems are typically designed as independent individual-specific detectors, each one implemented with a template matcher, 1-, or 2-class classification system per individual of interest [5]. During enrollment, each detector may be modeled using reference still ROI(s) from target individuals, and possibly still ROIs from the cohort or other non-target persons, as well as, trajectories of video ROIs from unknown (non-

* Corresponding author.
E-mail addresses: bashbaghi@gmail.com, bashbaghi@livia.etsmtl.ca (S. Bashbaghi), eric.granger@etsmtl.ca (E. Granger), robert.sabourin@etsmtl.ca (R. Sabourin), gabilodeau@polymtl.ca (G.-A. Bilodeau).

http://dx.doi.org/10.1016/j.patcog.2017.04.014
0031-3203/

target) individuals. The benefits of designing individual-specific detectors are the feasibility to add, update, and remove detectors from the system, as well as, to select specialized feature subsets, and decision thresholds for each corresponding individual [6].

Watch-list screening is challenging for still-to-video FR systems, because the number of representative reference still ROIs (high-quality mugshots or ID photos) available during enrollment of a target individual is very limited [6]. It is typically too costly or unfeasible to collect and analyze several reference ROIs. In particular, only one or few still ROIs are available for enrollment of an individual, and also a restricted or small number of individuals (cohort) are enrolled to the system. Furthermore, the appearance of ROIs captured from reference stills may differ significantly from ROIs captured from videos, and vary due to capture conditions (e.g. illumination, pose, scale, blur, expression, and occlusion) [7].

Given this single sample per person (SSPP) problem, state-of-the-art systems for still-to-video FR may achieve a low level of performance due to difficulties in designing robust facial models [8]. Different techniques specialized for SSPP problems have been proposed to improve robustness to intra-class variability, such as using multiple face representations, synthesizing virtual faces, and incorporating auxiliary sets to enlarge the training data [6,9,10]. However, multiple representations and synthetic generation techniques alone are only effective to the extent where reference target ROIs captured in the enrollment domain (ED) are representative of an operational domain (OD).

An important issue in still-to-video FR is that probe ROIs are captured over multiple distributed surveillance cameras, where each camera represents a different non-stationary OD. Capture conditions may vary dynamically within an OD according to environmental conditions and individual behaviors. Accordingly, their data distribution differs significantly from ROIs captured with a still camera in the ED, degrading system performance [11]. Designing a robust face model for still-to-video FR is a challenging task due to the difference of faces captured in the ED and OD [8].

Several transfer learning methods have been proposed to design accurate recognition systems that will perform well in the OD using the knowledge taken from the ED [12]. Since the learning tasks and feature spaces between the ED and OD are the same, but their data probability distributions are different, watch-list screening corresponds to domain adaptation (DA) [13]. According to the information transferred between the domains, two unsupervised DA approaches are relevant for still-to-video FR: instance-based and feature representation-based approaches [12]. The former methods attempt to exploit parts of the ED for learning in the OD, while the latter methods exploit OD to find a desired common representation space that reduces the difference between domain spaces and subsequently, the classification error.

Recently, multi-classifier systems (MCSs) have been shown to provide a high level of accuracy and robustness in watch-list screening applications [6,8]. In particular, classifier ensembles can increase the accuracy and robustness of still-to-video FR by integrating diverse pools of classifiers generated using multiple representations of reference facial ROIs. Furthermore, during operations, dynamic classifier selection/weighting methods allow to exploit the most competent classifiers from the pool for a given input probe [14–16]. Dynamic selection (DS) has been shown to be an effective tool to address ill-defined classification problems, where the training data is limited and imbalanced [17,18]. To the best of authors' knowledge, DS has not been exploited in these SSPP problems without using several other target samples to form a validation set.

In this paper, an efficient and robust MCS is proposed for still-to-video FR. Multiple face representations and domain adaptation are exploited to generate an individual-specific ensemble of e-SVMs (Ee-SVM) per target individual using a mixture of facial ROIs captured in the ED (the single labeled high-quality still of target and cohort captured under controlled conditions) and the OD (i.e., an abundance of unlabeled facial trajectories captured by surveillance cameras during a calibration process). Facial models are adapted to the OD by training the Ee-SVMs using a single labeled target still ROI versus cohort still ROIs, along with unlabeled non-target video ROIs. Several training schemes are considered for DA of ensembles according to utilization of labeled ROIs in the ED and unlabeled ROIs in the OD.

During enrollment of a target individual, semi-random feature subspaces corresponding to different face patches and descriptors are employed to generate a diverse pool of classifiers that provides robustness against different perturbations frequently observed in real-world surveillance environments. In this paper, two application scenarios are investigated to design individual-specific ensembles. In the first scenario, a validation set is employed together with a global criterion (measuring the significance of each patch on the overall performance) in order to rank and select patches and subspaces. In contrast, a local distance-based criterion is used in the second scenario to rank subspaces without employing a validation set. In particular, various ranked feature subspaces are sampled from face patches represented using state-of-the-art face descriptors, instead of randomly sampling from the entire ROIs. Pruning of the less accurate classifiers is performed to store a compact pool of classifiers in order to alleviate computational complexity.

During operations, a subset of the most competent classifiers is dynamically selected/weighted and combined into an ensemble for each probe using a novel distance-based criteria. Internal criteria are defined in the e-SVM feature space that rely on the distances between the input probe to the target still and non-target support vectors. In addition, persons appearing in a scene are tracked over multiple frames, where matching scores of each individual are integrated over a facial trajectory (i.e., group of ROIs linked to the high-quality track) for robust spatio-temporal FR. The proposed system is efficient, since the criteria to perform DS and weighting allows to combine a lower restrained number of the most relevant classifiers within the individual-specific ensembles.

Videos from the COX-S2V [19] and Chokepoint [20] datasets are employed to evaluate and compare the performance of the proposed system against state-of-the-art methods. These datasets contains a high-quality reference still from the ED and low-quality videos of individuals captured under uncontrolled conditions in different ODs. Experimental results are obtained at the transaction- and trajectory-levels in the ROC and precision-recall spaces. The results indicate that the proposed system provides state-of-the-art accuracy, yet with a significantly lower computational complexity.

This paper is organized as follows. Section 2 provides some background on still-to-video FR, and its challenges, and on state-of-the-art systems developed to address this SSPP problem. Section 3 presents a review of techniques proposed in the literature for ensemble generation, dynamic selection and weighting of classifiers. Section 4 presents a detailed description of the proposed system. The experimental methodology and simulation results are presented and interpreted in Sections 5 and 6, respectively.

2. Background on still-to-video face recognition

2.1. A generic spatio-temporal system

A spatio-temporal system for still-to-video FR is mainly comprised of the following components, face segmentation (detection), person tracking, face classification and spatio-temporal fusion. In such a system, each surveillance camera captures individuals appearing in its field of view (FoV). Segmentation is performed in each frame to isolate the facial ROIs and then the features are ex-

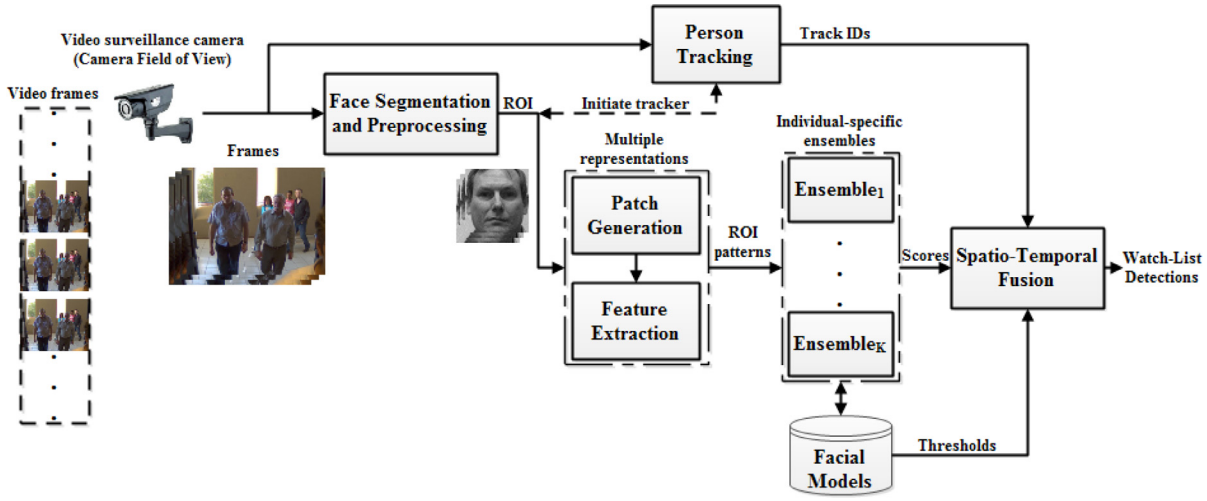


Fig. 1. A multi-classifier system for still-to-video FR using multiple face representations.

tracted and combined into ROI patterns, as well as, initiating the person tracker. Input ROI patterns are then matched against the facial models of each enrolled individuals along a trajectory through a face classification module.

Video-based FR systems can make use of spatial information (e.g. face appearance) along with the location of persons and variations of faces over time to perform a robust spatio-temporal recognition. In watch-list screening applications, spatio-temporal fusion can be performed using a person tracker that regroups ROIs belonging to the same person into a trajectory. The matching scores obtained for each enrolled individual along a trajectory can be accumulated and compared to decision thresholds in order to provide a list of likely target individuals associated with each trajectory. For instance, an adaptive appearance model tracking has been proposed for still-to-video FR [3] to learn track-face-model for each different individual appearing in the scene during operations. Sequential Karhunen–Loeve technique is employed within a particle filter-based tracker for online learning of track-face-models that are matched against the face models of individuals enrolled in the system.

2.2. State-of-the-art still-to-video face recognition

A common mapping space for matching face stills and videos has been learned using partially weighted linear discriminant analysis based on a single high-quality still and a set of low-quality videos of each individual [19]. Since the characteristics of stills and videos are different, it could be an inefficient approach to build a common discriminant space. As a result, a weighted discriminant analysis method has been proposed in [21] to learn a separate mapping for stills and videos by incorporating the intra-class compactness and inter-class separability as the learning objective.

To match image sets in unconstrained environments, a regularized least square regression method has been proposed in [22] based on heuristic assumptions (i.e. still faces and video frames of the same person are identical according to the identity space), as well as, synthesizing virtual face images. In addition, a point-to-set correlation learning approach has been proposed in [23] for either still-to-video or video-to-still FR tasks, where Euclidean points are matched against Riemannian elements in order to learn maximum correlations between the heterogeneous data. Recently, a Grassmann manifold learning method has been proposed in [24] to address the still-to-video FR by generating multiple geodesic flows, to connect the subspaces constructed in between the still images and video clips.

2.3. Techniques for multiple face representation

MCS specialized for spatio-temporal still-to-video FR contains individual-specific ensembles of classifiers generated for multiple face representations (see Fig. 1) [25]. Facial ROIs in each frame are isolated using segmentation and preprocessing module. Meanwhile, the person tracker is initiated to regroup the facial ROIs captured for a same person into a trajectory. Then, multiple face representations are obtained by generating patterns that correspond to different patches and feature extractions to train a diverse pool of base classifiers. An individual-specific ensemble of classifiers is employed for multiple face representations. The fusion module combines the classification scores obtained using comparison of probe ROI pattern against facial models designed for each individual of interest.

Generating multiple face representations from the target reference still can improve robustness in watch-list screening applications. To provide diverse representations for ensembles, extracting different face descriptors and patches can be exploited. To that end, facial ROIs are first divided into several sub-regions (patches) with or without overlapping, then different feature extraction techniques (face descriptors) can be applied on each patch. Patch-based methods allow to recognize faces in partially occluded unconstrained environments through local matching. In addition, they provide robustness to changes in pose and appearance [26]. Hence, patching makes use of local structural information to effectively deal with variations in uncontrolled surveillance conditions. Extracting features from local facial regions for local matching may lead to a robust and accurate FR systems [6].

Exploiting several discriminant face descriptors to generate multiple representations can be effective in still-to-video FR system [6,8]. Each descriptor is specialized to address some nuisance factors (e.g., illumination, pose, blur, etc.) encountered in video surveillance. Hence, the choice of descriptors is based on the complementary information that they provide, where combining classifiers trained with different descriptors into an ensemble can achieve a high-level of robustness. Furthermore, generating synthetic faces through morphology or 3-D reconstruction may be used to provide multiple virtual face views [27].

2.4. Domain adaptation

Domain adaptation (DA) methods have been applied in systems either in still-to-video or video-to-video FR [11,13]. Capturing faces in unconstrained environments and different locations translates to large differences between in the source and target distribu-

tions, due to different camera viewpoints, pose, illumination conditions, etc. Real-world scenarios for watch-list screening and person re-identification are most specially pertinent for unsupervised DA, because it is costly and requires human efforts to provide labels for faces in the target domain [28,29]. Recently, a discriminative transfer learning approach has been proposed for the SSPP problem that relies on exploiting a generic training set (source domain) to learn a feature projection and then transfer into the single sample gallery set (target domain) through performing discriminant analysis [30]. It attempts to minimize the differences between the source and target domains, and employs sparsity regularization to provide robustness against outliers and noise. In addition, an extended sparse representation classification approach through DA (ESRC-DA) has been proposed for still-to-video FR incorporating matrix factorization and dictionary learning [31].

Ensemble-based algorithms have been also proposed to address the problem of cross-resolution face matching, where low-resolution probes in the OD are matched against high-resolution gallery faces in the ED. In such a case, the classifiers that are trained only on the ED may not classify the test instances effectively, because of the variations in domain distributions [12,32]. For instance, an ensemble of pre-computed SVMs trained independently on the labeled samples from multiple EDs has been employed to enforce the decision boundary of a target classifier toward them using a domain-dependent regularizer [32]. An ensemble-based co-transfer learning framework has been developed in [33], where a semi-supervised approach is employed to integrate transfer learning and co-training in order to efficiently transfer the knowledge from the ED to the OD. This approach leverages few labeled and large unlabeled probe instances to enable knowledge transferring by combining the ensemble of SVM classifiers trained on the ED and OD. Followed by ensemble-based methods, individual-specific ensembles of e-SVMs are proposed in this paper for robust still-to-video FR, where the classifiers are trained using a single labeled high-resolution still from the ED versus few labeled high-resolution stills from the ED (cohort or other non-target individuals) and abundant unlabeled low-resolution videos from the OD.

3. Generation and selection of individual-specific ensembles

Techniques introduced in the literature that are relevant for the generation and selection of individual-specific ensembles are briefly presented including random subspace methods, classification systems, dynamic classifier selection and weighting. To overcome the challenges of designing a robust MCS according to the watch-list screening constraints, different techniques can be applied for ensemble generation. Bagging, boosting, and random subspace method (RSM) are well-known resampling techniques developed to efficiently tackle the small sample size problem [34]. The main idea of these methods is to enlarge the sample size by manipulating the training data. It is worth noting that bagging and boosting methods are not applicable to watch-list applications, since they require more than one target sample in the training set. Moreover, the selection and weighting of classifiers can be employed to consider the most competent classifiers within a pool [35] to adapt to the capturing context of each given probe.

The accuracy and diversity of classifiers within ensembles are key issues in ensemble-based systems [36,37]. Assuming both diversity and accuracy is not trivial, because diversity means disagreement among classifier predictions, whereas accuracy implies agreement on the predictions [38–40]. Individual-specific ensembles designed in this paper generate a diverse pool of classifiers per each individual of interest using processed face patches and random subspaces [41,42]. In particular, using faces from different domains (diversity of domains), considering different face patches without overlapping, exploiting different feature extraction techniques, as well as, randomly resampling different feature subsets allow to generate a diverse pool of classifiers.

3.1. Random subspace methods

RSMs randomly sample different feature subspaces from the original feature space of the input sample to create an ensemble of classifiers [43]. Let $\mathcal{F} = \{f_1, f_2, ..., f_d\}$ be the d-dimensional original feature space. To create a random subspace $\mathcal{R}$, s features are randomly sampled from $\mathcal{F}$. A feature vector belonging to the subspace $\mathcal{R}$ is denoted by $\mathbf{a} = [a_1, a_2, ..., a_s]$ and is used to train a classifier. This sampling process is repeated K times to create an ensemble of classifiers $C = \{c_1, \ldots, c_l, \ldots, c_K\}$, where using different subsets $\mathcal{R}$ encourage diversity among the classifiers c_l. The ensemble of classifiers C is therefore more suitable than a single classifier constructed with an instance from the complete feature space $\mathcal{F}$. Since RSM generates many redundant features, one of them may achieve higher accuracy compared to the original feature space. In the SSPP context, RSMs can provide different representations of the single training sample and inherit accuracy from classifier aggregation. RSM also helps avoiding over-fitting and is more robust to noisy data [44].

Ensemble of randomized linear discriminant analysis [45] is proposed based on constructing an ensemble using randomly selected features from the single training sample, where ensemble of these low-dimensional subspaces provides efficient functionality. Although this approach has achieved great success in some extents, it fails to deal with the existing variations in video surveillance environments, mostly due to feature extraction from a whole face rather than local parts of the face. However, semi-random subspace method has been proposed in [37] to randomly sample features from local regions of the face in a deterministic way instead of completely random way. In this regard, a face image is first divided into several regions, and then a set of base classifiers are constructed on different randomly sampled features extracted from each region, and finally all base classifiers are fused to provide the final decision. Recently, random subspaces was used to perform ensemble learning for makeup-robust FR by sampling features randomly from a set of features extracted from patches of before-makeup and after-makeup facial images [46]. Hence, random sampling on each local patch can gain more diversity between classifiers within the ensembles, because it generates multiple classifiers on different feature distributions, as well as, because it exploits local structure information to recognize faces efficiently under changes in expression, illumination, and occlusions [37].

3.2. Classification systems

Designing accurate classifiers for a MCS under imbalanced data situation is a challenging issue [8]. SVM is a well-known and widely used discriminative classifier that finds the optimal hyperplane to separate data patterns into binary classes. Thus, specialized 2-class SVMs are used to generate a pool of classifiers. Conventional 2-class SVM classifiers typically fail to find an optimal decision boundary in case of imbalance data [47]. However, different error costs (DEC) method [48] can be used to assign two misclassification cost values C^+ and C^- to manipulate the SVM objective function as follows:

$$min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^2 + C^+ \sum_{[i|y_i=+1]}^{l} \xi_i + C^- \sum_{[i|y_i=-1]}^{l} \xi_i \quad (1)$$

where $\mathbf{w}$ is the weight vector, b is the bias term, C^+ and C^- are the positive and negative misclassification costs to control the weight, respectively.

In the specialized approach proposed according to the existing constraints, classifiers are trained using a single target reference stills against many non-target samples. A method called exemplar-SVM (e-SVM) [49] has been proposed to train a separate SVM classifier with DEC for each individual of interest. It has shown effectiveness and generalization to design an individual-specific ensembles for still-to-video FR, where diversity of an e-SVM pool is provided using multiple representations [8]. It is worth mentioning that training many different e-SVM classifiers based on multiple representations and then combining their scores may avoid the issue of over-fitting. Since there is only a single positive sample in the training set, its error should be weighted much higher than the negative samples to avoid the skewness toward negatives. Let **a** be the target ROI pattern, **x** and U be sets of non-target ROI patterns (either labeled still ROIs or unlabeled video ROIs depending on the different training schemes) and their number, respectively. The cost function of e-SVM using a linear kernel is formalized as follows:

$$min_{\mathbf{w},b}\mathbf{w}^2+C_1\max(0,1-(\mathbf{w}^T\mathbf{a}+b))+C_2\sum_{\mathbf{x}\in U}\max(0,1-(\mathbf{w}^T\mathbf{x}+b)) \quad (2)$$

where C_1 and C_2 define the regularization weights, **w** is the classifiers weight vector, b is the bias.

3.3. Dynamic selection and weighting of classifiers

Selection of diverse and discriminative classifiers is a fundamental task in MCSs, where it can favorably decrease the risk of classifier over-generalization. The key idea of classifier selection is to select a set of classifiers $C^* \in C$ that contains the most appropriate classifiers for a given input pattern **t**. This task can be broadly categorized into static and dynamic selection methods [35]. Methods that select the classifiers statically are performed offline with a validation set, while dynamic selection methods exploit operational time information [50]. The latter is preferred to select the most locally accurate set of classifiers based on the context knowledge for each input pattern **t**.

A major issue to achieve a reliable dynamic selection scheme is to determine an accurate criterion in order to measure the level of competence among the base classifiers c_l within the pool C. Considering the SSPP problem, it is challenging to define desirable criteria to dynamically select and weight the most competent classifiers during operations. The notion of competence as a selection approach indicates the capability of classifiers to best fit the given classifier selection process. In other words, it reveals a measure to select the best classifiers regarding to different classification tasks [14]. In order to calculate the competence level of a base classifier, three different approaches were proposed in the literature [35]: (1) the local neighborhood accuracy (over a region around the input test pattern **t** in the feature space) [51–53], (2) decision templates or profiles (over a space declared by the base classifiers' output) [17,52], and (3) extent of consensus [18].

DS methods that have been proposed in the literature are mostly based on the local accuracy concept [35]. Thus, the accuracy of each classifier in the pool is estimated within the local region defined in a neighborhood of the pattern to be classified in the feature space (region of competence) [51]. Techniques using local accuracy to measure the competence are highly reliant on performance of the methods employed to define the neighborhood, such as k-NN. In contrast, as an advantage of decision templates techniques, it can be highlighted that they are not dependent on the quality of the region of competence over the feature space, while the decision space is considered to compute the similarity. Nevertheless, they only exploit global information of the base classifiers instead of the local expertise of them [17]. Furthermore, techniques based on extent of consensus are independent from the region of competence information, contrary to the local neighborhood accuracy techniques. However, since there are some ties among different members of the pool, an ensemble of classifiers with an acceptable consensus (level of confidence) may not be selected and the system may perform a random selection [18].

Dynamic weighting (DW) methods are similarly related to DS techniques, because they rely on the competence of classifiers [54]. Consequently, defining the appropriate competence is a key factor in the design of these techniques. A set of competent classifiers are dynamically selected from the ensemble to classify each input pattern in DS, while the scores of classifiers in the ensemble are weighted in DW. Previous studies reveal that using only one criterion as a level of competence is typically capable of selecting or weighting the classifiers dynamically and achieve a higher level of performance [14,55]. However, multiple criteria can be considered to measure the competence of classifiers in order to appropriately select or weight them.

In contrast with conventional ensemble-based fusion methods [56], static and dynamic classifier combination techniques cannot be directly utilized in the watch-list screening application with only a single reference still per target individual during the design [54,56]. Since each classifier in the proposed system is trained specifically for the target individual, classifiers dedicated to other individuals may not be a desired candidate. However, it is crucial to define suitable levels of competence that are not based on the closeness of a given probe (local neighborhood) to other instances of the target individual. Hence, the properties of e-SVMs can be exploited to define internal criteria in order to find the most competent ones according to the capturing context.

4. Dynamic individual-specific Ee-SVMs through domain adaptation

A novel ensemble learning approach is proposed in this paper to design accurate classification systems for each target individual enrolled to a still-to-video FR system. In particular, to improve robustness to intra-class variations, individual-specific Ee-SVMs models the single reference still ROI for the OD using several diverse e-SVMs based on multiple face representations and domain adaptation. During enrollment, each patch-wise e-SVM is trained for a different patch, descriptor and feature subset extracted from the single reference still ROI of the target individual (in the ED) versus those extracted from the abundance of still and video ROIs of non-target individuals (in either ED and OD). Several training schemes are proposed for unsupervised DA according to assumptions made for unlabeled video ROIs from the OD.

Two different scenarios are investigated for the design phase to select the most discriminant among a large number of representation subspaces (descriptors and feature subsets of a patch) for enrollment of target individuals (Ee-SVMs design). In the first design scenario, a validation set, containing stills and videos of some random non-target individuals, is exploited with a global criterion to effectively adapt the system to the actual context. Thus, the most accurate e-SVM classifiers (i.e., discriminative representation subspaces) are selected by ranking trained e-SVMs using a criterion based on the area under precision-recall curve [57], where these subspaces are used for enrollment of a target individual. In the second design scenario, the most informative representation subspaces are selected without considering a validation set. A local distance-based criterion is applied to rank and prune them, where the best subspaces are selected for enrollment of a target individual.

Since capture conditions change over time, the best ensemble to recognize the target individual will vary according to the given probe ROI. Pre-selection of the most discriminative representation subspaces during the design phase, as well as, selecting or weight-

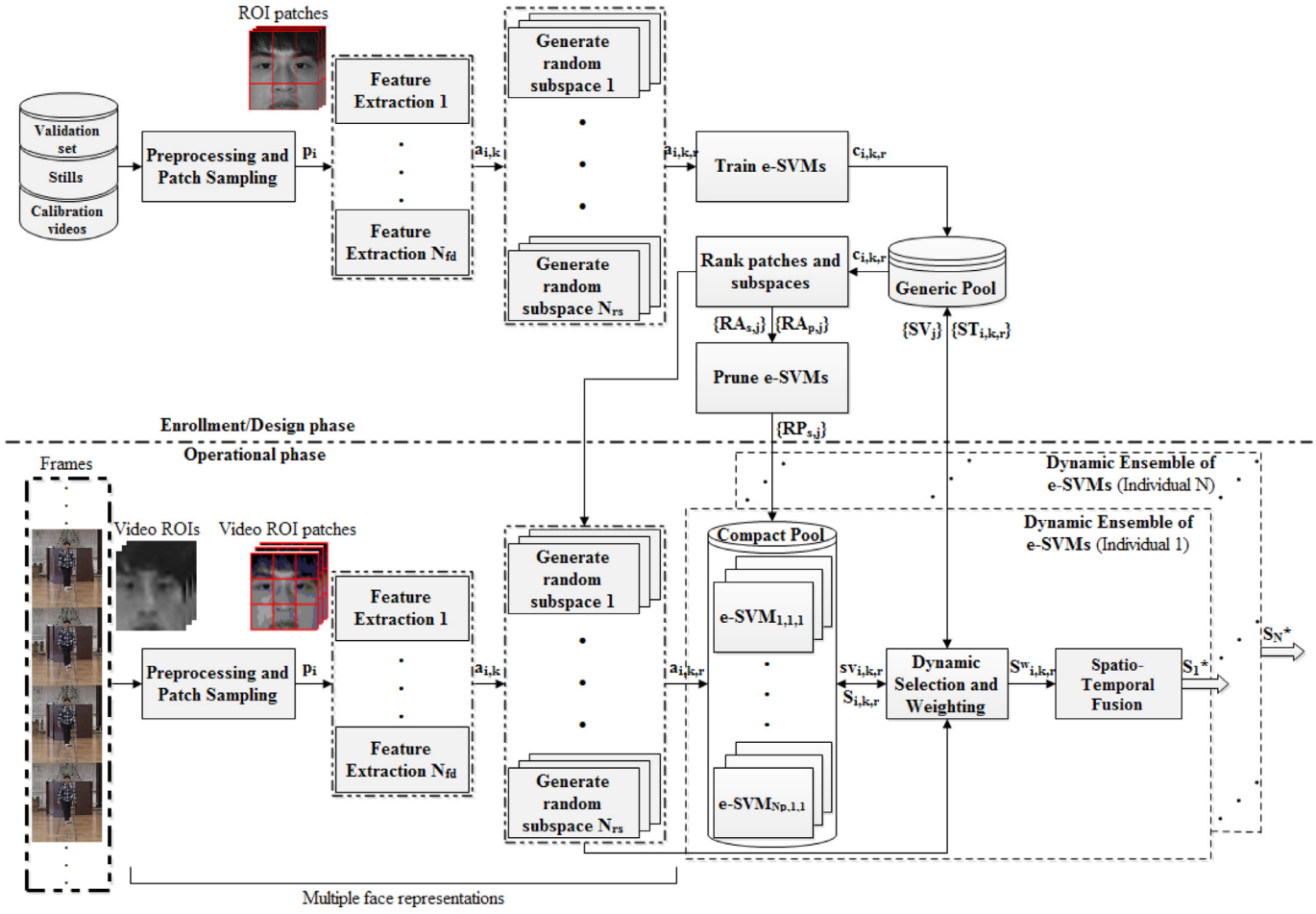


Fig. 2. The enrollment and operational phases of the proposed multi-classifier system for accurate still-to-video FR.

ing the most competent classifiers during the operational phase can provide a higher level of performance at a lower computational complexity in such a real-time application, unlike employing fusion over the entire pool.

4.1. System overview

A block diagram of the proposed MCS for still-to-video FR is shown in Fig. 2. It generates a diverse and compact pool of classifiers during the design phase, and selection and weighting ensembles dynamically during the operational phase. Each step of the proposed system is described in the following subsections.

During the design phase (Enrollment/Design phase), a pool of diverse e-SVM classifiers is generated per individual of interest. Multiple different facial representations are produced over all patches for several face descriptors and random subspaces. The parameters of the proposed system, such as number of patches, number and size of feature subspaces are defined in this phase. Different number of classifiers are trained for each patch based on their significances on performance using the best subspaces (representations) that were already ranked.

During the operational phase, classifiers of the pool are selected or weighted dynamically according to competence for classifying the given input probe (ROI), and then their scores are combined to obtain the final score. The proposed system exploits two levels of information fusion. First, the fusion of subspace-wise classifiers selected during operations from corresponding face descriptor (patch-level fusion), and then the fusion of patch-wise classifiers generated by the face descriptors (descriptor-level fusion).

4.2. Design phase (first scenario)

In this scenario for the design phase, a compact pool of e-SVM classifiers is generated using semi-random subspaces pruned based on the most informative pre-ranked patches. This phase is performed off-line, and as shown in Fig. 2 (Enrollment/Design phase), it consists of patch-wise feature extraction, training patch-wise e-SVMs, as well as, ranking patches and pruning subspaces to select the best subspaces (representations). Note that in this scenario, the labeled stills and video trajectories correspond to some unknown individuals or actors appearing in the scene, and are used to estimate system parameters and pre-selection of the best subspaces. Then, the pre-selected subspaces are used to design an Ee-SVMs for individuals of interest based on a single labeled still.

The validation set D consists of labeled high-quality stills and unlabeled low-quality videos defined as $D = \{ST_1^l, \ldots, ST_j^l, \ldots, ST_{N_a}^l \cup T_1^l, \ldots, T_j^l, \ldots, T_{N_a}^l \cup T_1^u, \ldots, T_v^u, \ldots, T_{N_v}^u\}$, where ST_j^l and T_j^l represent the labeled still and video trajectory of individual j, respectively, and T_v^u denotes the unlabeled video trajectory of unknown person v. N_a indicates the number of unknown non-target individuals in the validation set, where the number of videos is equal to N_v. All the stills and videos are segmented and scaled to the resolution of M_cxN_c. As illustrated in Fig. 2, all still ROIs of ST_j^l and video ROIs of T_j^l and T_v^u are first divided into m_cxn_c pixels patches $P_j^l = \{p_i^l\}$ and $P_v^u = \{p_i^u\}$, where $i = [1, 2, \ldots, N_p]$ and $N_p = (M_c/m_c) \times (N_c/n_c)$ is the total number of patches. Afterwards, feature extraction techniques (face descriptors) $FD = \{f_k\}$ are applied to extract feature sets $F_j^l = \{\mathbf{a}_{i,k}^l\}$ and $F_v^u = \{\mathbf{a}_{i,k}^u\}$ from patch

p_i, for $k = 1, 2, \ldots, N_{fd}$ and N_{fd} is the number of face descriptors. Thus, $\mathbf{a_{i,\,k}}$ defines the descriptor f_k extracted from patch p_i. Then, different random subspaces $RS = \{s_r\}$ with the dimension N_d are randomly selected from F_j^l and F_v^u to generate random subspaces $R_j^l = \{\mathbf{a}_{i,k,r}^l\}$ and $R_v^u = \{\mathbf{a}_{i,k,r}^u\}$, for $r = 1, 2, \ldots, N_{rs}$, and N_{rs} is the total number of random subspaces. Hence, $\mathbf{a_{i,\,k,\,r}}$ denotes the feature subspaces s_r randomly selected from $\mathbf{a_{i,\,k}}$.

To construct a compact pool of classifiers $P_c = \{Ej | 1 \le j \le N_a\}$, ensemble of e-SVM classifiers $E_j = \{C_l | 1 \le l \le N_P \cdot N_{fd} \cdot N_{rs}\}$ are trained to enroll a target individual j. The number of random subspaces $RP_{s,j} = \{s_r | 1 \le r \le N'_{rs}\}$ is determined based on the significance of patches $RA_{p,j}$ and their rankings $RA_{s,j}$ to train accurate classifiers $c_{i,\,k,\,r}$ (See Algorithm 3). However, all the subspaces $RS = \{s_r\}$ are employed to construct a generic pool of classifiers $P_g = \{Ej | 1 \le j \le N_a\}$, where $E_j = \{C_l | l = 1, 2, \ldots, N_P \cdot N_{fd} \cdot N_{rs}\}$ as formalized in Algorithm 1.

Algorithm 1 Generic pool generation.

Input: Validation set $D = \{ST_1^l, \ldots, ST_j^l, \ldots, ST_{N_a}^l \cup T_1^l, \ldots, T_j^l, \ldots, T_{N_a}^l \cup T_1^u, \ldots, T_v^u, \ldots, T_{N_v}^u\}$
Output: Generic pool of e-SVM classifiers $P_g = \{Ej | 1 \le j \le N_a\}$
▷ Constructing an ensemble of e-SVMs
for each individual j in D **do**
  Divide ST_j^l, and T_v^u into patches P_j^l and P_v^u of size $m_c x n_c$
  for each patch $i = 1 \ldots N_p$ **do**
    for each face descriptor $k = 1 \ldots N_{fd}$ **do**
      ▷ Patch-wise feature extraction
      $\mathbf{a}_{i,k} \leftarrow$ extract face descriptors f_k from patch p_i
      for each random subspace $r = 1 \ldots N_{rs}$ **do**
        $\mathbf{a}_{i,k,r} \leftarrow$ randomly sample subspaces s_r from $\mathbf{a}_{i,k}$
        ▷ Training patch-wise e-SVM classifiers
        $E_j \leftarrow$ train a classifier $c_{i,k,r}$
      end for
    end for
  end for
end for

As formulated in the Algorithm 1, labeled still ST_j^l and unlabeled video ROIs T_v^u in the validation set D are employed to train patch-wise e-SVM classifiers and subsequently, to build a generic pool of classifier $P_g = \{Ej | 1 \le j \le N_a\}$ based on DA using multiple face descriptors. To that end, an ensemble of e-SVMs E_j is constructed for each individual in D and stored within the generic pool.

Semi-random subspaces selected during this phase are utilized to increase the probability of generating representative facial models that are robust to nuisance factors existing in the surveillance environments. However, due to a loss of information in some of the subspaces, selecting a suitable size of patches and random subspaces are essential. The time complexity and accuracy are dependent to these parameters. Smaller rate of random sampling causes to perform faster, but simultaneously it may miss useful discriminant features subsets. On the other hand, larger rate may also cause less diversity among classifiers.

4.2.1. Patch-wise feature extraction

In this paper, the patches in each face are represented using LPQ and HOG descriptors [58,59], although many other face descriptors may be suitable. The choice of face descriptors is based on the complementary robustness that they provide to the nuisance factors in surveillance environments [6]. Previous study suggests that the combination of these descriptors is capable of providing a high level of discrimination on the SSPP problem [8,25].

LPQ extract texture features of the face images from frequency domain through Fourier transform and has shown high robustness to motion blur. LPQ is based on the blur insensitive property of the Fourier phase spectrum. The phase is computed in local rectangular M-by$M$ neighborhoods N_x at each pixel position x of the image $f(x)$ using a short-term Fourier transform defined by:

$$F(\mathbf{u}, x) = \sum_{y \in N_x} f(x - y) e^{-j2\pi \mathbf{u}^T y} = \mathbf{w_u}^T \mathbf{f_x} \tag{3}$$

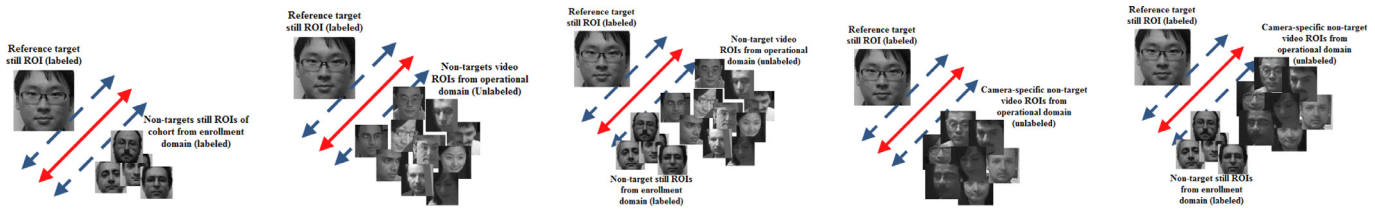
where $\mathbf{w_u}$ is the basis vector of the 2-D discrete Fourier transform at frequency $\mathbf{u}$, and $\mathbf{f_x}$ is another vector containing all M^2 values of f in N_x. It is examined for all positions $x \in \{x_1, x_2, \ldots, x_N\}$ at four frequency points $\mathbf{u} \in \{\mathbf{u_1}, \ldots, \mathbf{u_4}\}$ that results in a vector $\mathbf{F_x}$. The phase information is obtained using the signs of each component in the $\mathbf{F_x}$ by a simple scalar quantizer $q_j(x)$, where $q_j(x)$ is the jth component of the Fourier coefficients. Then, the label image $f_{LPQ}(x)$ with blur invariant LPQ values is represented by eight binary coefficients $q_j(x)$ as integer values between 0-255 using the binary coding $f_{LPQ}(x) = \sum_{j=1}^{8} q_j(x) 2^{j-1}$. Finally, the histograms of labels $f_{LPQ}(x)$ from different non-overlapping rectangular regions are concatenated to build the 256-dimensional LPQ face descriptor.

On the other hand, HOG extract gradients, and it is more robust to pose and scale changes, as well as, rotation and translation. In particular, the occurrences of gradient orientations are counted in each local neighborhood of an image. The image is divided into different blocks and cells (small connected regions) for a block spacing stride of l pixels. Then a histogram of gradient orientations is computed for each cell within the blocks. According to the sign of gradients, the channels of each histogram can be varied over $0 - 180^\circ$ or $0 - 360^\circ$ for unsigned and signed, respectively with 9 orientation bins. The histograms are normalized using color and Gamma correction with L2-Hys threshold for robustness against illumination and scale. Finally, the combination of normalized group of histograms in all cells and blocks represents the HOG face descriptor.

4.2.2. Training patch-wise e-SVM classifiers

In order to learn the individual-specific Ee-SVM for target individual j based on DA, the 5 training schemes have been considered by employing either labeled still ROIs ST_j^l from the cohort or other non-target individuals or unlabeled video ROIs T_v^u captured from the operational domain.

1. Scheme 1 (target still ROI vs non-target still ROIs): The single labeled target still and non-target still ROIs from cohort model are employed to train e-SVMs without exploiting unlabeled video ROIs. Thus, videos in the OD are not employed for DA (see Fig. 3(a)).
2. Scheme 2 (target still ROI vs non-target video ROIs): The single labeled target still ROI are considered with an abundance of unlabeled non-target video ROIs from the OD (see Fig. 3(b)).
3. Scheme 3 (target still ROI vs non-target stills and video ROIs): Labeled non-target still ROIs from the cohort model are considered in addition to video ROIs from the OD (see Fig. 3(c)).
4. Scheme 4 (target still ROI vs unlabeled non-target camera-specific video ROIs): Unlabeled video ROIs captured using a specific camera FoV are exploited along with the labeled target still ROI in order to construct a camera-specific pool. Thus, several camera-specific pools equivalent to the number of surveillance cameras are constituted (see Fig. 3(d)).
5. Scheme 5 (target still vs non-target stills and camera-specific video ROIs): Labeled non-target still ROIs with unlabeled camera-specific video ROIs are considered versus the single target still ROI in order to build several camera-specific pools (see Fig. 3(e)).



(a) Scheme 1 (b) Scheme 2 (c) Scheme 3 (d) Scheme 4 (e) Scheme 5

Fig. 3. A 2-D illustration of e-SVM in the feature space trained using different classification schemes according to DA. (a) a target still vs labeled non-target still ROIs of ED, (b) a target still vs unlabeled non-target video ROIs of Od, (c) a target still vs labeled non-target still ROIs of Ed and video ROIs of OD, (d) a target still vs unlabeled non-target camera-specific video ROIs of Od, and (e) a target still vs labeled non-target still ROIs of Ed and unlabeled non-target camera-Specific video ROIs of Od.

To assess the 5 aforementioned training schemes, all the classifiers in the generic pool are tested to obtain the system performance. However, the best scheme is adopted to learn the individual-specific Ee-SVMs in the proposed system. To accomplish DA, unlabeled video ROIs captured from the OD allow to incorporate the knowledge of operational domain during generation of the pool. Therefore, an unsupervised DA approach is considered, where labeled still ROIs from the cohort model and unlabeled video ROIs captured from the OD are employed to train classifiers in the enrollment domain. As illustrated in Fig. 3(c), this training scheme favors the transfer of knowledge from either ED or OD to the classifiers trained specifically for each individual of interest.

4.2.3. Ranking patch-wise and subspace-wise e-SVMs

During the design prior to the enrollment, $N_p \cdot N_{rs}$ classifiers are trained for individuals in the validation set according to each face descriptor f_k. Then, these classifiers are combined using the mean fusion function over the random subspaces s_r (patch-level fusion). Subsequently, N_p classifiers are evaluated and ranked $RA_{p,j}$ using the global system performance based on the area under precision-recall (AUPR) as formulated in Algorithm 2. Noted that N_{rs} constant subspaces are selected from each patch, because it is tended to rank the significance of patches p_i based on the information encapsulated in each one.

In addition, to rank the subspaces s_r selected randomly from each patch p_i, the $N_p \cdot N_{rs}$ classifiers in the P_g are combined over the patches and the corresponding performance is similarly evaluated as in Algorithm 2. Thus, each feature subset is ranked and its corresponding classifier retained in $RA_{s,j}$ according to ranking of patches already preserved in $RA_{p,j}$.

These ranking processes allow the pre-selection of e-SVM classifiers according to the best representations (feature subsets) during the design. It allows generating the less number of more accurate classifiers for each patch through patch ranking during the enrollment of target individuals.

4.2.4. Pruning subspaces-wise e-SVMs

After ranking patches and subspaces, a pruning process is used to select a variable numbers of the ranked subspaces from each patch as shown in Algorithm 3. A larger the number of subspaces are selected for the most relevant patches. In order to select different number of subspaces for each patch, a criterion is deployed as follows according to the overall AUPR performance obtained using all the classifiers in the pool $c_{i,k,r}$ and AUPR performance gained by corresponding all the classifiers of each patch $c_{i,k}$:

$$N'_{rs} = \left\lceil N_{rs} . \frac{AUPR(c_{i,k})}{AUPR(c_{i,k,r})} \right\rceil \tag{4}$$

where R_{pruned} contains N'_{rs} ranked subspaces s_r (integer values using a ceiling function) for each patch p_i. It allows to constitute the compact pool and accordingly, the dynamic classifier selection can

Algorithm 2 Ranking of patch-wise and subspace-wise e-SVMS.

Input: Validation set D and generic pool P_g
Output: Ranking of patches $RA_{p,j}$ and subspaces $RA_{s,j}$
for each individual j in D **do**
  for each face descriptor $k = 1...N_{fd}$ **do**
    ▷ Ranking patch-wise classifiers
    for each patch $i = 1...N_p$ **do**
      $RA_{p,j} \leftarrow \{\emptyset\}$
      Combine classifiers $c_{i,k}$ over random subspaces s_r using the mean fusion function
      $RA_{p,j} \leftarrow$ rank patches p_i in descending order of the AUPR obtained using $c_{i,k}$
    end for
    ▷ Ranking subspace-wise classifiers
    for each random subspace $r = 1...N_{rs}$ **do**
      $RA_{s,j} \leftarrow \{\emptyset\}$
      Combine classifiers $c_{k,r}$ over patches p_i using the mean fusion function
      $RA_{s,j} \leftarrow$ rank subspaces s_r in descending order of the AUPR obtained using $c_{k,r}$
    end for
  end for
end for

Algorithm 3 Pruning subspace-wise e-SVMs and compact pool generation.

Input: Validation set D, generic pool P_g, ranked patches $R_{p,j}$, ranked subspaces $R_{s,j}$, and phase *phase*
Output: Compact pool of e-SVM classifiers $P_c = \{Ej | 1 \le j \le N_a\}$
for each individual of interest $j = 1...N_a$ **do**
  for each face descriptor $k = 1...N_{fd}$ **do**
    if design phase **then** ▷ Pruning subspace-wise e-SVMs
      for each patch $i = 1...N_p$ in the R_{patch} **do**
        $N'_{rs} \leftarrow \left\lceil N_{rs} . \frac{AUPR(c_{i,k})}{AUPR(c_{i,k,r})} \right\rceil$
        $RP_{s,j} \leftarrow$ select N'_{rs} subspaces from $R_{s,j}$ for each patch p_i
      end for
    end if
    if enrollment phase **then** ▷ Constructing a compact pool (enrollment)
      for each random subspace $r = 1...N'_{rs}$ in the $RP_{s,j}$ **do**
        $E_j \leftarrow$ train $c_{i,k,r}$ to construct a compact pool of classifiers
      end for
    end if
  end for
end for

be accomplished with the lowest number of classifiers during operations. However, the best subspaces are found during the design phase and those subspaces are employed to train e-SVMs for each individual in the watch-list during the enrollment phase.

4.3. Design phase (second scenario)

In this scenario relies on the over-produce and select paradigm, where a large number of subspaces are generated for each individual of interest during the design phase of the system. Then, e-SVM classifiers are trained and the best subspaces are selected during the enrollment phase. In the proposed system, several feature subspaces are randomly produced for each patch, and these subspace are ranked $RA_{s,j}$ based on a distance-based local criterion to select the best set of subspaces ($N'_{rs} \ll N_{rs}$). They can be employed to construct a compact pool of classifiers as presented in Algorithm 4.

Algorithm 4 Ranking subspace-wise e-SVMs and compact pool generation.

1: Input: Labeled still ROIs of target individuals $ST_1^l, \ldots, ST_j^l, \ldots, ST_{N_a}^l$ and unlabeled video ROIs of non-target individuals $T_1^u, \ldots, T_v^u, \ldots, T_{N_v}^u$, and phase *phase*
2: Output: Compact pool of e-SVM classifiers $P_c = \{Ej | 1 \leq j \leq N_a\}$
3: **for** each individual of interest $j = 1 \ldots N_a$ **do**
4: **for** each patch $i = 1 \ldots N_p$ **do**
5: **for** each face descriptor $k = 1 \ldots N_{fd}$ **do**
6: **if** *phase* = *design* **then** ▷ Over-producing subspaces
7: **for** each random subspaces $r = 1 \ldots N_{rs}$ **do**
8: $\mathbf{a}_{i,k,r} \leftarrow$ randomly sample subspaces s_r from $\mathbf{a}_{i,k}$
9: **end for**
10: **end if**
11: **if** *phase* = *enrollment* **then** ▷ Training classifiers and ranking subspace-wise e-SVMs
12: $E_j \leftarrow$ train a classifier $c_{i,k,r}$
13: $RA_{s,j} \leftarrow$ rank subspaces in descending order based on the $\text{dist}(\mathbf{ST}_{i,k,r}, \mathbf{sv}_{i,k,r})$
14: ▷ Constructing a compact pool (enrollment)
15: **for** random subspaces $r = 1 \ldots N'_{rs}$ in the $RA_{s,j}$ **do**
16: $E_j \leftarrow$ preserve $c_{i,k,r}$ to constitute a compact pool of classifiers
17: **end for**
18: **end if**
19: **end for**
20: **end for**
21: **end for**

The proposed ranking criterion is based on distance of the still ROI and the support vectors of e-SVMs dist($\mathbf{ST}_{i,k,r}$, $\mathbf{sv}_{i,k,r}$) in the feature space. It is assumed intuitively that those subspaces used for training are the most relevant ones, where the corresponding e-SVM classifiers have a larger distance to the target still than others. Subspaces are thereby ranked in descending order based on distance between the target still $\mathbf{ST}_{i,k,r}$ and e-SVM support vectors $\mathbf{sv}_{i,k,r}$ in the feature space (see Fig. 4). N_{rs} set the number of over-produced subspaces, and N'_{rs} be the number of ranked subspaces.

4.4. Operational phase (dynamic classifier selection and weighting)

An important challenge is to derive accurate measures for classifier competence in the context of the SSPP problem. The proposed approach allows the still-to-video FR system to select the classifiers that are most competent for the capture conditions. A new distance-based DS approach is proposed to provide the best classifiers to discriminate between the target and non-target ROIs.

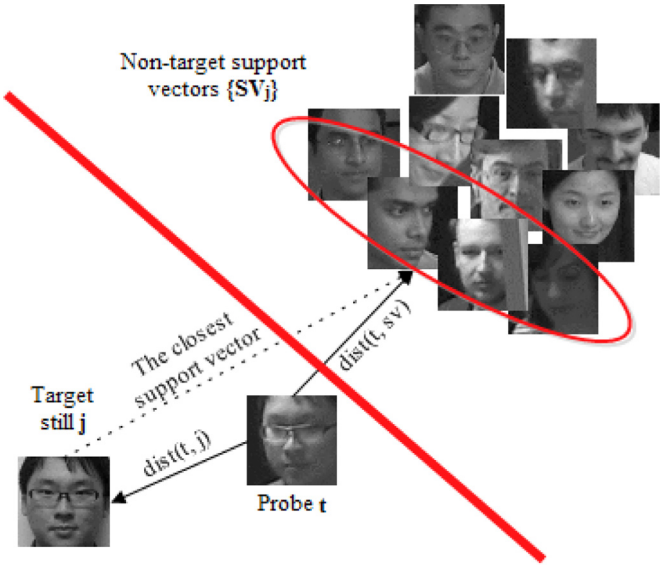


Fig. 4. A 2-D illustration of the proposed dynamic classifier selection approach in the feature space.

In order to dynamically select the most competent classifiers for the design of a robust ensemble, the proposed internal criteria (levels of competence) per given probe ROI relies on the: (1) distance from the non-target support vectors ROI patterns, and (2) closeness to the target still ROI pattern. The key idea is to select the classifiers that effectively locate the given probe ROI pattern close to the target still in the feature space. If the distance between the probe and the target still ROI pattern is lower than the distance to support vectors, then those classifiers are selected dynamically as competent classifiers for the given probe ROI pattern.

The distance from support vectors can be defined based on the distance to the closest support vector to the target still. On the other hand, the classifiers with support vectors that are far from the ROI test patterns of individuals of interest can be also suitable candidates, because they may classify probe ROI patterns correctly. In the proposed DS approach (illustrated in Fig. 4), all the non-target support vectors were sorted based on their distance to the target still (the target support vector) in an offline processing. Then, the closest support vector to the target still is used to compare with the input probe.

During operations, each given probe ROI **t** is projected in the feature space and those classifiers form the pool that verify the selection criteria (locate the input near the target still and far from support vectors) are selected dynamically, and their scores are combined using score-level fusion. In contrast to the approaches that use local neighborhood accuracy for measuring the level of competence, it is not mandatory in the proposed method to define neighborhood using all the validation data, like with method based on, e.g., kNN. Thus, different distance metrics, such as Euclidean, CityBlock, Hamming, etc., can be employed to measure the distances between ROI patterns and support vectors. The algorithm of proposed classifier selection method is formalized in Algorithm 5.

As described in the Algorithm 5, each given input ROI t is first divided into patches p_i. Then, feature extraction technique f_k is applied on each patch to form a feature vector $\mathbf{a}_{i,k}$ per patch. Afterwards, the ranked subspaces stored in the $RA_{s,j}$ are sampled from $\mathbf{a}_{i,k}$ and then $\mathbf{a}_{i,k,r}$ is projected into the feature space containing support vectors $\{SV_j\}$ of classifiers and the reference still $\mathbf{ST}_{i,k,r}$ of target individual j. Finally, those classifiers c_l in E_j that satisfy the levels of competence criteria (line 13) are selected to constitute C_j^* in order to classify testing sample **t**. Subsequently, the scores of selected classifiers $S_{i,k,r}$ are combined using mean function to provide

Algorithm 5 Operational phase with DS.

Input: Pool of e-SVM classifiers E_j for individual of interest j, the set of support vectors $\{SV_j\}$ per E_j
Output: Scores of dynamic ensembles based on a subset of the most competent classifiers C_j^*
for each probe ROI t **do**
  Divide testing ROI t into patches after preprocessing
  for each patch $i = 1...N_p$ **do**
    for each face descriptor $k = 1...N_{fd}$ **do**
      $\mathbf{a}_{i,k} \leftarrow$ extract features f_k from patch p_i
      for each subspace $r = 1...Nrs$ **do**
        $\mathbf{a}_{i,k,r} \leftarrow$ sample subspaces s_r from $RA_{s,j}$
        $C_j^* \leftarrow \{\emptyset\}$
        for each classifier c_l in C_j **do**
          if $\text{dist}(\mathbf{a}_{i,k,r}, \mathbf{ST}_{i,k,r}) \leq \text{dist}(\mathbf{a}_{i,k,r}, \mathbf{sv}_{i,k,r})$ **then**
            $C_j^* \leftarrow c_l \cup C_j^*$
          end if
        end for
      end for
    end for
  end for
  if C_j^* is empty **then**
    $S_j^* \leftarrow$ Use mean scores of E_j to classify t
  else
    $S_j^* \leftarrow$ Use mean scores of C_j^* to classify t
  end if
end for

final score S_j^*. All the classifiers in E_j are combined to classify ROI t when none of classifier fulfill the competence criteria.

In the proposed system, the ground-truth tracks are also exploited allowing to accomplish a robust spatio-temporal recognition. To that end, ROI captures for different individuals are regrouped through facial trajectories. In particular, decision fusion module accumulates the scores S_j^* of each individual-specific ensemble over a fixed size window W to make a decision d_j^* as follows:

$$d_j^* = \sum_{w=0}^{W-1} S_j^*\left[S_{i,k,r(W-w)}\right] \in [0, W] \tag{5}$$

Dynamic weighting of e-SVMs is suitable for rapid adaptation of individual-specific ensembles to tackle the variations within the operational domains. In this case, a distance-based combination strategy is also proposed to dynamically weight the scores of e-SVMs, where it relies on the distance of the probe instance to the support vectors of each classifier, as well as, to the target reference still in the feature space. This approach aims to reduce the effect of non-competent classifiers when their support vectors are closer to the given probe than the target still. Higher weights are assigned to the scores of classifiers with larger distance to the probe with respect to closeness to the single target still, and vice versa. Hence, each probe ROI pattern is compared to that of the single target still, and to that of the support vector of each classifier. If distance with the target still is closer than the closest support vector, then those classifiers are attributed higher weights. The proposed DW strategy is formalized in Algorithm 6.

Algorithm 6 Dynamic classifier weighting strategy.

Compute the distances of the probe with the closest support vector of each e-SVM and the target still, then store these distances $\text{dist}(t, sv)$, and $\text{dist}(t, j)$, respectively
Weight the scores of a classifiers s_k and create the weighted scores $s_k^w = s_k.w_k$, where the w_k is the relative competence of the classifier c_k on its corresponding weighted scores s_k^w estimated as $w_k = \frac{\text{dist}(t,sv)^2}{\text{dist}(t,sv)^2 + \text{dist}(t,j)^2}$
Use the mean fusion of weighted scores s_k^w to obtain the final score after score normalization

5. Experimental methodology

Several aspects of the proposed system are assessed experimentally using real-world video surveillance data. First, different e-SVM training schemes are compared for the individual-specific ensembles. Second, different pool generation scenarios are evaluated in terms of accuracy and time complexity. Finally, the impact of applying DS and DW are analyzed on the performance.

5.1. Video data

To evaluate performance, two publicly-available still-to-video datasets called COX-S2V[1] [19] and Chokepoint[2] [20] are employed. COX-S2V contains high-quality still images and low-quality video sequences of 1000 subjects. A digital camera is utilized to capture the frontal face images of subjects under controlled conditions, while the video cameras capture videos of subjects in uncontrolled conditions using two different off-the-shelf camcorders. In these videos, subjects walking through a designed-S curve with changes in illumination, expression, scale, viewpoint, and blur. Thus, four video sequences are recorded per subject simulating video surveillance scenario. An example of one subject is demonstrated in Fig. 5, showing the differences between ROIs captured in the ED and OD. This is also challenging, because there are only about 25 facial captures for each sequence.

Another publicly available dataset that can be used to validate the proposed system is Chokepoint. It consists of high-quality faces captured with a still camera, and videos captured with three video cameras under controlled and uncontrolled conditions, respectively. During four sessions, 29 subjects walk through different portals, and videos were recorded using an array of three cameras located above the portals.

5.2. Experimental protocol

In experiments on COX-S2V, the high-quality stills for $N_{wl} = 20$ individuals are randomly chosen to populate the watch-list, as well as, $N_{wl} = 10$ for evaluation of different training schemes. In addition, N_{ntd} video sequences of non-target persons from the OD are selected as calibration videos for the design phase. Moreover, N_{ntu} video sequences of unknown persons are considered for the operational phase. Hence, different subsets of COX-S2V are separated as demonstrated in Fig. 6 according to design scenarios, validation, and operational phases of the proposed system. Validation set D as required in the first design scenario is separated to define the system parameters containing $N_a = 20$ stills and videos of some random individuals along with $N_{ntd} = 100$ (to calibrate for cameras and scores) and $N_{ntu} = 100$ testing videos of other unknown persons for the design and operational phases, respectively. Design set to create facial models (generating a pool of classifiers) including high-quality stills of watch-list individuals $N_{wl} = 20$ and low-quality calibration videos of non-target persons $N_{ntu} = 100$. Operational set (test set) to assess the system performance that consists of N_{ntu} videos belonging to another set of unknown persons, as well as, videos of a target individual. During operations, one target

[1] http://vipl.ict.ac.cn/resources/datasets/cox-face-dataset/COX-S2V.
[2] http://arma.sourceforge.net/chokepoint/.

Fig. 5. An example of a still image belonging to one subject and corresponding four video sequences in the COX-S2V.

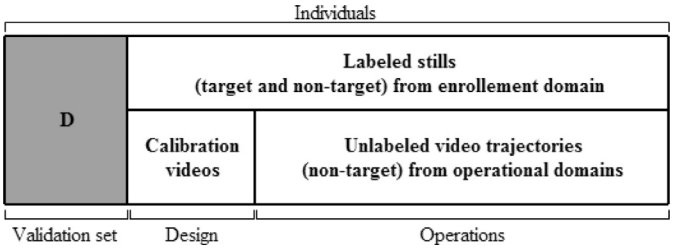


Fig. 6. The separation of COX-S2V dataset for validation, design, and operational phases of the proposed system.

individual is considered at a time along with non-targets in the operational scene. In order to achieve statistically significant results, these experiments are replicated 5 times with considering different stills and videos of individuals of interest as watch-list persons.

In experiments on Chokepoint, stills of $N_{wl} = 5$ individuals of interest are considered to constitute the watch-list. Videos of $N_{ntd} = 10$ unknown persons are used as calibration videos to construct a pool of e-SVM classifiers, and videos of $N_{ntu} = 10$ other non-target individuals are associated for the operations along with videos of watch-list individuals.

The facial ROIs appearing in reference stills and video frames were isolated in the COX-S2V and Chokepoint using the viola-Jones face detection. The reference stills and video ROIs are all converted to grayscale and scaled to a common size of 48x48 pixels for computational efficiency [23]. Histogram equalization is used to enhance contrast, as well as, to eliminate the effect of illumination changes. Then, an uniform non-overlapping patch configurations is applied to divide each ROI into 9 blocks of 16x16 pixels as in [8,46]. HOG and LPQ feature extraction techniques are utilized to extract discriminating features with the dimensions of 192 and 256, respectively. For HOG face descriptor, 3x3 pixel cells are considered with unsigned gradients, spacing stride of $l = 2$, and the default value of L2-Hys threshold. In addition, numbers and dimensions of feature subspaces are shown in Fig. 7. Libsvm library [60] is used in order to train e-SVMs, where the same regularization parameters $C_1 = 1$ and $C_2 = 0.01$ are considered for all exemplars (w of a target sample is 100 times greater than non-targets) [8]. Random subspace sampling with replacement is also employed to generate different subspaces randomly from feature space.

Ensemble of template matchers (TMs) and e-SVMs using multiple face representations [6,8], specialized kNN adapted for video surveillance (VSkNN) [5], sparse variation dictionary learning (SVDL) [61], and ESRC-DA [31] are considered as the base-line and state-of-the-art FR systems to validate the proposed system. In kNN experiment, PCA is applied for ROIs [62] are employed to compute the VSkNN using $k = 3$ (1 target still from the cohort model along with 2 nearest non-target video ROIs). To that end, distances of the probe ROI t are calculated from the target still ST_j, as well as, two nearest non-target T_1 and T_2 from the calibration videos. Thus, VSkNN score (S_{VSkNN}) is obtained as follows [5]:

$$S_{VSkNN} = \frac{\text{dist}(t, ST_j)}{\text{dist}(t, ST_j) + \text{dist}(t, T_1) + \text{dist}(t, T_2)} \quad (6)$$

where dist(t, ST_j) is the distance of the probe face t from the target still ST_j, dist(t, T_1) and dist(t, T_2) are the distances of the given probe t from the two nearest non-target captures, respectively.

In SVDL experiment, high-quality stills belonging to the individuals of interest are considered as a gallery set and low-quality videos of non-target individuals are employed as a generic training set to learn a sparse variation dictionary. Three regularization parameters λ_1, λ_2, and λ_3 set to 0.001, 0.01, and 0.0001, respectively, and also the dimensionality of faces is reduced to 90 using PCA according to the default values defined in [61]. The number of dictionary atoms are initialized to 100 based on the number of stills in the gallery set, where it is a trade-off between the computational complexity and the level of sparsity.

5.3. Performance metrics

The performance of still-to-video FR systems are typically assessed at the transaction-level to evaluate matching of Ee-SVMs for each ROI pattern (target versus non-target). Transaction-level analysis can be shown in the receiver operating characteristic (ROC) curves, in which true positive rates (TPRs) are plotted as a function of false positive rates (FPRs) over all threshold values. The proportion of target ROIs that correctly classified as individuals of

interest over the total number of target ROIs in the sequence is considered as TPR. Meanwhile, FPR is the proportion of non-target ROIs incorrectly classified as individuals of interest over the total number of non-target ROIs. In a ROC space, a global scalar metric of the detection performance is the area under ROC curve (AUC), which can be interpreted as the probability of classification over the range of TPR and FPR. In other words, the AUC indicates correct ranking of positive-negative pairs in terms of class separation. For instance, AUC=100% shows an accurate discrimination among samples, where all positive are perfectly ranked higher than negatives.

In still-to-video FR system scenario, class priors of targets and non-targets may vary over time in each sequence. However, conventional ROC curves and AUC allow for evaluating the performance that is independent of mis-classification costs and class priors between classifiers. Thus, the precision-recall space can be employed in order to estimate the performance of the system at transaction-level, where it can characterize performance as the fraction of the correctly detected target ROIs against the total number of ROIs predicted belonging to an individual of interest. It is suitable to measure the system performance under highly imbalanced data situation during operations. Recall can be corresponded as TPR and precision (P) is computed as follows $P = TP/(TP + FP)$.

In transaction-level analysis, performance of the watch-list screening system is provided using partial AUC (pAUC) and area under precision-recall (AUPR). Thus, pAUC(20%) is calculated using the AUC at $0 < FPR \leq 20\%$ in the ROC curve. The AUPR is desirable to illustrate the global accuracy of the system in the skewed imbalanced data circumstances. Experiments are iterated for each individual of interest in the watch-list for all video sequences, and then the average values are reported along with standard errors.

Moreover, the ground-truth track is employed to gradually group the captured ROIs over consecutive frames to create a trajectory due to trajectory-level analysis. To that end, captured ROIs of each individual in the operational scene are processed separately and the spatio-temporal fusion module accumulates ensemble scores over a window of fixed size to obtain the highest value inside the window in order to plot a ROC curve. Then the entire AUC is reported as a trajectory-level performance.

There exist several measures to estimate the ensemble diversity, that are computed based on classifier predictions (correct or incorrect for the class label) of base classifiers [36]. To assess the diversity of the proposed individual-specific ensembles, kappa (k) is calculated as a widely used diversity measure that is related to Kohavi–Wolpert variance and disagreement measures. The value of k ranges from -1 to 1, where its lower values show greater diversity. The positive values indicate that the classifiers tend to classify the same object correctly, whereas the negative values correspond to negative correlation [41].

5.4. Computational complexity

In practical video surveillance applications, FR systems must be computationally efficient, and scale well to a growing number of cameras, watch-list individuals, and clutter in the scene. The generation of e-SVM classifiers comprised of training e-SVMs, ranking patches and subspaces, as well as, pruning the e-SVMs were performed off-line. Since e-SVMs trained for different patches, descriptors, and random subspaces are generated and ranked independently from one another, they can be processed in parallel. Computational complexity of the proposed system is therefore relevant to the operational phase, and affected by the feature extraction techniques, classification process, and dynamic selection and weighting of each input ROI probe with the size of nxn.

Extraction of face descriptors using HOG and LPQ is related to their transformation functions, where their complexities are $O(n)$ and $O(nlogn)$, respectively [58,59]. Classification has been performed using e-SVM which employs a linear SVM kernel function using a dot product with the complexity of $O(N_d \cdot N_{sv})$ [60], where N_d and N_{sv} are the average dimensionality of the face descriptors and the average number of support vectors, respectively. Finally, dynamic selection and weighting is based on Cityblock distance which is a linear distance metric, therefore, this process requires $O(N_d \cdot N_c \cdot N_{sv})$ computations, where N_c is the total number of classifiers in the pool.

Memory complexity of the proposed system mainly depends on the number of watch-list persons N_{wl} and size of the pool. Thus, complexity of the pool (number of classifiers N_c) for each individual of interest can be considered as $O(N_p \cdot N_{fd} \cdot N_{rs})$, where N_p is the number of patches, N_{fd} and N_{rs} are the number of face descriptors and the average number of random subspaces, respectively. Hence, the overall memory complexity can be computed as $O(N_{wl} \cdot N_p \cdot N_{fd} \cdot N_{rs} \cdot N_d)$. More specifically, the worst case of computational complexity of the proposed individual-specific Ee-SVMs in the operational mode to process an input ROI pattern can be formulated as $N_p \cdot N_{fd} \cdot N_{rs} \cdot N_{sv} \cdot N_d$ according to the dot products required by each e-SVM classifier.

6. Results and discussion

6.1. Number and size of feature subspaces

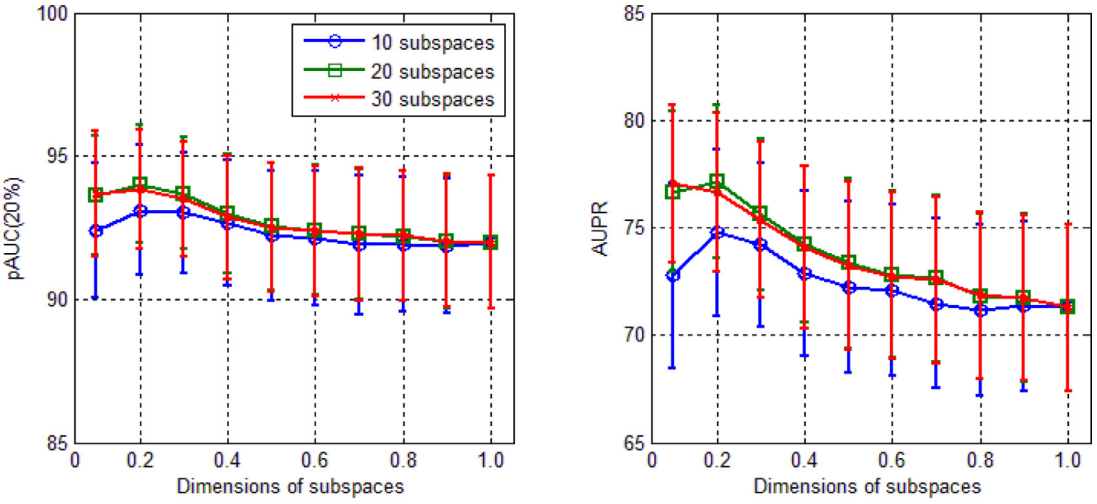
The critical parameters of the proposed system need to be defined precisely to select the best values using the generic pool. The impact of different numbers and dimensions of feature subspaces are statistically analyzed for each face descriptors extracted from each patch using a validation set during the design phase. In this analysis, different numbers of subspaces (N_{rs}) are considered w.r.t. different proportions of feature dimensions (N_d). In this section, experiments were conducted with a generic pool that uses RSM to generate individual-specific Ee-SVMs combined through score averaging based on the third training scheme. The transaction-level analysis (pAUC(20%) and AUPR with standard errors) of different numbers and dimensions of subspaces for HOG and LPQ are depicted in Fig. 7.

Fig. 7(a) implies that performance obtained using 20% of features is slightly higher than other dimensions in term of both pAUC(20%) and AUPR for HOG descriptor. Results suggest that it is better to select the 20% of original feature space as a dimensions of HOG descriptor (39 features). In addition, 20 random subspaces as the number of subspaces achieves the highest performance.

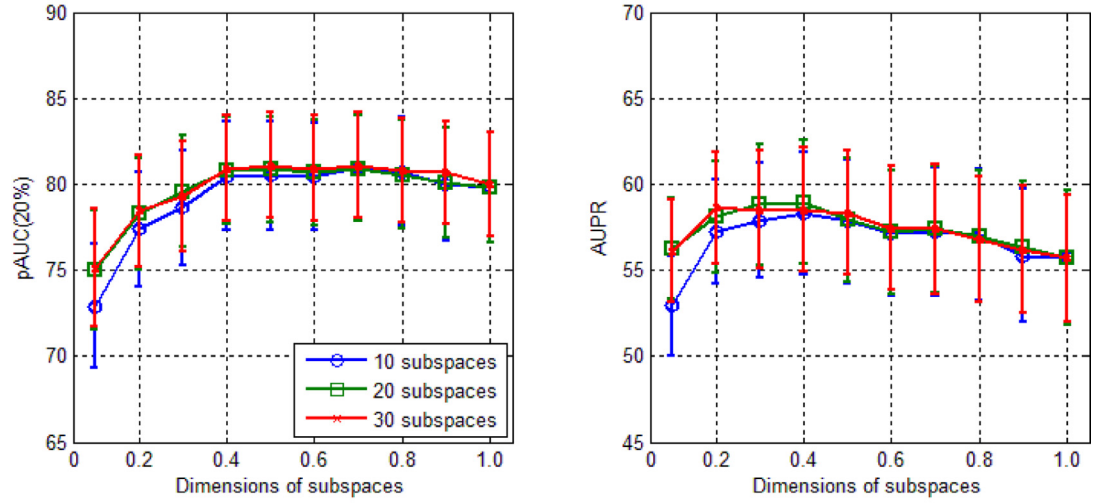
As shown in Fig. 7(b), 40% of the LPQ descriptor can be a suitable value as dimension of LPQ subspaces. Moreover, the best number of subspaces can be defined as 20 subspaces. It can be seen that performance of the system is not greatly affected by the numbers and dimensions of feature subspaces, where either pAUC(20%) or AUPR first raise and then stabilize. This suggests that increasing the number of subspaces may transfer more diversity among classifiers in the pool, but it cannot improve the accuracy. Noted that, performance is stabilized for the values higher than 20 subspaces. Hence, it can be concluded that the proposed system is not highly sensitive to the number of subspaces (see Fig. 7(a)).

Another experiment that was performed prior to design is to rank patches using the validation set D. The sensitivity analysis on the performance of using each patch separately in order to rank them based on their importance is illustrated in Fig. 8.

As shown in Fig. 8, each patch performs differently from other patches for each descriptor. Selecting a different number of semi-random subspaces from each patch based on its importance for overall performance therefore can lead to a robust system.



(a) HOG face descriptor



(b) LPQ face descriptor

Fig. 7. The impact of different numbers and size of feature subspaces on performance of using HOG and LPQ face descriptor.

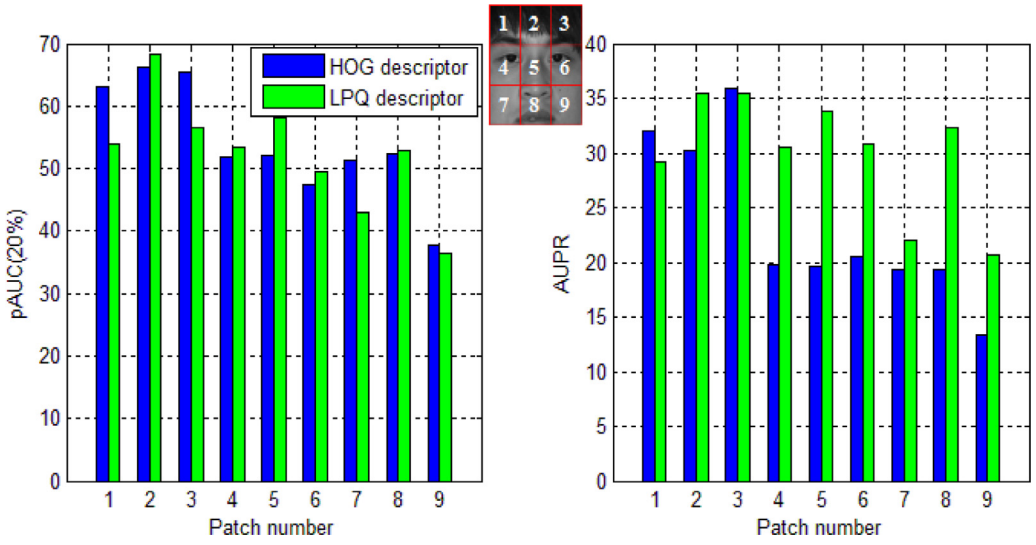


Fig. 8. The analysis of system performance based on each patch over COX-S2V.

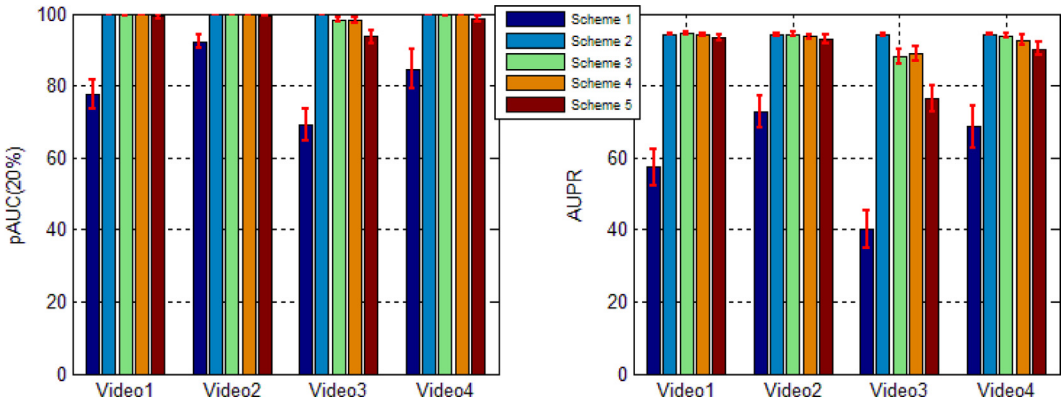


Fig. 9. Average pAUC(20%) and AUPR transaction-level performance of different training schemes at with COX-S2V.

6.2. Training schemes

Fig. 9 presents the average transaction-level performance of using the generic pool for different training schemes as described in Section 4.2.2 over each video of COX-S2V. Results were produced using a generic pool of 360 e-SVMs (9 patches × 2 descriptors × 20 subspaces) per each target individual.

Results in Fig. 9 indicate that training schemes 2, 3, 4, and 5 greatly outperform scheme 1, due to DA using knowledge transferred from all of the surveillance cameras in the target domain. The results also suggest that exploiting a few non-target stills from the source domain during training e-SVMs in the third scheme can provide slight improvements, especially in AUPR values according to video1, video2, and video4 comparing to the second scheme [8]. Knowledge of the ED is therefore incorporated in the third scheme due to combination of feature representations across domains using a mixture of labeled still ROIs from the ED and unlabeled calibration videos from the OD [12].

Camera-specific training schemes 4 and 5 provide higher performance in comparison to scheme 1, where they also exploit knowledge of the operational domain. However, they are also outperformed by schemes 2 and 3 in terms of both accuracy and complexity, because videos from all of the cameras are considered in schemes 2 and 3 to generate a general pool, while several camera-specific pools must be generated in the schemes 4 and 5 using videos of each specific camera. Meanwhile, scheme 4 performs slightly better than scheme 5, because all of the video ROIs captured from a specific camera FoV have the same pose and angle, while adding frontal stills with significant differences in data distributions may subsequently degrade the training performance. Noted that only the classifiers from pool #1 trained using camera #1 is employed to classify the probe captured using camera #1 during operational phase.

Therefore, other experiments on the proposed system are accomplished using the third training scheme. Since the characteristics of capturing devices are different, it has a significant impact on the system performance according to each video. The differences between pAUC(20%) and AUPR observed in Fig. 9 reveal that the large number of e-SVMs classify the non-target ROIs as non-targets, but only some of them classify the target ROIs correctly. Therefore, the FPR values are very low in the all cases.

Another test that can be also used in order to assess the performance of the training schemes is the Friedman test with a post-hoc test, where it is basically incorporated to find a significance difference between several methods according to their ranks averaged across datasets. The Friedman test is typically followed by a post-hoc test, such as Nemenyi test to indicate whether the difference in ranks is above a critical distance (CD) [63]. Fig. 10 shows

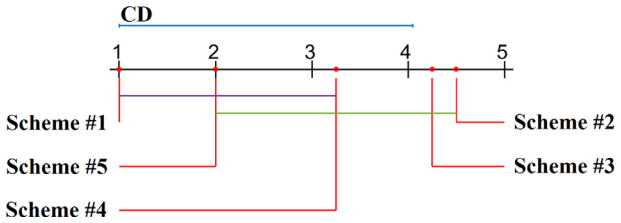


Fig. 10. Training schemes by significant differences according to the post-hoc Nemenyi test over COX-S2V.

the results of Nemenyi's post-hoc test, where the schemes linked by colored lines are not significantly different by the test for a significance level of $\rho = 0067$.

Fig. 10 demonstrates with a more visual insight as differences of the training schemes, where the lowest average rank is associated to the worst training scheme and vice versa. According to this test, schemes that exploit DA are significantly different than scheme #1, meaning that training through DA provides significantly higher performance than the training without considering DA.

6.3. Number of training and testing stills and trajectories

The impact of employing different number of non-target videos from the background model (videos of non-target persons), as well as, different number of non-target stills from the cohort model (stills of non-target persons) on the performance is illustrated in Fig. 11. In this regard, the third training scheme is employed considering the first $N_{wl} = 10$ persons of COX-S2V as watch-list individuals. The number of low-quality videos of non-target persons N_{ntd} considered for training during the design phase is varied from 10 to 100 according to the number of non-target stills belonging to other persons in the cohort.

As shown in Fig. 11, growing the number of non-target persons participating in the design phase can slightly improve the performance. Since it may be costly and impractical to employ plenty of training data in the real-world application, the proposed system provides convincing results even with limited non-target video data. Thus, knowledge of the target domain can be appropriately transferred by considering the limited number of non-target video data.

Fig. 11 also demonstrates that growing the number of high-quality non-target stills during training degrades the performance significantly. Since these still ROIs are close to the still of the target individual, most of the support vectors are selected from them and subsequently, these classifiers could not successfully classify the low-quality input probes. Hence, the larger the number of non-target stills, the higher the number of inappropriate support vec-

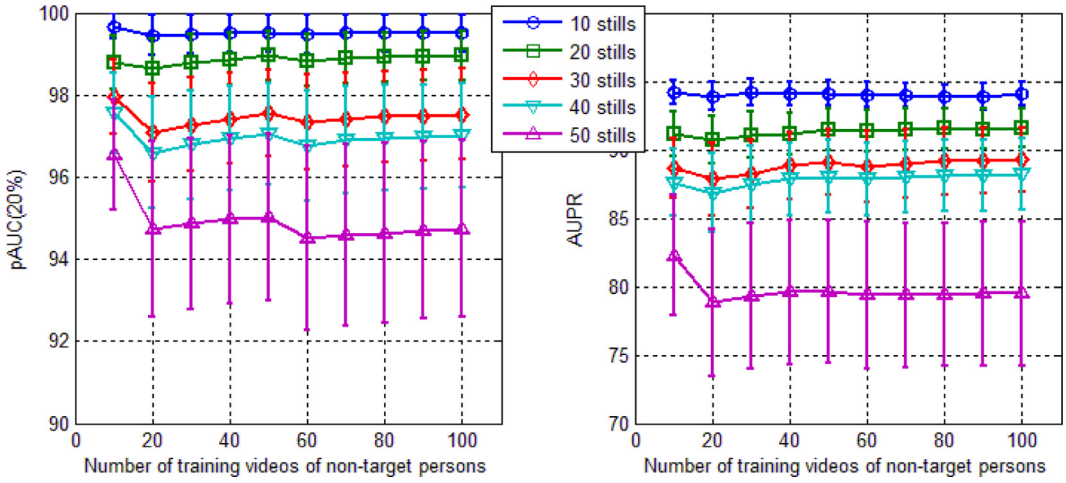


Fig. 11. The analysis of system performance using different number of training non-target persons over COX-S2V.

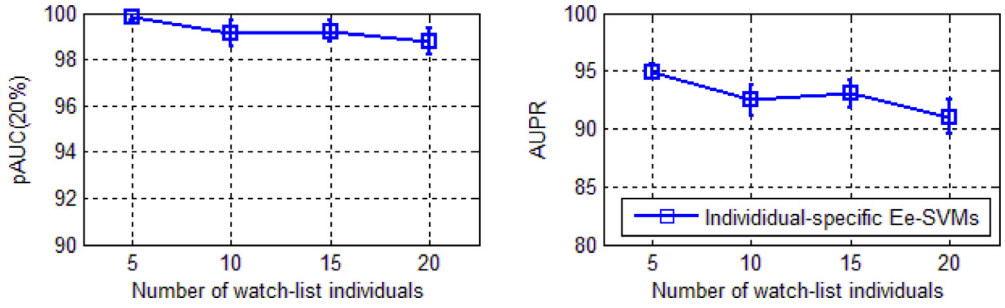


Fig. 12. The analysis of Ee-SVMs performance using different number of watch-list persons during operations over COX-S2V.

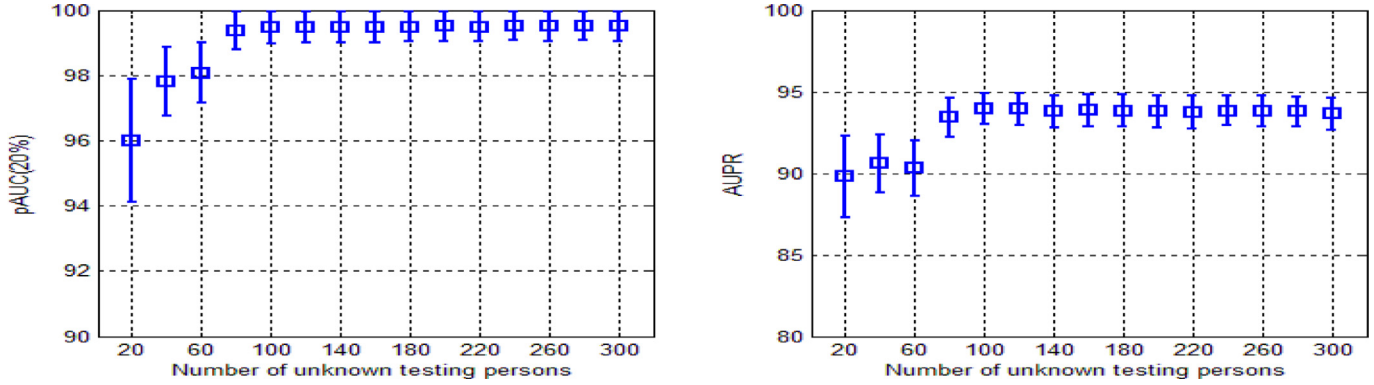


Fig. 13. The analysis of system performance using different number of unknown persons during operations over COX-S2V.

tors, and therefore the capability of classifiers reduces to classify the given probe as if employing a lower number of non-target stills. Nevertheless, employing the lower number of stills from the cohort along with videos of non-target persons provides higher classification performance as shown in Fig. 9.

To analyze the performance considering different number of watch-list individuals enrolled to the system, N_{wl} is varied from 5 to 20 as illustrated in Fig. 12.

Fig. 12 shows that enlarging the list of watch-list persons does not have a significant impact on the system performance. Since the proposed system is comprised of individual-specific ensembles, and each one seeks to detect one watch-list individual at a time, there should not be significant differences in increasing the number of watch-list persons.

The impact of considering different number of non-target videos of unknown persons from the test set on performance is displayed in Fig. 13. In this regard, the number of unknown persons N_{ntu} appearing in the surveillance environment along with the target person during the operational phase is altered to see its influence on the system performance.

As illustrated in Fig. 13, the number of unknown persons participating in the operational phase is varied from 20 to 300 persons. Since the FP values for each threshold in the ROC and inverted precision-recall curves increase slower than the total number of negatives, then the FPR values decrease slightly and it subsequently leads to a higher values of area under ROC and precision-recall curves. It can be concluded that the proposed system can perform well even with severely imbalanced data according to observation of many unknown persons during operations.

To obtain the transaction-level performance of the proposed system using pAUC, values of FPR are varied from 5% to 100% as demonstrated in Fig. 14.

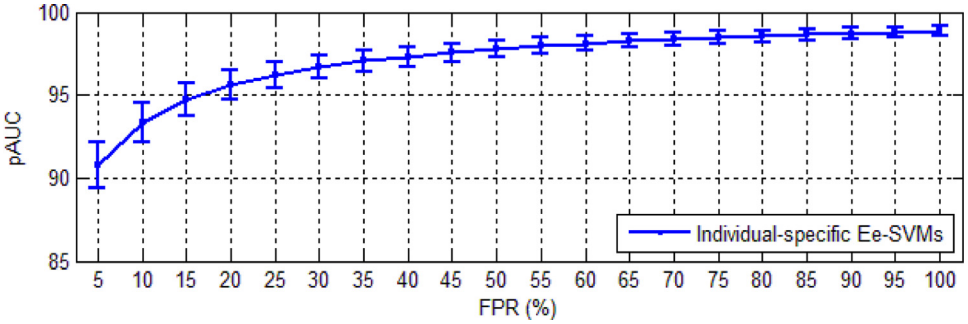


Fig. 14. The analysis of system performance using different number of unknown persons during operations over COX-S2V.

Table 1
Average pAUC(20%) and AUPR performance of the system with generic pool and different design scenarios at transaction-level over COX-S2V.

Systems	Video 1		Video 2		Video 3		Video 4		Complexity
	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	(# dot products)
Generic pool	99.19 ± 0.44	93.18 ± 0.88	99.43 ± 0.16	91.39 ± 0.82	92.01 ± 1.11	70.95 ± 2.20	96.08 ± 1.09	84.89 ± 2.08	460,080
Scenario 1	99.97 ± 0.03	94.86 ± 0.18	99.40 ± 0.22	92.60 ± 0.78	97.77 ± 0.52	87.23 ± 1.13	93.12 ± 0.90	81.18 ± 0.87	127,800
Scenario 2	99.08 ± 0.40	92.64 ± 0.69	99.32 ± 0.17	90.44 ± 1.01	91.02 ± 1.28	68.54 ± 2.36	96.21 ± 1.02	84.37 ± 2.11	230,040

Table 2
Average pAUC(20%) and AUPR performance of the system with generic pool and different design scenarios at transaction-level over Chokepoint.

Systems	Session 1		Session 2		Session 3		Session 4		Complexity
	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	(# dot products)
Generic pool	97.67 ± 0.92	96.63 ± 1.21	96.93 ± 1.43	95.33 ± 2.21	100 ± 0.00	99.64 ± 0.07	75.33 ± 6.04	71.85 ± 6.66	460,080
Scenario 1	99.74 ± 0.12	99.25 ± 0.25	99.99 ± 0.01	99.81 ± 0.01	100 ± 0.00	99.74 ± 0.05	91.81 ± 0.92	90.56 ± 1.16	127,800
Scenario 2	98.81 ± 0.49	98.15 ± 0.56	98.07 ± 0.82	96.89 ± 1.36	100 ± 0.00	99.74 ± 0.08	77.00 ± 5.59	73.52 ± 6.37	230,040

As shown in Fig. 14, increasing the FPR thresholds can slightly achieve higher AUC, while the real-world watch-list screening systems must perform on a certain operating point that has been considered as FPR=20% in this paper. Thus, the rate of false positives must be limited by considering an appropriate operating point w.r.t. the application.

6.4. Design scenarios

Performance of the proposed system in terms of considering different design scenarios is presented in Tables 1 and 2 using the third training scheme over videos of COX-S2V and Chokepoint, respectively.

The results in Table 1 indicate that generating a compact pool of classifiers based on the first design scenario can yield higher performance, where the baseline performance is obtained using the generic pool. Hence, pre-selection of e-SVMs by ranking patches and subspaces achieves better performance with a lower computational complexity. Moreover, system with a compact pool generated according to the second design scenario cannot improve the performance effectively, since no priori knowledge is taken into account and all system choices are performed during the enrollment phase. Consequently, generating a compact pool according to the first design scenario using the criteria based on overall AUPR through a validation set is more accurate and efficient.

Table 2 also confirms that the results obtained with the first design scenario are higher than generic pool and compact pool generated according to the second design scenario among all the sessions. On the other hand, performance of the system with the second design scenario is slightly better than the baseline using the generic pool.

It is worth pointing out that, the number of classifiers in the generic pool for each individual of interest is 360 ($9 \cdot 2 \cdot 20$), while each target individual has about 100 and 180 classifiers in the compact pool of first and second scenarios. Meanwhile, the average number of support vectors for each classifier and the dimension of each feature vector are 18 and 71, respectively. Thus, the time complexity as described in Section 5.4 for generic pool is about 460,080 ($360 \cdot 18 \cdot 71$) dot products for processing a given probe ROI, while the compact pool based on the first and second scenarios requires around 127,800 and 230,040 computations, respectively. Hence, the proposed system based on the first scenario is effective in terms of either accuracy or computational complexity.

Furthermore, the impact of different numbers of ranked subspaces in the system with a pool generated based on the second design scenario is shown in Fig. 15. In this scenario, over-produce and select paradigm is considered, where 50 subspaces are generated for each patch and then they are ranked using the local distance-based criteria (see Section 4.3).

As shown in Fig. 15, both systems perform equally in terms of pAUC(20%) values, while the system designed with the second scenario outperforms the generic pool specifically for the first 10 ranked subspaces. Moreover, the pAUC performance is stable starting from $N'_{rs} = 10$ subspaces. The system with the generic pool performs better in terms of AUPR values. It can be concluded that the local criteria exploited to select the best subspaces in the second design scenario cannot be a desired metric consistently in contrast to the global criteria utilized in the first design scenario.

The diversity among classifiers within each individual-specific ensemble is computed using kappa (k) diversity measure for $N_{wl} = 20$ individuals with 5 replications. The value of k is 0.0065 ± 0.0005, where it can be concluded that the classifiers within the ensembles are relatively diverse.

6.5. Dynamic selection and weighting

The performance of applying dynamic selection and weighting approaches on the proposed system with generic pool, the first

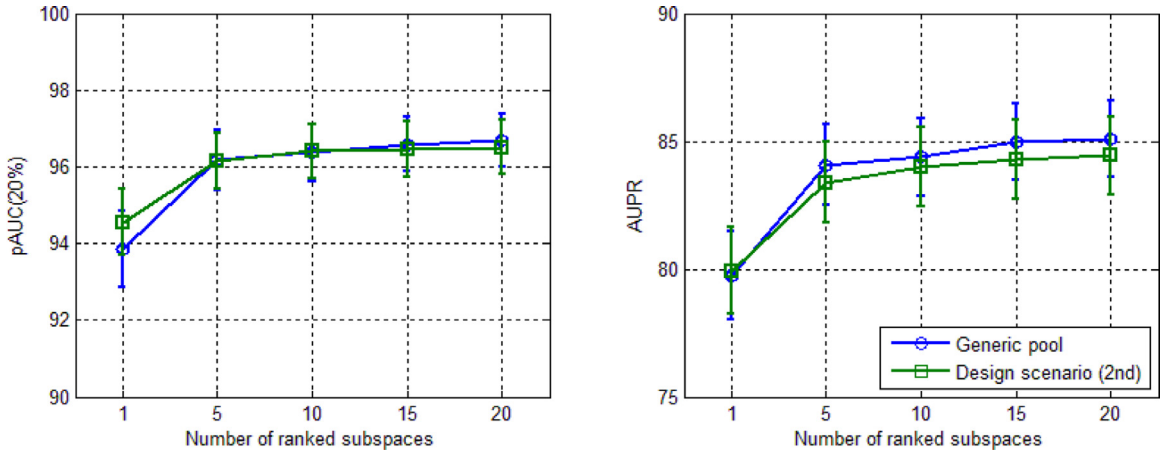


Fig. 15. The analysis of system performance using different numbers of ranked subspaces based on the second design scenario of compact pool generation over COX-S2V.

Table 3
Average pAUC(20%) and AUPR performance at transaction- and trajectory-level after applying dynamic selection and weighting on the system with generic pool and different design scenarios over COX-S2V.

Systems	Transaction-level		Trajectory-level	Complexity
	pAUC	AUPR	AUC	(# dot products)
Generic pool	96.68 ± 0.70	85.10 ± 1.49	99.72 ± 0.05	$(9\cdot 2\cdot 20\cdot 18\cdot 71) = 460{,}080$
Generic pool with DS	98.21 ± 0.45	86.40 ± 1.17	99.93 ± 0.04	$(9\cdot 2\cdot 20\cdot 18\cdot 71) + (9\cdot 2\cdot 20\cdot 2\cdot 71) = 511{,}200$
Generic pool with DW	97.52 ± 0.59	87.27 ± 1.38	99.91 ± 0.04	$(9\cdot 2\cdot 20\cdot 18\cdot 71) + (9\cdot 2\cdot 20\cdot 2\cdot 71) = 511{,}200$
Generic pool with DS and DW	96.89 ± 0.64	85.39 ± 1.47	99.90 ± 0.05	$(9\cdot 2\cdot 20\cdot 18\cdot 71) + 2\cdot(9\cdot 2\cdot 20\cdot 2\cdot 71) = 562{,}320$
Scenario 1 with DS	93.47 ± 0.76	77.32 ± 1.66	99.52 ± 0.14	$(100\cdot 18\cdot 71) + (100\cdot 2\cdot 71) = 142{,}000$
Scenario 1 with DW	98.11 ± 0.49	88.60 ± 1.24	99.93 ± 0.05	$(100\cdot 18\cdot 71) + (100\cdot 2\cdot 71) = 142{,}000$
Scenario 1 with DS and DW	95.60 ± 0.72	84.08 ± 1.39	99.77 ± 0.10	$(100\cdot 18\cdot 71) + 2\cdot(100\cdot 2\cdot 71) = 156{,}200$
Scenario 2 with DS	98.02 ± 0.47	86.14 ± 1.25	99.87 ± 0.07	$(9\cdot 2\cdot 10\cdot 18\cdot 71) + (9\cdot 2\cdot 10\cdot 2\cdot 71) = 255{,}600$
Scenario 2 with DW	97.38 ± 0.82	87.36 ± 1.84	99.89 ± 0.05	$(9\cdot 2\cdot 10\cdot 18\cdot 71) + (9\cdot 2\cdot 10\cdot 2\cdot 71) = 255{,}600$
Scenario 2 with DS and DW	96.37 ± 0.98	85.12 ± 1.88	99.76 ± 0.08	$(9\cdot 2\cdot 10\cdot 18\cdot 71) + 2\cdot(9\cdot 2\cdot 10\cdot 2\cdot 71) = 281{,}160$

design scenario (compact pool), and the second design scenario are demonstrated in Table 3 using the third training scheme at transaction- and trajectory-level along with the time complexity.

Table 3 indicates that applying proposed DS method can improve the performance in contrast to combining all of classifiers in the system with generic pool and the second design scenario. It implies that combining a subset of competent classifiers leads to a system with higher accuracy. In addition, the proposed DW approach performs better in comparison with DS in terms of AUPR, where only two distances (distance to the target still and distance to the closest non-target support vector) are measured in the both selection and weighting strategies. Moreover, applying the proposed DW approach on the scores of classifiers selected dynamically cannot achieve a better performance, due to elimination of classifiers.

As observed in Table 3, DW can also magnify performance of the system with the first design scenario slightly, while applying the dynamic selection approach deteriorates its performance. Since a pre-selection scenario was already applied to the compact pool, applying DS can diminish the ensemble diversity. It can be concluded that using the compact pool and weighting the classifiers dynamically achieves the highest level of performance considering the AUPR values.

The trajectory-level performance of the proposed systems with DS and DW are also presented in Table 3 as the result of spatio-temporal FR. Thus, scores of individual-specific ensembles are gradually accumulated over a window of $W = 10$ consecutive frames using a trajectory defined by the tracker. To assess the overall performance, the corresponding ROC curve can be then plotted for each individual of interest by varying the thresholds from 0 to 10 (size of the window) over the accumulated scores, and the AUC are computed as overall performance.

As shown in Table 3, spatio-temporal recognition applied on the proposed systems leads to a near perfect face screening system. An example of accumulated scores over the generated trajectory is shown in Fig. 16 using the systems with the best AUPR values. Video1 of COX-S2V is thus employed in this example, where individual ID#001 is considered as the watch-list target individual along with $N_{ntu} = 100$ unknown non-target individuals.

As shown in Fig. 16, the accumulated scores for target individual (ID#001) is significantly higher than all non-targets individuals. It can be observed that the accumulated scores of some non-target individuals are high, due to appearance similarity to the target individual. The proposed system based on the first scenario with DW performs more reliable in trajectory-level, where it provides higher accumulated scores for the target, and simultaneously lower accumulated scores for non-target individuals.

Table 4 presents the complexity in terms of the number of dot products required during operations to process a probe ROI. The proposed selection and weighting approaches are desirable for the screening application in terms of operational time complexity. On the other hand, the distance measures can influence on the computational time based on their complexity. However, the CityBlock distance measure can be a suitable candidate due to its efficiency and linear computability. For example, the proposed system with DW over COX-S2V data needs $9\cdot 2\cdot 20\cdot 18\cdot 71$ dot products for fusion in the worst case, where all of the classifiers are dynamically selected, and $9\cdot 2\cdot 20\cdot 2\cdot 71$ for selection. It is worth pointing out that the average number of support vectors N_{sv} (the fourth element in the complexity formulation) for COX-S2V and Chokepoint

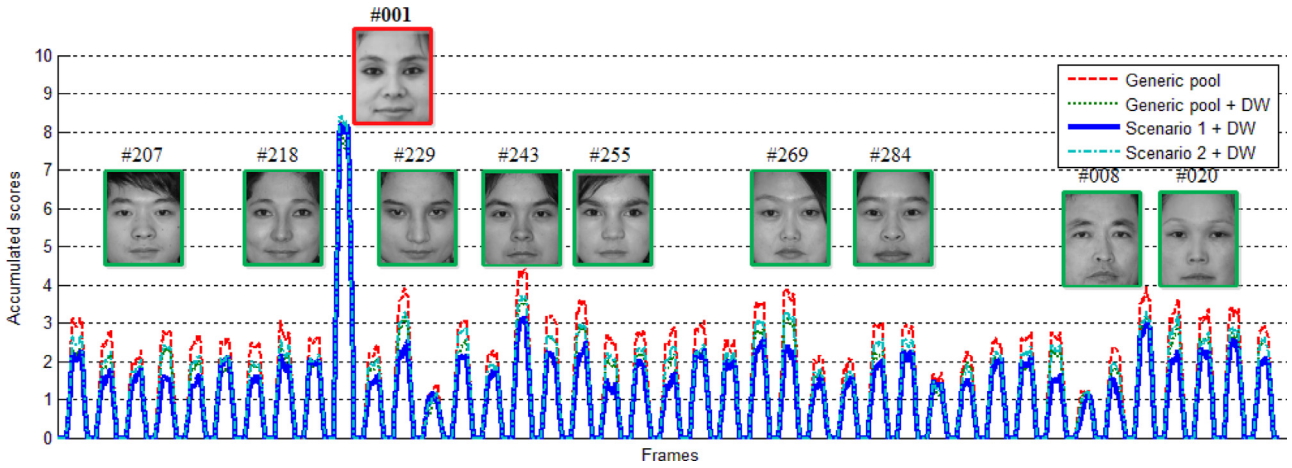


Fig. 16. An example of the scores accumulated over windows of 10 frames over video1 of COX-S2V.

Table 4
Average pAUC(20%) and AUPR performance and time complexity of the proposed system at transaction-level over COX-S2V and Chokepoint videos against the state-of-the-art systems.

Systems	COX-S2V			Chokepoint		
	pAUC	AUPR	Complexity	pAUC	AUPR	Complexity
VSkNN [5]	56.80 ± 4.02	26.68 ± 3.58	671,744	19.00 ± 0.40	16.48 ± 0.90	671.744
SVDL [61]	69.93 ± 5.67	44.09 ± 6.29	810,000	74.91 ± 4.03	65.09 ± 4.82	810,000
ESRC-DA [31]	99.00 ± 1.13	63.21 ± 4.56	228,614,400	97.16 ± 1.28	76.97 ± 6.73	432,224,100
Ensemble of TMs [6]	84.00 ± 0.86	73.36 ± 9.82	1,387,200	85.60 ± 1.04	82.78 ± 7.06	1,387,200
Ensemble of e-SVMs [8]	99.02 ± 0.15	88.03 ± 0.85	2,281,472	100 ± 0.00	99.24 ± 0.38	2,235,392
Scenario 1 with DW	98.11 ± 0.49	88.60 ± 1.24	142,200	97.52 ± 0.50	96.86 ± 0.72	113,600
Scenario 2 with DW	97.38 ± 0.82	87.36 ± 1.84	255,600	93.36 ± 1.97	91.79 ± 2.45	204,480

data are not the same (18 and 14 support vectors, respectively), so that the computational complexity over these datasets is different.

Results are compared with the state-of-the-art and baseline systems in Table 4 according to the average transaction-level performance over the COX-S2V and Chokepoint data.

It can be seen from Table 4 that ensemble of e-SVMs significantly outperforms ensemble of TMs, VSkNN, SVDL, and ESRC-DA. Performance of the screening system using VSkNN and SVDL is poor, mostly because of the notable differences between quality and appearances of the target face stills in the gallery set and video faces in the generic training set, as well as, severely data imbalance of target versus non-target individuals observed during operations. It is worth noting that both VSkNN and SVDL are more suitable for close-set FR problems, such as face identification. Since each faces captured should be assigned to one of the target still in the gallery, therefore, many false positive occur. Moreover, SVDL can only apply as a complex global N-class classifier in contrast to the proposed ensemble of SVMs, due to sparse optimization and classification during the operational phase. However, sparsity concentration index [64] is used as a rejection threshold to reject the probes not appearing in the training.

The results observed from Table 4 confirm that the proposed system using the first design scenario along with DW approach is efficient and can achieve an equivalent performance comparing to Bashbaghi et al. [8] with a significant decrease in computational complexity. In addition, the system design with the second scenario and DW can perform almost equivalent to state-of-the-art systems performance. However, the systems proposed in this paper employ two different face descriptors, whereas ensemble of e-SVMs utilizes four different face descriptors along with PCA with $O(N_d^3)$ for feature selection. Meanwhile, ensemble of TMs and VSkNN employ Euclidean distance with $O(N_d^2)$ to calculate the similarity among templates, therefore, they need more computations.

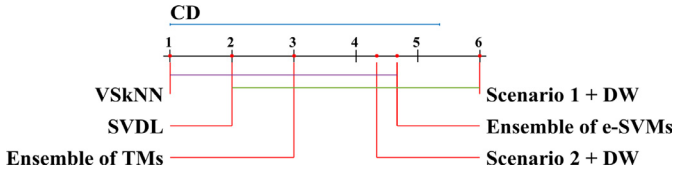


Fig. 17. State-of-the-art systems by rank and significant differences according to the post-hoc Nemenyi test over COX-S2V.

The proposed system is also validated using Chokepoint dataset, where the results observed from Table 4 confirm that the proposed system can achieve promising performance compare to state-of-the-art systems with a significantly lower computational complexity.

A Friedman test is also conducted on the comparison of the proposed systems against state-of-the-art and found significant with a significance level of ρ-value $\rho = 0.012$. The results of the Nemenyi post-hoc test is shown in Fig. 17. These systems are ranked in an ascendant order, where the highest average rank is assigned to the best system. It indicates that the other four systems (ranked 1 to 5) are not significantly different, while the proposed system using design scenario 1 with DW is slightly different than the others and above the critical distance.

7. Conclusion

In this paper, a robust MCS is proposed for still-to-video FR that is specialized for watch-list screening applications, where individual-specific Ee-SVMs are designed to model a single reference still of target individuals. A novel ensemble-based learning is

utilized, where multiple random subspaces are generated for different face descriptors extracted from face patches to effectively provide ensemble diversity and address the SSPP constraints. Unlike conventional RSM that completely select the feature subspaces randomly from the entire ROI, semi-random subspaces are employed to either consider the distribution of face descriptors and to make use of the local spatial relation among each patch. Furthermore, an unsupervised DA method is used to train e-SVM classifiers in the ED through several training schemes, where video ROIs of non-target individuals are exploited versus a single still ROI to transfer knowledge from the ODs. This paper also investigates the impact of using different training schemes for DA, as well as, the validation set of non-target faces extracted from stills and video trajectories of unknown individuals in the OD. Thus, such a system can incorporate knowledge of the ODs and improve the robustness against several nuisance factors frequently observed in video surveillance operational environments.

During enrollment of a target individual, a pool of diverse classifiers is generated through two design scenarios to select the most representative subspaces. The first scenario exploits additional knowledge acquired from a validation set and a global criterion, while the second scenario employs a local criterion. Hence, the best subspaces are selected using the first scenario and it can construct an efficient system with a compact pool. In addition, distance-based dynamic selection and weighting approaches are also proposed based on the SSPP issue to either select or weight the classifiers dynamically during operations. Since there is no other target still during design, different internal criteria are defined using distances of the input probe from the support vectors of e-SVMs and the reference still in the feature space. The final output of ensembles is obtained using two levels of fusion among the classification scores of patches and eventually scores from face descriptors to perform spatio-temporal recognition.

Extensive evidences are provided using the COX-S2V and Chokepoint datasets that the proposed method is effective and comparable against the state-of-the-art methods. The proposed ensemble-based system was validated under extremely imbalanced screening situation. Experimental results indicate that integration of the ranked semi-random subspaces into an individual-specific Ee-SVMs, in the construction of a compact and diverse pool demonstrates a higher level of performance than using a generic pool without pruning. Although the proposed dynamic selection and weighting approaches generate better performance in terms of accuracy, but they impose some overhead and computational burden to such a time-bounded application. However, the proposed system with a compact pool of e-SVMs and dynamic weighting can achieve state-of-the-art performance with a significantly lower computational complexity.

Acknowledgment

This work was supported by the Fonds de Recherche du Québec - Nature et Technologies.

References

[1] F. Matta, J.-L. Dugelay, Person recognition using facial video information: A state of the art, J. Visual Lang. Comput. 20 (3) (2009) 180–187.
[2] M.D. la Torre, E. Granger, P.V. Radtke, R. Sabourin, D.O. Gorodnichy, Partially-supervised learning from facial trajectories for face recognition in video surveillance, Inf. Fusion 24 (2015) 31–53.
[3] M.A.A. Dewan, E. Granger, G.-L. Marcialis, R. Sabourin, F. Roli, Adaptive appearance model tracking for still-to-video face recognition, Pattern Recognit. 49 (2016) 129–151.
[4] M.D. la Torre, E. Granger, R. Sabourin, D.O. Gorodnichy, Adaptive skew-sensitive ensembles for face recognition in video surveillance, Pattern Recognit. 48 (11) (2015) 3385–3406.
[5] C. Pagano, E. Granger, R. Sabourin, G. Marcialis, F. Roli, Adaptive ensembles for face recognition in changing video surveillance environments, Inf. Sci. 286 (2014) 75–101.
[6] S. Bashbaghi, E. Granger, R. Sabourin, G.-A. Bilodeau, Watch-list screening using ensembles based on multiple face representations, ICPR, 2014.
[7] R. Chellappa, M. Du, P. Turaga, S.K. Zhou, Face tracking and recognition in video, in: S.Z. Li, A.K. Jain (Eds.), Handbook of Face Recognition, Springer London, 2011, pp. 323–351.
[8] S. Bashbaghi, E. Granger, R. Sabourin, G.-A. Bilodeau, Ensembles of exemplar-svms for video face recognition from a single sample per person, AVSS, 2015.
[9] M. Kan, S. Shan, Y. Su, D. Xu, X. Chen, Adaptive discriminant learning for face recognition, Pattern Recognit. 46 (9) (2013) 2497–2509.
[10] F. Mokhayeri, E. Granger, G.-A. Bilodeau, Synthetic face generation under various operational conditions in video surveillance, ICIP, 2015.
[11] V. Patel, R. Gopalan, R. Li, R. Chellappa, Visual domain adaptation: A survey of recent advances, IEEE Signal Process. Mag. 32 (3) (2015) 53–69.
[12] S.J. Pan, Q. Yang, A survey on transfer learning, KDE, IEEE Trans. 22 (10) (2010) 1345–1359.
[13] S. Shekhar, V. Patel, H. Nguyen, R. Chellappa, Generalized domain-adaptive dictionaries, CVPR, 2013.
[14] R.M. Cruz, R. Sabourin, G.D. Cavalcanti, T.I. Ren, Meta-des: A dynamic ensemble selection framework using meta-learning, Pattern Recognit. 48 (5) (2015) 1925–1935.
[15] T. Gao, D. Koller, Active classification based on value of classifier, in: Advances in Neural Information Processing Systems 24, Curran Associates, Inc., 2011, pp. 1062–1070.
[16] P. Matikainen, R. Sukthankar, M. Hebert, Classifier ensemble recommendation, ECCV, Workshops and Demonstrations, Springer Berlin Heidelberg, 2012.
[17] P.R. Cavalin, R. Sabourin, C.Y. Suen, Logid: An adaptive framework combining local and global incremental learning for dynamic selection of ensembles of hmms, Pattern Recognit. 45 (9) (2012) 3544–3556.
[18] P. Cavalin, R. Sabourin, C. Suen, Dynamic selection approaches for multiple classifier systems, Neural Comput. Appl. 22 (3-4) (2013) 673–688.
[19] Z. Huang, S. Shan, H. Zhang, S. Lao, A. Kuerban, X. Chen, Benchmarking still–to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset, in: ACCV, Springer, 2013, pp. 589–600.
[20] Y. Wong, S. Chen, S. Mau, C. Sanderson, B.C. Lovell, Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition, CVPR, Biometrics Workshop, 2011.
[21] X. Chen, C. Wang, B. Xiao, C. Zhang, Still-to-video face recognition via weighted scenario oriented discriminant analysis, IJCB, 2014.
[22] H. Wang, C. Liu, X. Ding, Still-to-video face recognition in unconstrained environments, in: Proc. SPIE, Image Processing: Machine Vision Applications, 2015.
[23] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, X. Chen, A benchmark and comparative study of video-based face recognition on cox face database, IP, IEEE Trans. 24 (12) (2015) 5967–5981.
[24] Y. Zhu, Y. Li, G. Mu, S. Shan, G. Guo, Still-to-video face matching using multiple geodesic flows, Inf. Forensics Secur. IEEE Trans. 11 (12) (2016) 2866–2875.
[25] S. Bashbaghi, E. Granger, R. Sabourin, G.-A. Bilodeau, Robust watch-list screening using dynamic ensembles of svms based on multiple face representations, Mach. Vision Appl. 28 (1) (2017) 219–241.
[26] S. Liao, A. Jain, S. Li, Partial face recognition: Alignment-free approach, PAMI, IEEE Trans. 35 (5) (2013) 1193–1205.
[27] B. Kamgar-Parsi, W. Lawson, B. Kamgar-Parsi, Toward development of a face recognition system for watchlist surveillance, PAMI, IEEE Trans. 33 (10) (2011) 1925–1937.
[28] Q. Qiu, J. Ni, R. Chellappa, Dictionary-based domain adaptation for the re-identification of faces, in: Person Re-Identification, Advances in Computer Vision and Pattern Recognition, 2014, pp. 269–285.
[29] A. Ma, J. Li, P. Yuen, P. Li, Cross-domain person reidentification using domain adaptation ranking svms, IP, IEEE Trans. 24 (5) (2015) 1599–1613.
[30] J. Hu, Discriminative transfer learning with sparsity regularization for single-sample face recognition, Image Vision Comput. (2016).
[31] F. Nourbakhsh, E. Granger, G. Fumera, An extended sparse classification framework for domain adaptation in video surveillance, ACCV, Workshop on Human Identification for Surveillance, 2016.
[32] L. Duan, D. Xu, I. Tsang, Domain adaptation from multiple sources: A domain-dependent regularization approach, Neural Netw. Learn. Syst. IEEE Trans. 23 (3) (2012) 504–518.
[33] H. Bhatt, R. Singh, M. Vatsa, N. Ratha, Improving cross-resolution face matching using ensemble-based co-transfer learning, IP, IEEE Trans. 23 (12) (2014) 5654–5669.
[34] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, F. Herrera, A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches, Syst. Man Cybern. Part C 42 (4) (2012) 463–484.
[35] A.S. Britto, R. Sabourin, L.E. Oliveira, Dynamic selection of classifiers - a comprehensive review, Pattern Recognit. 47 (11) (2014) 3665–3680.
[36] L. Kuncheva, C. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach. Learn. 51 (2) (2003) 181–207.
[37] Y. Zhu, J. Liu, S. Chen, Semi-random subspace method for face recognition, Image Vision Comput. 27 (9) (2009) 1358–1370.
[38] Q. Gao, J. Liu, K. Cui, H. Zhang, X. Wang, Stable locality sensitive discriminant analysis for image recognition, Neural Netw. 54 (2014) 49–56.
[39] Q. Gao, J. Ma, H. Zhang, X. Gao, Y. Liu, Stable orthogonal local discriminant embedding for linear dimensionality reduction, IP, IEEE Trans. 22 (7) (2013) 2521–2531.
[40] Q. Gao, J. Liu, H. Zhang, J. Hou, X. Yang, Enhanced fisher discriminant criterion for image recognition, Pattern Recognit. 45 (10) (2012) 3717–3724.

[41] M. Galar, A. Fernandez, E. Barrenechea, F. Herrera, Eusboost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling, Pattern Recognit. 46 (12) (2013) 3460–3471.
[42] G. Yu, G. Zhang, C. Domeniconi, Z. Yu, J. You, Semi-supervised classification based on random subspace dimensionality reduction, Pattern Recognit. 45 (3) (2012) 1119–1135.
[43] N. Chawla, K. Bowyer, Random subspaces and subsampling for 2-d face recognition, CVPR, 2005.
[44] X. Wang, X. Tang, Random sampling for subspace face recognition, Int. J. Comput. Vision 70 (1) (2006) 91–104.
[45] Y. Li, W. Shen, X. Shi, Z. Zhang, Ensemble of randomized linear discriminant analysis for face recognition with single sample per person, IEEE Automatic Face and Gesture Recognition (FG), 2013.
[46] C. Chen, A. Dantcheva, A. Ross, An ensemble of patch-based subspaces for makeup-robust face recognition, Inf. Fusion (2015) 1–13.
[47] Z.-Q. Zeng, J. Gao, Improving svm classification with imbalance data set, Neural Information Processing, 2009.
[48] K. Veropoulos, C. Campbell, N. Cristianini, et al., Controlling the sensitivity of support vector machines, IJCAI, 1999.
[49] T. Malisiewicz, A. Gupta, A. Efros, Ensemble of exemplar-svms for object detection and beyond, ICCV, 2011.
[50] I. Misra, A. Shrivastava, M. Hebert, Data-driven exemplar model selection, WACV, 2014.
[51] L. Didaci, G. Giacinto, F. Roli, G.L. Marcialis, A study on the performances of dynamic classifier selection based on local accuracy estimation, Pattern Recognit. 38 (11) (2005) 2188–2191.
[52] A.H. Ko, R. Sabourin, A.S. Britto, Jr., From dynamic classifier selection to dynamic ensemble selection, Pattern Recognit. 41 (5) (2008) 1718–1731.
[53] R. Caruana, A. Munson, A. Niculescu-Mizil, Getting the most out of ensemble selection, ICDM, 2006.
[54] M. Galar, A. Fernandez, E. Barrenechea, F. Herrera, Drcw-ovo: Distance-based relative competence weighting combination for one-vs-one strategy in multi–class problems, Pattern Recognit. 48 (1) (2015) 28–42.
[55] B. Krawczyk, B. Cyganek, Selecting locally specialised classifiers for one-class classification ensembles, Pattern Anal. Appl. (2015) 1–13.
[56] L.I. Kuncheva, Combining pattern classifiers: methods and algorithms, John Wiley & Sons, 2004.
[57] V. Cheplygina, D. Tax, Pruned random subspace method for one-class classifiers, Multiple Classifier Systems, 6713, 2011.
[58] T. Ahonen, E. Rahtu, V. Ojansivu, J. Heikkila, Recognition of blurred faces using local phase quantization, ICPR, 2008.
[59] O. Deniz, G. Bueno, J. Salido, F.D. la Torre, Face recognition using histograms of oriented gradients, Pattern Recognit. Lett. 32 (12) (2011) 1598–1603.
[60] C.-C. Chang, C.-J. Lin, Libsvm: A library for support vector machines, ACM TIST 2 (3) (2011) 1–27.
[61] M. Yang, L. Van Gool, L. Zhang, Sparse variation dictionary learning for face recognition with a single training sample per person, ICCV, 2013.
[62] J. Zhang, Y. Yan, M. Lades, Face recognition: eigenface, elastic matching, and neural nets, IEEE 85 (9) (1997) 1423–1435.
[63] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[64] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, Y. Ma, Toward a practical face recognition system: Robust alignment and illumination by sparse representation, PAMI, IEEE Trans. 34 (2) (2012) 372–386.

Saman Bashbaghi received the B.Sc. degree in computer engineering and M.Sc. in artificial intelligence from Bu-Ali Sina University, Hamedan, Iran, in 2010 and 2012, respectively. He is currently pursuing the Ph.D. in Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA) at the École de Technologie Supérieure (ETS). His main research interests are pattern recognition, computer vision, adaptive classification systems, video surveillance and deep learning.

Eric Granger obtained a Ph.D. in Electrical Engineering from the École Polytechnique de Montréal in 2001. From 1999 to 2001, he was a Defence Scientist at Defence R&D Canada in Ottawa. Until then, his work was focused primarily on neural networks for fast classification of radar signals in Electronic Surveillance (ES) systems. From 2001 to 2003, he worked in R&D with Mitel Networks Inc. on algorithms and electronic circuits to implement cryptographic functions in Internet Protocol (IP) based communication platforms. In 2004, he joined the ETS, Université du Québec, where he has developed applied research activities in the areas of patterns recognition, computer vision and microelectronics. He is presently Full Professor in System Engineering. Since joining ÉTS, he has been a member of the Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), and his main research interests are adaptive classification systems, incremental learning, change detection, and multi-classifier systems, with applications in biometrics, video surveillance, and computer and network security.

Robert Sabourin joined the physics department of the Montreal University in 1977 where he was responsible for the design, experimentation and development of scientific instrumentation for the Mont Mégantic Astronomical Observatory. His main contribution was the design and the implementation of a microprocessor based fine tracking system combined with a low light level CCD detector. In 1983, he joined the staff of the École de Technologie Supérieure, Université du Québec, in Montréal where he cofounded the Dept. of Automated Manufacturing Engineering where he is currently a Full Professor, and teaches Pattern Recognition, Evolutionary Algorithms, Neural Networks and Fuzzy Systems. In 1992, he joined also the Computer Science Department of the Pontifícia Universidade Católica do Paraná (Curitiba, Brazil) where he was, co-responsible for the implementation in 1995 of a master program and in 1998 a PhD program in applied computer science. Since 1996, he is a senior member of the Centre for Pattern Recognition and Machine Intelligence (CENPARMI, Concordia University). Since 2012, he is the Research Chair holder specializing in Adaptive Surveillance Systems in Dynamic Environments. Dr. Sabourin is the author (and coauthor) of more than 400 scientific publications including journals and conference proceeding. He was co-chair of the program committee of CIFED'98 (Conférence Internationale Francophone sur l'Écrit et le Document, Québec, Canada) and IWFHR'04 (9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan). He was nominated as Conference co-chair of ICDAR'07 (9th International Conference on Document Analysis and Recognition) that has been held in Curitiba, Brazil in 2007. His research interests are in the areas of adaptive biometric systems, adaptive surveillance systems in dynamic environments, intelligent watermarking systems, evolutionary computation and biocryptography.

Guillaume-Alexandre Bilodeau received the B.Sc.A. degree in computer engineering and the Ph.D. degree in electrical engineering from Université Laval, QC, Canada, in 1997 and 2004, respectively. He was appointed as an Assistant Professor with Polytechnique Montréal, QC, Canada, in 2004, where he was an Associate Professor in 2011. Since 2014, he has been a Full Professor with Polytechnique Montréal. His research interests encompass image and video processing, video surveillance, object recognition, content-based image retrieval, and medical applications of computer vision. Dr. Bilodeau is a member of the Province of Québec's Association of Professional Engineers and REPARTI research network.