Accepted in Machine Vision and Applications, 2017

Robust Watch-List Screening Using Dynamic Ensembles of SVMs Based on Multiple Face Representations

Saman Bashbaghi[1], Eric Granger[1], Robert Sabourin[1], and Guillaume-Alexandre Bilodeau[2]

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

[2] Laboratoire d'interprétation et de traitement d'images et vidéo, Polytechnique Montréal, Montréal, Canada

bashbaghi@livia.etsmtl.ca, eric.granger@etsmtl.ca, robert.sabourin@etsmtl.ca, gabilodeau@polymtl.ca

Abstract. Still-to-video face recognition (FR) is an important function in video surveillance, where faces captured over a network of video cameras are matched against reference stills of target individuals. Screening faces against a watch-list is a challenging video surveillance application because the appearance of faces vary due to changing capture conditions and operational domains. The facial models used for matching may not be representative of faces captured with video cameras because they are typically designed a priori with only one reference still. In this paper, a multi-classifier framework is proposed for robust still-to-video FR based on multiple and diverse face representations of a single reference face still. During enrollment of a target individual, the single reference face still is modeled using an ensemble of SVM classifiers based on different patches and face descriptors. Multiple feature extraction techniques are applied to patches isolated in the reference still to generate a diverse SVM pool that provides robustness to common nuisance factors (e.g., variations in illumination and pose). The estimation of discriminant feature subsets, classifier parameters, decision thresholds, and ensemble fusion functions is achieved using the high-quality reference still and a large number of faces captured in lower quality video of non-target individuals in the scene. During operations, the most competent subset of SVMs are dynamically selected according to capture conditions. Finally, a head-face tracker gradually regroups faces captured from different people appearing in a scene, while each individual-specific ensemble performs face matching. The accumulation of matching scores per face track leads to a robust spatio-temporal FR when accumulated ensemble scores surpass a detection threshold. Experimental results obtained with the Chokepoint and COX-S2V datasets show a significant improvement in performance w.r.t. reference systems, especially when individual-specific ensembles (1) are designed using exemplar-SVMs rather than one-class SVMs, and (2) exploit score-level fusion of local SVMs (trained using features extracted from each patch), rather than using either decision-level or feature-level fusion with a global SVM (trained by concatenating features extracted from patches).

Keywords: Video Surveillance, Watch-List Screening, Face Recognition, Single Sample Per Person, Ensemble Methods, Exemplar-SVMs, Dynamic Ensemble Selection.

The final publication is available at Springer via http://dx.doi.org/10.1007/s00138-016-0820-4

1 Introduction

Given the covert and non-intrusive nature of many security and surveillance applications, face recognition (FR) is widely employed in law enforcement, forensics, access controls, information security and video surveillance (VS) [25]. Systems for FR in VS attempt to compare the faces captured over a network of cameras against facial models[3] of target individuals enrolled to the system.

In VS, capture conditions typically range from semi-controlled with one person in the scene (e.g. passport inspection lanes and portals at airports), to uncontrolled free-flow in cluttered scenes (e.g. airport baggage claim areas, and subway stations). Two common types of applications in VS are: (1) watch-list screening (that requires a system for still-to-video FR), and (2) face re-identification or search and retrieval (that requires a system for video-to-video FR) [41], [47]. In the former application, reference face images or stills of target individuals of interest are used to design facial models, while in the latter, facial models are designed using faces captured in reference videos. This paper focuses on systems for still-to-video FR in semi-controlled conditions.

During enrollment of target individuals, facial regions of interests (ROIs) are isolated in reference images that were captured under controlled condition, and used to design facial models. Then, during operation, the ROIs of faces captured in videos are matched against the facial models of each individual enrolled to the watch-list. In VS, a person in a scene may be tracked over several frames, and matching scores may be accumulated over a facial trajectory (a group of ROIs that correspond to the same high-quality track of an individual) for robust spatio-temporal FR. An alarm is triggered if accumulated matching scores linked to a watch-list individual surpasses an individual-specific threshold [13].

In watch-list screening, the number of representative reference stills per target individuals is limited. Moreover, ROIs isolated from reference stills may differ significantly from those captured in video frames, due to camera inter-operability, and to the different capture conditions. The appearance of faces captured in video under semi- or uncontrolled conditions may also vary considerably according to changes in ambient illumination, pose, sharpness, scale, resolution, expression, occlusion, and blur [5].

It is assumed in this paper that only one high-quality reference face still is available to enroll a target individual to the still-to-video FR system. This is a common situation in watch-list screening applications due to the cost and feasibility of capturing reference stills, and managing facial models over time. In pattern recognition literature [46], this challenging situation is referred to as a "single sample per person" (SSPP) problem, and the resulting lack of representativeness of facial models can yield poor FR performance. Techniques specialized for a SSPP problem in FR rely on multiple face representations (employing various face descriptors), synthetic generation (2D morphing and 3D reconstruction), and enlarging the training set using auxiliary data [28], [46], [39]. In this paper, the SSPP problem found in watch-list screening is addressed by exploiting multiple face representations, particularly through multiple patch configurations and multiple

[3] A *facial model* is defined as either a set of samples extracted from one or more reference face images (stored in a gallery for a template matcher), or a set of classifier parameters estimated from reference samples (for a pattern classifier).

feature extraction techniques to design the individual-specific classification system. It is also addressed by employing an auxiliary dataset comprised of faces extracted from a large number of operational videos with non-target individuals captured in the scene.

Ensemble methods have been shown in many studies to improve the accuracy and robustness of a classification systems, where there is limited design data [19], [20] . To design still-to-video FR systems from a limited number of reference face stills, a diversified pool of base classifiers can be generated to design an individual-specific ensemble through multiple representations. Multiple face representations of a single target ROI pattern has been shown to significantly improve the overall performance of basic template-based still-to-video FR system [6], [33]. Moreover, modular systems designed using individual-specific ensembles have been successfully applied to the detection of target individuals in VS [41], [47].

The framework proposed in this paper provides insights for the design of individual-specific ensembles that are robust in still-to-video FR when only one reference still is available to represent face models. Given the target and non-target data available for design, one-class Support Vector Machine (SVM) and the exemplar SVM (2-class) [37] are considered for the base classifiers. They follow a discriminant approach that is robust to limited reference data and class imbalance. These specialized ensembles of SVMs model the variability in facial appearances by generating multiple and diverse face representations that are robust to various nuisance factors commonly found in VS environments, like variations in pose and illumination.

During enrollment of a target individual, the corresponding facial model is encoded into an ensemble of specialized SVMs using a ROI extracted from a single high-quality reference still. A pool of diverse SVM classifiers is generated from multiple face representations of the reference ROI obtained by extracting face descriptors from patches. In particular, uniform non-overlapping patches are isolated in the reference ROI to improve robustness to occlusion [34]. Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Histogram of Oriented Gradients (HOG), and Haar features are considered to extract information from patches to provide robustness to local changes in illumination, blur, etc [2], [3], [9], [16]. During operations, ROIs of faces captured in videos are classified by each individual-specific ensemble of the system, and the ensemble scores are combined. Then, these scores are accumulated over trajectories of each person appearing in the scene for robust spatio-temporal recognition.

In particular, this paper focuses on the analysis of different face patches and descriptors, one- and two-class SVM classifiers, and ensemble fusion strategies that are most suitable for the application constraints. Thus, SVM ensembles are trained using a single reference target ROI obtained from a high-quality generic still reference versus many non-target ROIs captured from low-quality videos. These non-target ROIs acquired from specific camera viewpoints, and of video cameras belonging to unknown people in the environment (background model) are used throughout the design process to estimate classifier parameters and ensemble fusion functions, to select discriminant feature subsets and decision thresholds, and to normalize the scores.

To form discriminant ensembles, this paper considers the benefits of selecting and combining patch- and descriptor-based classifiers with fusion at a feature-, score-, and decision-level, and by following a new dynamic selection strategy. To improve perfor-

mance, specialized strategy allows to perform dynamic selection of ensembles based on patch ROIs with SVM properties, by measuring the distance between the target ROI patterns and support vectors.

The performance of the still-to-video FR systems designed according to the proposed framework are compared to reference systems [6], [41], [53] using videos from the publicly-available Chokepoint [50] and COX-S2V [22] datasets. Accuracy and efficiency are measured at the transaction-level (matching of input probe ROI against reference ROI) and at the trajectory-level (the entire FR system over multiple frames).

The rest of this paper is organized as follows. Section 2 presents an overview of state-of-the-art systems for still-to-video FR. Section 3 presents detailed description of the proposed framework. Experimental methodology and results are presented and interpreted in Section 4.

2 Systems for Still-to-Video FR

Systems designed for FR in VS applications aim to detect the presence of target individuals of interest from faces captured with a network of video cameras. During enrollment of a target individual, some features are extracted from ROIs isolated in the reference face stills to create representative facial models and then to construct a gallery. During operation, probe ROI patterns captured from videos are matched against facial models within the gallery. In this context, FR is typically achieved using a system comprised of modules for segmentation, tracking, feature extraction, classification, and spatio-temporal fusion (see Figure 1) [42].

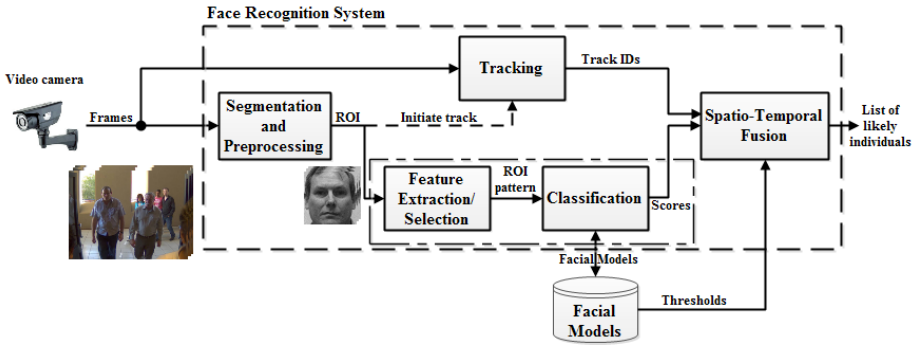


Fig. 1: Generic system of spatio-temporal FR in video surveillance.

In these systems, each video camera in a network captures the environment in its field of view that may contain individuals during operational phase. The segmentation and preprocessing module first detects faces and isolates the ROI in video frames. Then, a face track is initiated for each new individual appearing in the scene. Afterwards, some

feature extraction/selection module extracts an invariant and discriminative set of features. Once features are extracted, they are assembled into an ROI pattern and processed by the classification system. Finally, classification allows to compare the similarity of a each probe ROI pattern against facial models of individuals enrolled to the system to generate matching [13].

A face tracker follows the location of facial region of people appearing in the scene. This module regroups probe ROIs of a same individual captured over consecutive frames into facial trajectories. Finally, a spatio-temporal fusion module integrates the responses of the classifier and tracker. Accumulation of classifier responses for successive ROI patterns over multiple frames can lead to robust spatio-temporal FR. In watch-list screening, a person of interest is often detected in the scene with spatio-temporal recognition by comparing the matching scores over a trajectory with a predefined threshold [47].

2.1 State-of-the-Art Techniques

Multi classifier systems are often used for FR in VS specialized for open-set problems, where the number of non-target samples outnumber target samples of individuals of interest [8]. Video-based FR systems should be modeled as an individual-specific detection problem [42], where each detector is implemented with a mixture of reference ROI patterns of the individual of interest and a set of ROI patterns of non-target individuals. In these modular architectures, it is easier to add and remove individuals over time, and also different decision threshold, feature subsets and classifiers can be assigned to a specific individual. Given the limited number of reference samples, and the complexity of environments in VS, individual-specific detectors have also been implemented using ensemble methods to improve the overall system accuracy [20], [42].

Ensemble of 2-class ARTMAP classifiers per individual has been designed in [42]. In contrast, a modular system based on SVMs has been provided in [18] for real-time FR in real-world surveillance settings. More recently, an adaptive ensemble-based system has been proposed to self-update the facial models. An individual-specific ensemble is updated if its recognition over a trajectory is with high confidence [47].

Few specialized systems have been proposed for still-to-video FR as required for watch-list screening. A probabilistic tracking-and-recognition approach called sequential importance sampling [59] has been proposed for still-to-video FR by converting still-to-video into video-to-video using frames satisfying required scale and pose criteria during tracking. Similarly, a probabilistic mixture of Gaussians learning algorithm using expectation-maximization (EM) from sets of static images is presented for video-based FR system which is partially robust to occlusion, orientation, and expression changes [57]. A matching-based algorithm employing several correlation filters is proposed for still-to-video FR from a gallery of a few still images in [52], where it was assumed that the poses and viewpoints of the ROIs in video sequences are the same as corresponding training images. A local facial feature based framework performing the matching of stills against video frames with different features (e.g., manifold to manifold distance, affine hull method, and multi-region histogram) has been proposed in [45], where these features are extracted from a set of stills driven by utilizing spatial and temporal video information.

Recently, sparse representation-based classification (SRC) methods have been shown to provide a high-level of performance in FR [51]. The conventional SRC method is not capable of operating with one reference still, yet an auxiliary training set has been exploited in extended SRC (ESRC) [15] to enhance robustness to the intra-class variation. Similarly, an auxiliary training set has been exploited with the gallery set to develop a sparse variation dictionary learning (SVDL), where an adaptive projection is jointly learned to connect the generic set to the gallery set, and to construct a sparse dictionary with sufficient variations of representations [53]. In addition, an ESRC approach through domain adaptation (ESRC-DA) has been lately proposed in [40] for still-to-video FR incorporating matrix factorization and dictionary learning. Despite their capability to handle the SSPP problem, they are not fully-adapted for still-to-video FR systems. Indeed, they are relatively sensitive to variations in capture conditions (e.g., considerable changes in illumination, pose, and especially occlusion). In addition, samples in the generic training set are not necessarily similar to the samples in the gallery set due to the different cameras. Hence, the intra-class variation of training set may not translate to discriminative information regarding samples in the gallery set. They may also suffer from a high computational complexity, because of the sparse coding and the large and redundant dictionaries [15], [53].

Another approach is to identify the decision region of individual faces in the feature space, a specialized feed-forward neural network using morphing to synthetically generate variations of a reference still is trained for each target individual for watch-list surveillance, where human perceptual capability is exploited to reject previously unseen faces [27]. Recently, in [22] partial and local linear discriminant analysis has been proposed using samples containing a high-quality still and a set of low resolution video sequences of each individual for still-to-video FR as a baseline on the COX-S2V dataset. Similarly, coupling quality and geometric alignment with recognition [23] has been proposed, where the best qualified frames from video are selected to match against well-aligned high-quality face stills with the most similar quality. Low-rank regularized sparse representation is adopted in a unified framework to interact with quality alignment, geometric alignment, and face recognition.

However, watch-list screening is a challenging problem, and performance of the state-of-the-art still-to-video FR systems decline due to semi- or uncontrolled conditions and camera inter-operability [10], [22]. Nuisance factors that cause changes in facial appearance are mostly variations in illumination, pose, scale, resolution, expression, motion blur, and occlusion [5]. In this context, face models are not typically representative of faces captured in the operational environment. Beyond the limited representativeness of reference stills, still-to-video FR is very challenging when only one face still is available per person for system design. However, selecting ROI patterns from videos with non-target individuals in the scene may be explored to optimize performance necessary to overcome the problem of imbalanced data and to estimate classifier parameters, selecting discriminant features, and decision thresholds.

2.2 Single Sample Per Person Solutions

To compensate the limited representativeness of facial models due to a SSPP, several approaches have been proposed in the FR literature. They can be divided into three

main categories: multiple face representation, synthetic generation of virtual faces, and using auxiliary data from non-target people in the scene to enlarge the training set [46].

In multiple face representations, different feature extraction techniques are employed to generate multiple discriminant and robust representations from a single reference still [6], where the key issue with this type of approaches is combining those representations appropriately. In synthetic face generation, several virtual face images are synthesized using 2D morphing or 3D reconstructions to enhance the number of target samples with different pose and viewpoints [27], [39]. The problem with these approaches is to exploit prior knowledge to locate the facial components reliably. Finally, a generic auxiliary dataset containing non-target ROI patterns of other people possibly captured under different conditions can be used to provide intra-class variations [15], [28]. However, the main concern with this kind of approaches is the large differences between target faces in the gallery and non-target faces from the generic set.

2.3 Multiple Face Representations

To compensate for the impact of using only a single design sample, multiple face representations may be generated from the target reference still, using various feature extraction techniques and patch-based methods.

Feature Extraction Techniques In FR literature, feature extraction techniques may be classified into holistic and local approaches based on locations and the ways they have been applied to face images [1], [46].

1. Holistic Approaches: These methods characterize the appearance of the entire face, and use the whole ROI to extract features. For instance, each ROI can be represented as a single high-dimensional ROI pattern by concatenating the grayscale (intensities) or color values of all pixels. In these appearance-based methods, all pixels of a face image may be involved in the extraction process. Holistic methods are generally divided into two main types as follows.
   (a) Projection-based techniques: These methods typically transform the data from the original space to a new coordinate system in order to either reduce the dimensionality or classification process. Techniques such as PCA: principal component analysis (eigenface), LDA: linear discriminant analysis (fisherface), LPP: locality preserving projection (laplacianface), and LLE: local linear embedding [21], [43], [55] belong to this category. Due to the high-dimensional representation of face images, these techniques need sufficiently large training set to tackle the curse of dimensionality issue. Thus, they are not desired approacher to perform FR given a SSPP. However, they can be manipulated appropriately to provide either lower dimensional representation or feature selection.
   (b) Image processing techniques: In this category, image feature descriptors are exploited for providing face representation. These descriptors may scan either image regions and then extract features such as LBP or use image color histograms or mean/variance of grayscale values, and transformation such as Haar

and Gabor [2], [3], [35]. Dense computation can be also applied to extract features from regions such as HOG [16].

2. Local Approaches: These methods use local facial characteristics for generating face representation. Care should be taken when deciding how to incorporate global structural information into local face model. They are employed to characterize the information around a set of salient points, like eyes, nose, mouth, etc., or any local regions based on neighborhood or adjacency of pixels. They can be divided into two categories based on their definition of image locality.
   (a) Local facial feature-based techniques: These approaches first process the input image to locate and extract distinctive facial features such as the eyes, mouth, nose, etc., and then compute the geometric relationships among those facial points, thus reducing the facial image to a geometric feature vector. In other words, local facial features such as the eyes, nose, and mouth are taken into account along with their locations and local statistics (geometric and/or appearance). Therefore, these techniques extract structural information such as the width of the head, the distances between the eyes, etc. Thus, methods proposed based on extracting structural information aim to detect the eyes and mouth in real images, where various configurationally feature such as the distance between two eyes are manually derived from the single face image, such as Active Shape Model (ASM) [58].
   (b) Local appearance-based techniques: Local appearance-based methods extract information from defined local regions. Two steps are generally involved in these methods: (1) local region partitioning (to detect keypoints), and then (2) feature extraction from the neighborhood of those points. Local appearance-based face representation, like SIFT and SURF [17] are generic local approaches and do not require determining of any salient local facial region manually.

So far, no feature extraction technique can by itself overcome all the nuisance factors encountered in VS. In this paper, different holistic feature extraction techniques are exploited to generate multiple face representations that are robust to aforementioned nuisance factors.

Patch-Based Approaches With patch-based methods, several overlapping or non-overlapping sub-regions are extracted from facial ROIs, and then features are extracted within each patch for local matching FR [36], [61]. Patches can be defined uniformly using pyramid structures, saliency (detecting keypoints), or randomly. Local face matching is typically comprised of (1) alignment and partitioning, (2) extracting features from each partition, (3) and matching, as well as, combination [61]. Face alignment is a critical issue in FR of occluded and non-frontal faces, while a limited field of view and low-quality faces may significantly degrade the recognition performance [34]. Local matching with patch-based methods potentially offer higher discrimination, allowing to recognize either partially occluded faces or arbitrary poses that appear frequently in unconstrained VS environments.

Patch-based methods can be applied on the entire face image or local facial components (e.g., eye, node, and mouth) of the face image. Although, component-based

approaches, such as GMM and HMM part-based representations seem more convenient than other approaches in terms of their representativeness and robustness, they are typically complex. It may also be difficult to extract for real-time face screening system, because salient keypoints must first to be localized [34]. Furthermore, component-based approaches fail if facial components are occluded. Hence, in this paper, uniform patch configurations are exploited to mitigate the partial occlusions that may occurred in VS applications.

2.4 Face Classification Systems

Given only one target reference still ROI captured under controlled condition (from another scene and camera), and an abundance of non-target ROIs captured from videos, training classification system to address the variabilities in VS environment is challenging. Thus, a framework with an ensemble per person is considered. They have been shown to provide robust and accurate performance when training data is limited [47]. It is however challenging to train or generate a diverse pool of classifiers per target individual from the original data [33].

Techniques for classifier design under class imbalance can be broadly categorized into: (1) algorithms that take into account the importance of positive samples (internal or algorithm-level), (2) techniques that apply preprocessing steps to re-balance the data distribution (external or data-level), and (3), cost-sensitive methods that combine both internal and external approaches to deal with different misclassification costs [19]. In addition, ensemble methods are often combined with one of the techniques above, specifically data-level and cost-sensitive ones [33], [56].

For the design of classifier ensembles for watch-list screening under imbalanced data, specialized classification techniques are required. A simple approach for designing still-to-video FR system is using nearest neighbor classifier or template matching. In this case, the classification system performs hypothesis testing (or one-class classification) using a single reference face image per target individual. Template matching algorithms employ each facial model defined as a set of one or more templates stored in a gallery [6], [9]. It is also possible to consider a one-class classifier like Gaussian mixture modeling [29] or one-class SVMs [60] to learn from an abundance of non-target class samples that are somehow similar to the single target class sample.

In this paper, specialized SVM classifiers are considered for face matching as acquired in still-to-video FR. SVM is a widely used discriminative classifier that finds the optimal hyperplane to separate data patterns into two classes [54]. It requires a small number of training patterns to correctly model the boundary. Consider a training dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ in a 2-class classification problem, where $x_i \in R^n$ and $y_i \in \{-1, +1\}$ represent an n-dimensional data pattern and the classes of these data, respectively, for $i = 1, 2, \ldots, l$. These data patterns are typically mapped into a higher dimensional feature space using a mapping function ϕ to find the best separation of classes. Therefore, the soft-margin optimization problem is formulated as the following expression:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \tag{1}$$

$$\text{subject to } y_i \left(\mathbf{w}^T \phi\left(x_i\right) + b\right) \geq 1 - \xi_i, \\ \xi_i \geq 0$$

where slack variables ξ_i are introduced to account for misclassified examples. Thus, $\sum_{i=1}^{l} \xi_i$ can be considered as a misclassification amount, $\mathbf{b}$ is the bias, and $\mathbf{w}$ is the weight vector. Constant C is a misclassification cost of a training example, where it controls the trade-off between maximizing the margin, as well as, minimizing the number of misclassifications.

Traditional SVM classifiers fail to classify imbalanced datasets properly, so that the estimated boundary is skewed to the majority class patterns [7]. For classification of imbalanced datasets, the SVM objective function should be biased to push away the boundary from the majority class patterns in order to decrease the effect of class imbalance. The Different Error Costs (DEC) method [48] was proposed to modify the SVM objective function, where two misclassification cost values C^+ and C^- are assigned as follows:

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

where C^+ and C^- are the misclassification costs for the positive and negative classes, respectively. The optimal result is typically achieved when C^+/C^- equals the imbalance ratio [56].

To overcome the class imbalanced issue and high misclassification rate, another SVM strategy named z-SVM is proposed to automatically orient the skewed decision boundary [24]. This method adjust the trained decision boundary toward the minority class regardless of learning parameters, contrary to existing SVM classifiers that exploits additional parameters empirically. To that end, a multiplicative weight z is assigned to each support vectors belonging to the minority class as follows:

$$f\left(x, z\right) = z \sum_{x_p \in SV;[p|y_p=+1]} \alpha_p y_p K\left(x, x_p\right) + \sum_{x_n \in SV;[n|y_n=+1]} \alpha_n y_n K\left(x, x_n\right) + b \tag{3}$$

where K is the kernel function, α is the Lagrangian constant, and SV is the set of support vectors. p and n indicates the positive and negative classes, respectively. This method is not convenient according to the assumptions of this paper, because it requires more than one positive samples in the minority class.

In addition, one-class SVM (OC-SVM) classifier is designed to deal with data originating from only one class. It typically attempts to distinguish the samples of class of interest from all other outliers, where it defines a model using minimum volume contour (circle) to describe the target data [31], [44]. Basically, finding the optimal size of the volume is an indispensable issue due to the fact that a small volume may lead to an over-trained model, while a large volume size may accept outliers extensively.

Let $\chi = \{x_1, \ldots, x_m\}$ be the training dataset, where each x_j is a feature vector of a target sample. The goal of OC-SVM is to learn a function f_χ that assigns the data

in χ to the set $R_\chi = \{f_\chi(x \geq 0)\}$, while minimizing the volume of R_χ. This issue is so called as estimation of minimal volume set, where membership of x to the set of R_χ determines whether its overall estimated volume is close to χ or not. A radial basis function (RBF) is used as a kernel function to estimate the minimal volume set. The OC-SVM constructs the hyperplane W to separate the training data mapped into the artificial feature space H from the hypersphere with radius equal to one $S_{(R=1)}$, as well as, to maximize the distance from it. Thus, the OC-SVM decision function $f_\chi(x)$ can be estimated as follows:

$$f_\chi(x) = \sum_{j}^{m} \alpha_j k(x_j, x) - \rho, \tag{4}$$

where $0 \leq \alpha_\mathrm{j} \leq \frac{1}{\mathrm{m}}$, $\sum_\mathrm{j} \alpha_\mathrm{j} = 1$, and the value of ρ is computed using $\mathrm{f}_\chi(\mathrm{x_j}) = 0$. Therefore, $\mathrm{x_j} \in \chi$ are located in the decision boundary when they satisfy both $\alpha_\mathrm{j} \neq 0$ and $\alpha_\mathrm{j} \neq \frac{1}{\mathrm{m}}$.

More recently, a method called exemplar-SVM [37] has been proposed to train a separate linear SVM classifier for every single positive example versus millions of negatives. The idea behind this method is to learn a separate 2-class classifier for each exemplar within a class of interest, unlike category-based classification. It is worth mentioning that this method has been mostly applied to ensemble learning in object detection and visual recognition tasks [26], [38]. However, as described in Section 3, e-SVMs provide several advantages in the design of individual-specific ensembles for still-to-video FR. In particular, it can be trained with one target sample, and regardless of number of non-targets, as well as, it can rank the non-target support vectors w.r.t the target still. Furthermore, integrating multiple and diverse e-SVMs into an ensemble, with each e-SVM being specialized for a particular descriptor and facial zone provides a robust facial model.

Let $\mathbf{a}$ be the positive sample (target individual ROI pattern) and U be the number of non-target (negative) samples, respectively. The formulation of the e-SVM cost function is:

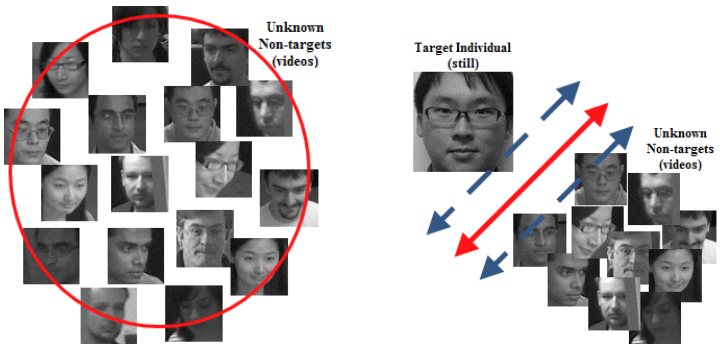
$$\min_{\mathbf{w},b} \left\{ \mathbf{w}^2 + C_1 \max(0, 1 - (\mathbf{w}^T a + b) + C_2 \sum_{x \in U} \max(0, 1 - (\mathbf{w}^T x + b)) \right\} \tag{5}$$

where C_1 and C_2 parameters control the weight of regularization terms, $\mathbf{w}$ is the weight vector, and b is the bias term. Since there is only one positive sample in the training set, its error is weighted much higher than the negative samples. The calibrated score of e-SVM for the given ROI pattern $\mathbf{a}$ and the learned regression parameters (α_a,β_a) is computed as follows:

$$f(x|\mathbf{w}, \alpha_a, \beta_a) = \frac{1}{1 + e^{-\alpha_a(\mathbf{w}_a^T - \beta_a)}}. \tag{6}$$

In the SSPP problems, OC-SVMs can be trained considering only the non-target samples obtained from unknown individuals (Figure 2 (a)), while e-SVMs can be trained

using a single target sample (still ROI pattern) along with many non-target samples (video ROI patterns) for each individual of interest as illustrated in Figure 2 (b). Thus, training can be performed by considering non-target ROIs as negative samples obtained from background model. Subsequently, the information of non-target individuals from the field of view may be exploited during training to enhance the capability to generalize during operation.



(a) OC-SVM (b) E-SVM

Fig. 2: Illustration of training OC-SVM and e-SVM for each individual of interest enrolled in the watch-list with a single ROI block.

3 Dynamic Ensembles of SVMs for Still-to-Video FR

A multi-classifier framework is proposed for robust still-to-video FR, where an ensemble of SVMs that encodes multiple discriminative face representations is assigned to each target individual (see Figure 3). Specifically, an individual-specific ensemble is designed using a diverse pool of specialized SVM classifiers to address the SSPP issue. This pool models the variability of faces by producing several face representations (various features extracted from patches) that are robust to common nuisance factors.

As illustrated in Figure 3, frames captured by a video camera may include several people. For each frame, preprocessing (e.g., grayscale conversion and histogram equalization) is first performed, and then segmentation is applied in order to isolate facial ROI(s). Then, the resulting ROIs are scaled into a predefined size and aligned based on the location of the eyes. Multiple ROI patterns are extracted from either the entire ROI, for i = 1, 2, ..., M number of feature extraction techniques, or from each patch p = 1, 2, ..., P. Each classifier (trained on the entire ROI or ROI patches) provides a matching score $S_{i,p}(\mathbf{a}_{i,p})$ between every ROI patch pattern $\mathbf{a}_{i,p}$ and the corresponding patch model $\mathbf{m}_{i,p^j}$ in the gallery index j=1, 2, ..., N indicates the number of individuals of interest enrolled to the system. Scores output from classifiers are fed into the fusion module after score normalization. A predefined threshold, $\gamma_{i,p}$ for each representation $\mathbf{a}_{i,p}$ is used to provide a decision d_i.

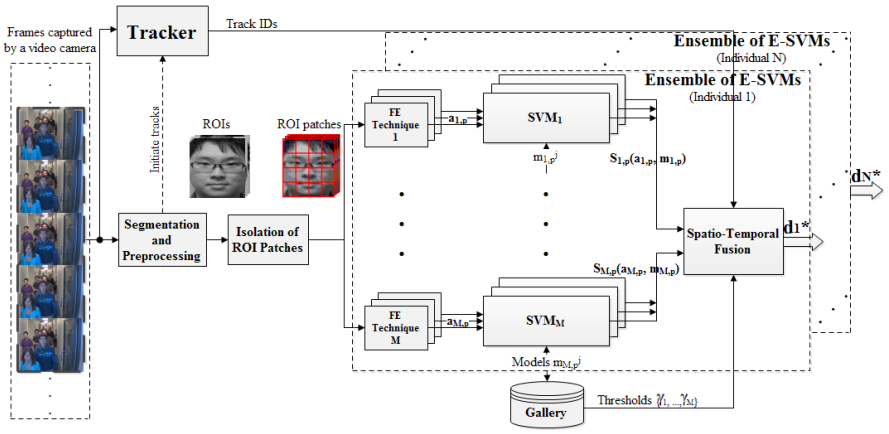


Fig. 3: Block diagram of the proposed system for still-to-video FR.

In order to improve accuracy and robustness of recognition, each ensemble of SVMs is trained during enrollment with a single reference face still versus many of non-target faces captured from video cameras in the scene. Hence, a diverse pool of SVM classifiers is generated during design and then combined dynamically during operation to provide the ensemble score. Finally, ensemble scores are accumulated over trajectories defined using a face tracker to provide robust spatio-temporal recognition. The following subsections present details on the proposed system.

3.1 Enrollment Phase

During enrollment of a target individual, multiple face representations are generated from the ROI isolated in a single reference still, and in unlabeled facial trajectories of unknown non-target individuals. ROI patterns randomly extracted from videos of non-target individuals allow to select discriminant features, to train individual-specific ensemble of SVM classifiers, to define decision thresholds and to normalize the scores. Several SVMs are trained to estimate the face models based on each representation (features extracted on each uniform patches).

A single still reference is first converted to grayscale and then a facial ROI is isolated using a face detection method [49]. Then, each ROI is scaled into a common predefined size, aligned, and then normalized for illumination invariance. Afterwards, different face descriptors are extracted from each patch in order to provide multiple face representations and generate a pool of diverse e-SVM and OC-SVM classifiers. For each representation, the ROI patch patterns of the target individual is combined with the corresponding ROI patterns of non-target individuals to train e-SVMs, while only ROI patterns of non-target individuals are employed to train OC-SVMs. For a system with P patches and M feature extraction techniques, enrollment involves generating a pool of MxP SVMs. Finally, decision thresholds computed from score distribution of non-targets [6] and preserved in the gallery.

Extraction of Multiple Face Representations Face descriptors (feature extraction techniques) and patch configurations employed in this paper provide robustness to at least one of the nuisance factors that may occur in VS environments, such as changes in illumination, pose, and scale, providing multiple discriminant representations that are uncorrelated is important to design a robust watch-list screening system. Different feature extraction techniques from FR literature have been categorized in Table 1.

Table 1: The main nuisance factors for FR in VS and some feature extraction techniques that have been proposed to provide robust representations.

Illumination	LDA, Direct LDA, Kernel LDA, Kernel PCA, LBP, Gabor filters, RIU-LBP, LPP, Haar, SIFT, LQP, SURF, Daisy, HOG, LPQ
Expression	PCA, 2DPCA, Discriminant PCA, KPCA, LBP, LDA, Direct LDA, ICA, E(PC)2A, LPP, HOG, DCT, LPQ, Daisy
Pose	Direct LDA, Haar, HOG, LPQ
Rotation	LBP, Gabor filters, SIFT, HOG, SURF
Occlusion	HOG, Haar/SURF (partial occlusion)
Scale	SIFT, SURF, Daisy, HOG
Motion Blur	LPQ
Aging(Time)	LBP, 2DPCA, ICA, DCT

LBP [2] and LPQ [3] are popular face descriptors that extract texture features of faces in different way. LPQ is more robust to motion blur because it relies on the frequency domain (rather than spatial domain) through the Fourier transform. LBP preserves the edge information, which remains almost the same regardless of illumination change. HOG and Haar features are selected to extract the information more related to shape. HOG [16] is able to provide a high level of discrimination on a SSPP because it extracts edges in images with different angles and orientations. Furthermore, HOG is robust to small rotation and translation. Wavelet transforms have shown convincing results in the area of FR [4]. In particular, Haar transform performs well with respect to pose changes and partial occlusion.

Finally, using multiple face representations generated through different face descriptors extracted from every face patch can increase the diversity among classifiers, robustness to variations, and tolerance to some occlusions. For each patch, a classifier is trained with the reference still patch versus the corresponding patches of non-target faces captured among universal background model. Features extracted from non-overlapping uniform patches from each ROI are used to train classifiers.

Generation of Diverse SVM Classifiers The diversity of SVMs in a pool is produced using multiple representations. It should be noted that the input features must be normalized between 0 and 1 through min-max normalization performed based on non-target face samples. Normalization of output scores for fusion is taken into account using min-max normalization as well.

E-SVMs possess some potential benefits in designing individual-specific classifier systems with multiple face representations from only one positive versus several negative samples. The large number of non-target samples appears to constrain the SSPP problem. Since this classifier finds support vectors that are highly similar to each individual when training e-SVMs, the amount of negative sample cannot affect the accuracy of the decision boundary [37]. Hence, it can be applied suitably even for large databases containing few exemplars in the training set, e.g., as acquired in watch-list screening.

Since each e-SVM is highly specialized to the target individual, the largest margin (decision boundary) will be obtained by training under imbalanced data exploiting different regularization parameters, which provides more freedom in defining the decision boundary. Therefore, this discriminative classifier is less sensitive to class imbalance than generative classifiers, or other classification techniques, such as neural networks and decision trees [54]. E-SVM as a passive learning approach impose no extra training overhead and compensate the imbalance data in the optimization process. Compared to other cost-sensitive SVMs, like z-SVM that apply active learning for classification of imbalanced data [24], the weights for classes are empirically determined during test mode. However, z-SVM requires more than one minority class sample to multiply the magnitude of positive support vectors by a particular small value of z estimated to bias the decision toward the majority negative class.

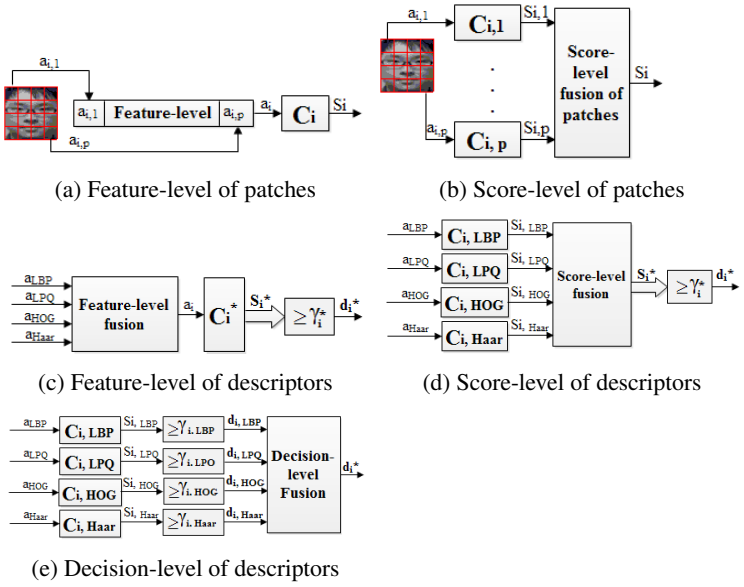
Since multiple representations can be generated from a single target to train these e-SVM classifiers, each classifier in an individual-specific ensemble is a different representation of an individual's face. Unlike similarity measurement methods, such as nearest neighbor schemes, e-SVMs do not necessarily compute distances to the other samples. Thus, combining e-SVMs into an ensemble may prevent over-fitting problems and simultaneously provides higher generalization performance [32].

This method can be interpreted as an approach to sort non-targets by visual similarity to the individual, because estimated support vectors also belong to the non-targets. However, in this case, since each e-SVM is supposed to correctly classify only visually similar faces, these faces can be used as an additional target samples that can be employed either in calibrating decision boundary or defining decision thresholds. As another advantage of using e-SVMs, the support vectors can be exploited as the closest non-target samples to the single target reference for selection of the most similar non-targets. Setting different regularization parameters during training, produce different number of support vectors. These support vectors can be ranked and used to define decision thresholds, although it could be difficult due to inter-operability of cameras.

As an alternative, a pool of OC-SVM classifiers may be generated using ROIs of non-target individuals selected to provide accurate decision boundaries. The main difference of this approach with conventional one-class classification is that SVMs are trained based on non-target class samples rather than samples from class of interest. In this context, contrary to template matching [6] that can be considered as a one-class classifier based on a single target reference still, OC-SVM can be defined as a method that either classifies non-target samples or rejects target samples during operation. Hence, the scores provided by OC-SVM classifiers can determine whether the input ROI patterns belong to non-target individuals or not and consequently target individuals are correctly detected.

3.2 Operational Phase

In the proposed system, different fusion approaches are applied to the proposed ensemble-based framework to achieve a higher level of generalization, and robustness [14]. Fusion techniques in such systems can be described as: (a) feature-level that aims to combine all the features extracted among patches into one feature vector in the feature space, (b) score-level attempts to combine the scores generated among patches using multiple classifiers trained per each patch, (c) feature-level concatenates several representations (descriptors), (d) score-level fusion of representations within the ensemble to provide the final score, and finally (e) decision-level of descriptors to produce the final response after applying decision thresholds as represented in Figure 4.



(a) Feature-level of patches (b) Score-level of patches

(c) Feature-level of descriptors (d) Score-level of descriptors

(e) Decision-level of descriptors

Fig. 4: Five approaches for fusion of responses after extracting features from multiple patches and descriptors of an individual j (for j = 1, 2, ..., N) considered in this paper.

With the feature-level fusion of patches, features extracted from the patches isolated within the ROI are concatenated to construct a long feature vector that is dimensionally equivalent to the number of patches multiplied by the dimensionality of feature extraction techniques. PCA is applied to project data in such that features may be ranked according to covariance, and the most correlated features may be reduced, where only one SVM classifiers is subsequently trained per ROI. In score-level fusion of patches, a separate SVM classifier is trained on the features extracted from each patch, so that a number of classifiers identical to the number of patches is trained per ROI. Moreover, multiple representations are concatenated after applying PCA and then a single classifier is trained to perform feature-level fusion of descriptors. Scores are combined among

multiple classifiers within the ensemble using the average function. Finally, decision-level fusion of descriptors consists in defining local decision thresholds for each descriptor specifically and exploiting majority vote to integrate their local decisions and produce the final decision. Decisions thresholds are defined using cumulative probability distribution function of non-target scores distribution at certain operating point of FPR=1% [6].

Dynamic Classifiers Selection In contrast to static approaches, the most competent classifiers in an individual's pool of classifiers that are trained over multiple face patches and representations can be selected and combined dynamically during operation in response to each probe ROI. Dynamic selection is used to improve the recognition accuracy by selecting the most competent classifiers and also to alleviate the computational cost. Hence, a novel approach is proposed to provide the best separation w.r.t. non-target samples in order to select an ensemble of classifiers based on a single high-quality target face still and many non-target low-quality video faces. Thus, the key idea is to allow the system to select those classifiers (face representations) that most properly discriminate target versus non-targets. In addition, this approach can improve the run-time speed in such applications by combining the selected classifiers rather than the entire pool. The proposed classifier selection method is formalized in Algorithm 1.

Algorithm 1 Dynamic ensemble selection method for individual j.

Input: Pool of diverse classifiers $P_j = \{c_{j,1}, \ldots, c_{j,M}\}$, the set of support vectors SV_j, the reference target still G_j, and the dataset of probe video ROIs D_{test}
Output: the set of the most competent classifiers $\{C^*\}$ for each testing sample t in D_{test}
for each probe ROI t in D_{test} **do**
    Divide t into uniform patches p
    $\mathbf{a}_{i,p} \leftarrow$ extract ROI pattern i from each patch p
    for each target individual j **do**
        Project $\mathbf{a}_{i,p}$ into the feature space of SV_j in P_j and the target still G_j
        for each classifier $c_{j,k}$ in P_j **do**
            if $Dist\left(\mathbf{a}_{i,p}, \mathbf{T}_{j,p}\right) \leq Dist\left(\frac{\sum_{s=1}^{s=||SV||} \mathbf{a}_{i,p}, \mathbf{V}_{i,p}}{||SV||}\right)$ **then**
                $\{C^*\} \leftarrow c_{j,k}$
            end if
        end for
        if $\{C^*\}$ is empty **then**
            Combine all classifiers C in the pool to classify t using mean function
        else
            Combine $\{C^*\}$ to classify t using mean function
        end if
    end for
end for

The selection criteria (level of competence) per given ROI pattern has two components: (1) distance from non-target support vectors, and (2) closeness to the target reference still, where if the distance between the input pattern and the target still is lower than the distance from support vectors (average distances from all support vectors), then those classifiers are selected dynamically. Contrarily to the conventional approaches that use local neighborhood accuracy for measuring the level of competence, it is not necessary in this approach to define neighborhood by measuring the distance from all the validation data. However, Euclidean distance is employed to measure the distances between the input pattern and either target still or non-target support vectors.

Spatio-Temporal Fusion In the proposed system, the head-face tracks are also exploited allowing for accumulation of scores associated with a same person to fulfill a robust spatio-temporal recognition. ROI captures for different individuals are regrouped into facial trajectories. Predictions for each individual are accumulated over time and if positive predictions surpass the detection threshold, then an individual of interest is detected. In particular, decision fusion module accumulates the ensemble scores S_j^* (obtained using score-level fusion) of each individual-specific ensemble over a fixed size window W according to:

$$d_j^* = \sum_{w=0}^{W-1} S_j^* \left[S_{i,p(W-w)} \right] \in [0, W] \tag{7}$$

4 Experimental Methodology

Different aspects of the proposed framework are evaluated experimentally using Chokepoint [50] and COX-S2V [22] still-to-video datasets. First, experiments assess the performance of classifiers trained on ROI patterns extracted using different feature extraction techniques. Second, experiments investigate the impact of patch configurations on the performance. Third, the performance of different levels and types fusion are compared. Finally, experiments show the effect of employing a tracker to form facial trajectories accumulate the ensemble predictions over consecutive frames in a trajectory and performing spatio-temporal recognition.

4.1 Video Database

In real-world scenarios such as portals, the videos produced by surveillance cameras have some variations containing changes in illumination, pose, expression, and motion of individuals, scales and occlusion. Chokepoint [4] and COX-S2V [5] dataset are selected based on these characteristics (see Table 2). These video surveillance datasets can be employed to emulate real-world watch-list screening applications. The main characteristics of these two datasets with respect to others [11], [30] are that they contain a high-quality still face images captured under controlled condition (with the same still

[4] http://itee.uq.edu.au/ uqywong6/chokepoint.html
[5] http://vipl.ict.ac.cn/resources/datasets/cox-face-dataset/COX-S2V

camera), and low-quality surveillance videos for each subject captured under uncontrolled conditions (with surveillance cameras).

Table 2: Characteristics of Chokepoint and COX-S2V datasets. Conditions include: indoor/outdoor (i/o); varying pose (p), illumination (l), expression (e), and scale (s); motion blur (b); occlusions (c); walking (w); random actions and/or motion (r); surveillance quality (v); and multiple people (m).

Characteristics	**Chokepoint**	**COX-S2V**
Number of persons	25 portal 1, 29 portal 2	1000
Resolution	800x600	1920x1080
Number of videos	54	4000
Frame Rate (fps)	30	25
Condition	i, p, l, e, s, b, c, w, v, m	i, p, l, e, s, b, c, w, v, m

Chokepoint dataset can be used as a benchmark for large-scale FR, especially in watch-list applications. An array of three cameras is placed above several portals to capture subjects walking through each portal in a natural way, used for simultaneously recording the entry of a person from different viewpoints. While a person is walking through a portal, a sequence of face images can be captured. Random examples of neutral still ROIs of target individuals and ROIs captured from different trajectories are depicted in Figure 5. The variations between viewpoints allow for variations in walking directions, facilitating the capture of a near-frontal face by one of the cameras. In the database, each testing video sequence is named according to the recording conditions, for example (P1E_S1_C1) where P, S, and C stand for portal, sequence and camera, respectively. E and L indicate subjects entering or leaving the portal.

Another publicly available still-to-video dataset called COX-S2V dataset is also employed to fulfill more experiments on watch-list screening. This dataset contains high-quality controlled still faces of 1000 subjects along with uncontrolled low-quality video sequences, where video clips are captured using two different off-the-shelf camcorders. In these videos, subjects walking naturally through a designed-S curve with changes in illumination, expression, scale, and viewpoint. Thus, four video clips with various resolutions are recorded per subject simulating VS scenario and located in the probe set. An example of four low-resolution video sequences is shown in Figure 6. It is worth noting that this dataset is much more challenging than Chokepoint, because there are only 25 captures are available for each sequence during operation, as well as, the ROIs captured are blurry.

4.2 Protocol for Validation Process

To validate the proposed system, all video sequences of Chokepoint data are chosen. In these experiments, 5 persons are selected randomly to be placed in the watch-list when only a single high-quality reference still and videos of 10 people that are assumed

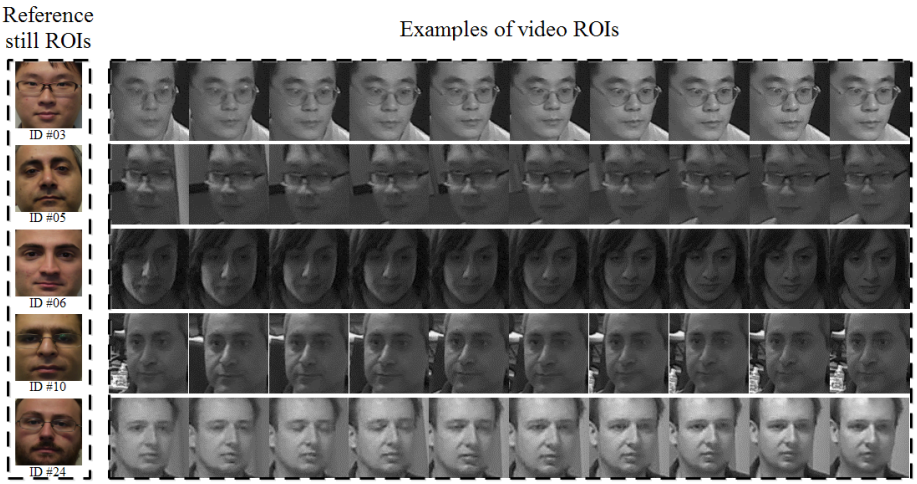


Fig. 5: Example of ROIs captured from the 'neutral' mugshot of 5 target individuals of interest and random examples of ROIs captured from 5 different trajectories in Chokepoint videos.

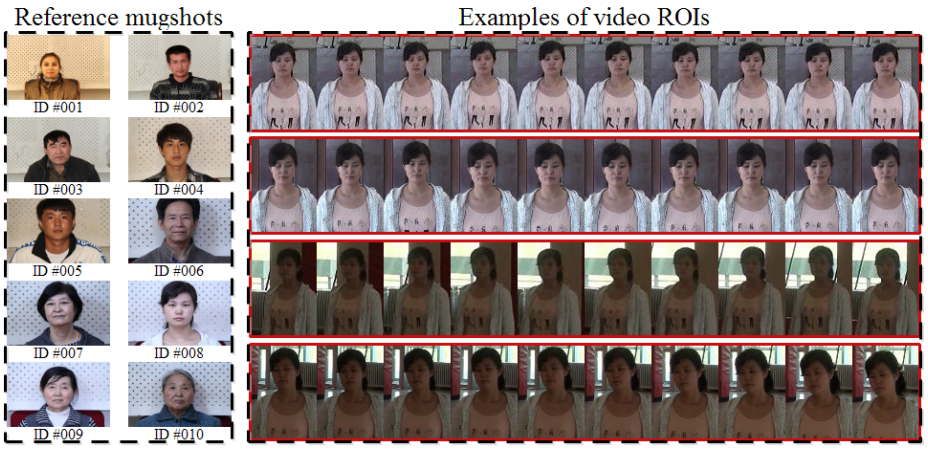


Fig. 6: Example of individuals of interest enrolled in the watch-list and low-resolution video sequences captured with off-the-shelf camcorders under uncontrolled conditions in the COX-S2V dataset.

to come from non-target persons are used during enrollment. Thus, the rest of videos including 10 other non-target persons and 5 videos of persons who are already enrolled in the watch-list are used for testing. This process is repeated 5 times with random selection of targets and non-targets.

Chokepoint data contains videos with 29 persons, while some of them do not appear in all video sequences. Therefore, in each test iteration, target individuals (one at a time) and unknown individuals within the test videos pass through the portal, where the system seeks to detect the target person during operation and this process is repeated for other video sequences.

In the experiments using COX-S2V, 20 high-quality stills are randomly chosen from Persons-for-Publication folder (see Figure 6) to participate in the watch-list along with video clips recorded from all videos for design phase. Videos of 100 other persons are considered as non-target individuals for the design phase, as well as, videos of 100 other unknown persons as testing videos. Hence, one target individual at a time and non-target persons in the testing videos are participating in the operating phase. In order to accomplish statistically significant results, these experiments are iterated 5 times with 20 different individuals of interest.

The size of the reference stills and captured ROIs are scaled to 48x48 pixels due to operational time. Libsvm library [12] is used in order to train e-SVMs and OC-SVMs. 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). Previous study [56] and experiments confirm that the optimal results will be achieved by choosing the misclassification costs (C_1 and C_2) based on the imbalance ratio. Differences greater than this will not improve the performance and on the other hand, the differences lower than this, may find worse decision boundary and degrade the performance.

Ground-truth tracks are employed to group captured ROIs over frames to create a trajectory in time-based analysis. To that end, captured ROIs of each individual in the scene are grouped and processed separately. In this regard, the decision system accumulates the scores over a window of fixed size 30 frames (1 second) to achieve the highest peaks of accumulated scores in order to fulfill FR at trajectory-level using fusion at frame-level.

Ensemble of TMs [6], ensemble of OC-SVMs, SVDL [53], and ESRC-DA [40] are considered as the baseline and state-of-the-art FR systems to validate the proposed framework. In kNN experiments, eigenfaces of ROIs [55] are employed to compute the specialized kNN adapted for VS (VSkNN) based on k equals to 3 (1 target still and 2 nearest non-targets captured from background model) [41]. To that end, the distance of the probe face are calculated from the target watch-list still along with the distance from the 2 nearest non-target captures from the training set. Thus, VSkNN score S_{VSkNN} is obtained as follows:

$$S_{VSkNN} = \frac{d_T}{d_T + d_{NT_1} + d_{NT_2}} \tag{8}$$

where d_T is the distance of the probe face from the target still, d_{NT_1} and d_{NT_2} are the distances from the nearest non-target captures, respectively.

Libsvm is also exploited in order to train OC-SVMs, where the regularization parameter n sets to 0.01 that indicates 1% of the non-target training data can be considered as support vectors. In SVDL experiment, 5 high-quality stills belonging to 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 according to the default values defined in SVDL. The number of dictionary atoms are initialized to 80 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.

4.3 Performance Metrics

The performance of watch-list screening systems are evaluated at the transaction-level by the Receiver Operating Characteristic (ROC) curve, in which TPRs are plotted as a function of FPRs over all threshold values. The proportion of target ROIs that correctly detected as individual of interest over the total number of target ROIs in the sequence is counted as True Positive Rate (TPR). Meanwhile, False Positive Rate (FPR) is the proportion of non-target ROIs detected as individual of interest over the total number of non-target ROIs. 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 order to estimate the performance of the system based on each sample (target and non-target), another curve called precision-recall (PR) can be used to characterize the performance of targets. It is suitable to measure the performance under the imbalanced data situation during operation. Recall can be considered as TPR and precision (PR) is computed as follows $\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$. TPR and PR can be combined locally in a well-known scalar value called F1 measure $\mathrm{F}_1 = 2\frac{\mathrm{TPR.PR}}{\mathrm{TPR+PR}}$.

In transaction-level analysis, performance of the chosen feature extraction techniques among patches and different levels of fusion are 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 suitable 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 along with standard deviations are reported.

4.4 Results and Discussion

The performance of different aspects of the proposed framework using different feature extraction techniques with feature- and score-level fusion among patches is shown in Table 3 and Table 4 for Chokepoint and COX-S2V videos. Experiments are provided for non-overlapping patch configurations with 1, 4, 9, and 16 blocks (48x48, 24x24, 16x16, and 12x12 pixels, respectively). The scores of SVM classifiers trained over each patch are combined to provide the final score for each representation using score averaging. Noted that dimension of the representations vary, for instance the dimension of HOG and Haar depends on the resolution of the image and they typically produce a longer feature vector. Due to complexity and to avoid over-fitting, the number of dimensions are also reduced using PCA. An example of the ROC and inverted-PR curves obtained using ensemble of e-SVMs (4 blocks and HOG descriptor) is shown in Figure 7 with P1E_S1_C1 videos of Chokepoint.

The average values of pAUC(20%) and AUPR along with standard errors are presented in the Table 3 and Table 4 for different patch configurations.

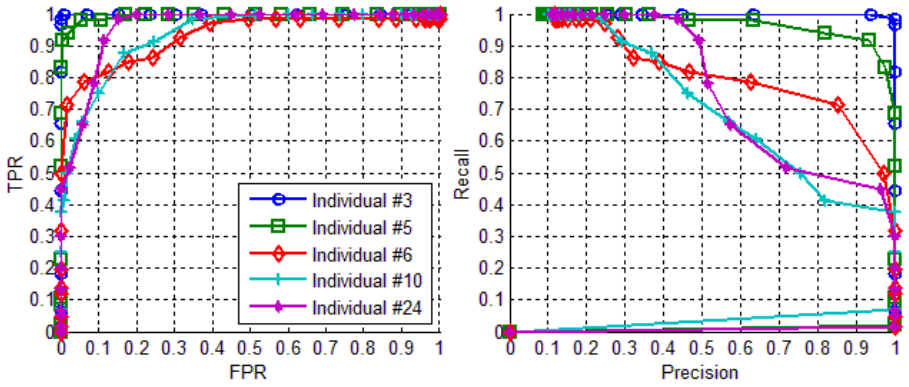


Fig. 7: ROC and inverted-PR curves for a randomly selected watch-list of 5 individuals with Chokepoint video P1E_S1_C1.

Table 3: Average pAUC(20%) and AUPR accuracy of proposed systems at the transaction-level using feature extraction techniques (w/o patches) and videos of the Chokepoint dataset.

ROI - Patch Configurations	**Face Representations**							
	LBP (59 features)		**LPQ (256 features)**		**HOG (500 features)**		**Haar (2304 features)**	
	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR
1 block (48x48 pixels)								
All features	77.86±2.53	72.12±7.18	83.60±2.72	79.98±6.84	91.50±2.30	88.46±4.18	78.20±2.62	76.56±8.16
PCA features (max 64)	77.86±2.53	72.12±7.18	77.93±1.80	69.13±7.10	86.08±1.70	81.71±6.34	71.12±3.08	67.54±8.92
4 blocks (24x24 pixels)								
Score-level	79.53±2.34	74.71±8.76	79.20±2.66	76.65±8.40	91.03±0.84	88.02±4.32	84.41±2.38	81.82±7.42
Feature-level	78.00±2.76	75.16±6.50	80.00±2.46	76.24±4.28	79.50±3.10	77.36±7.18	72.44±2.68	69.80±4.00
9 blocks (16x16 pixels)								
Score-level	81.68±2.04	77.38±6.37	85.03±1.12	82.18±6.90	98.44±0.78	96.64±2.12	82.50±1.16	80.46±6.20
Feature-level	51.70±2.82	48.92±6.14	80.90±3.22	79.14±7.72	77.60±2.24	74.38±4.24	80.00±3.06	77.62±4.68
16 blocks (12x12 pixels)								
Score-level	33.60±2.32	32.78±2.82	52.70±2.24	49.70±4.42	65.30±3.04	61.12±6.62	70.00±2.40	68.82±7.28
Feature-level	30.50±1.24	28.82±6.00	35.00±2.40	32.78±4.96	71.10±3.54	69.78±4.16	70.56±3.38	67.28±4.94

As shown in Table 3, using patch-based method with 4, and 9 blocks (24x24 and 16x16 pixels, respectively) outperforms cases without patches (1 block). Patches with 16x16 pixels significantly outperforms case with large patches, and HOG in most cases provides better performance, especially when 9 blocks are used. The performance obtained using the smaller patches (12x12 pixels) is substantially lower, because features extracted from these small sub-images are not discriminant enough to generate robust classifier ensembles.

Feature-level fusion is also performed, where features extracted from patches are concatenated into a long feature vector and only one classifier is trained per each ROI representation. To reduce complexity, the dimension of features extracted from each patch is first reduced using PCA and then they are concatenated into the higher dimen-

sional vector (for PCA projection, the 64 first eigenvectors are selected as features for LPQ, HOG and Haar descriptors). Concatenating features from larger blocks mostly provides higher performance. Longer ROI patterns obtained from more patches with smaller size may not perform well due to less of discriminative eigenvectors after applying PCA. However, training a separate classifier for each patch and combining local SVMs at a score-level typically achieves better performance, compared to training one global SVM based on the concatenated features extracted from all of the patches (feature-level fusion).

The experiments conducted over COX-S2V videos (Table 4) also suggest that the score-level fusion of patches can yield a better performance in contrast to the feature-level fusion of patches, due to encoding the pixels within each local patch into a different classifier separately.

Table 4: Average pAUC(20%) and AUPR accuracy of proposed systems at the transaction-level using feature extraction techniques (w/o patches) and videos of the COX-S2V dataset.

ROI - Patch Configurations	**Face Representations**							
	LBP (59 features)		**LPQ (256 features)**		**HOG (500 features)**		**Haar (2304 features)**	
	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR
1 block (48x48 pixels)								
All features	85.86±0.64	75.03±0.89	91.31±0.65	76.08±1.24	97.95±0.70	77.54±1.54	97.46±0.50	76.08±1.24
PCA features (max 64)	85.86±0.64	75.03±0.89	91.03±1.12	79.04±1.52	91.31±1.30	75.12±3.02	89.54±1.94	72.53±1.97
4 blocks (24x24 pixels)								
Score-level	92.31±1.14	80.38±1.30	95.40±1.14	84.26±1.46	98.47±1.58	86.70±1.80	97.70±0.47	80.73±1.72
Feature-level	84.59±1.10	73.00±1.58	89.64±0.26	82.00±0.73	88.08±2.30	68.30±1.86	89.90±1.32	78.04±0.89
9 blocks (16x16 pixels)								
Score-level	96.88±1.78	82.15±1.81	94.13±0.48	85.72±0.70	98.37±0.65	87.35±1.02	97.08±0.44	77.40±1.60
Feature-level	74.78±2.28	55.38±2.43	87.12±0.59	77.68±0.97	86.80±2.35	63.07±2.30	89.42±1.25	73.16±0.80
16 blocks (12x12 pixels)								
Score-level	76.10±2.28	49.98±3.37	86.62±0.82	75.60±0.92	92.95±1.06	80.01±1.46	92.96±0.38	76.77±1.42
Feature-level	69.93±1.05	48.44±3.85	80.72±1.05	69.98±0.74	91.64±1.47	75.02±2.14	93.68±0.48	64.63±2.15

Since each feature extraction technique performs inconstantly, applying fusion among them with dynamic classifier selection can provide higher level of performance. Table 5 presents a performance comparison for ensemble of classifiers designed with e-SVMs and OC-SVMs using feature- and score-level fusion of descriptors. The results of the proposed framework are also compared against baseline and state-of-the-art systems: VSkNN [41], SVDL [53], ESRC-DA [40], and ensemble of TMs [6] with the Chokepoint data. The performance achieved by combining the descriptors within static and dynamic ensembles at feature-level (concatenation) and score-level (mean function).

Using fusion of descriptors within the ensemble significantly improves performance over individual feature extraction techniques either with or without patches at transaction-level. Results indicate that the score-level fusion outperforms feature-level fusion, and 1 block (48x48 pixels) performs worse than other patch configurations. Feature-level fusion provides lower performance due to the effects of dimension of the concatenated vectors and the training of only one global SVM classifier. Accordingly, accurate lo-

cal SVM classifiers leads to a robust ensembles for face screening, where patches size 16x16 pixels performs slightly better.

It can be seen from Table 5 that ensemble of e-SVMs outperforms ensemble of OC-SVMs, ensemble of TMs, VSkNN, SVDL and ESRC-DA. Performance of the FR system using VSkNN and SVDL is poor, mostly because of the significant 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, imbalance of target versus non-target individuals observed during operation. 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 positives 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 operational phase.

Table 5: Average pAUC(20%) and AUPR performance of different implementations of the proposed framework at the transaction-level over all Chokepoint videos. Results are shown using feature-, score-level fusion of patches and descriptors within ensembles against reference state-of-the-art systems.

FR Systems		pAUC(20%)	AUPR
VSkNN [41]		19.00±0.40	16.48±0.90
SVDL [53]		74.91±4.03	65.09±4.82
ESRC-DA [40]		97.16±1.28	76.97±6.73
Ensemble of TMs [6]		85.60±1.04	82.78±7.06
Ensemble of OC-SVMs			
1 block (48x48 pixels)	**Feature-level**	71.34±5.78	64.07±5.96
	Score-level	86.10±1.06	81.62±7.82
4 blocks (24x24 pixels)	**Feature-level**	86.24±0.45	84.48±0.61
	Score-level	89.40±2.42	88.02±6.20
9 blocks (16x16 pixels)	**Feature-level**	96.73±0.43	91.55±4.43
	Score-level	97.40±0.40	95.72±2.64
16 blocks (12x12 pixels)	**Feature-level**	86.15±1.92	83.80±4.05
	Score-level	88.20±1.10	84.66±2.92
Dynamic Ensemble of OC-SVMs			
4 blocks (24x24 pixels)	**Score-level**	98.10±0.48	96.14±0.76
9 blocks (16x16 pixels)	**Score-level**	98.42±0.86	96.47±1.24
16 blocks (12x12 pixels)	**Score-level**	95.43±0.66	92.96±1.75
Ensemble of e-SVMs			
1 block (48x48 pixels)	**Feature-level**	92.90±0.82	90.20±5.06
	Score-level	92.28±0.54	90.95±2.84
4 blocks (24x24 pixels)	**Feature-level**	94.40±0.74	91.98±5.52
	Score-level	98.58±0.40	97.34±1.82
9 blocks (16x16 pixels)	**Feature-level**	89.80±0.12	89.24±0.44
	Score-level	100±0.00	99.24±0.38
16 blocks (12x12 pixels)	**Feature-level**	88.40±0.70	86.44±3.60
	Score-level	95.30±0.92	93.86±2.28
Dynamic Ensemble of e-SVMs			
4 blocks (24x24 pixels)	**Score-level**	100±0.00	98.86±0.90
9 blocks (16x16 pixels)	**Score-level**	100±0.00	99.31±0.46
16 blocks (12x12 pixels)	**Score-level**	97.71±1.06	94.60±3.12

Results indicate that, OC-SVM classifiers cannot classify target ROI patterns as discriminantly as e-SVM classifiers, because the target reference is not considered during training. Since the model (decision boundary) learned by OC-SVM is only based on low-quality non-target ROIs, and the quality of probe target ROIs are also similar to the training data, this model may fail to classify target ROIs precisely. In terms of number of blocks, ensemble of OC-SVMs using 9 blocks provides higher performance than others at score-level fusion. The proposed dynamic ensemble selection method is also assessed using 4, 9, and 16 blocks. The bottom of Table 5 shows that dynamic selection can improve accuracy and efficiency during operation by combining a lower number of classifiers. It slightly provides better results, where basically the larger the number of classifiers in the pool, the better the results achieved.

To validate the results, the aforementioned experiments are also repeated using the challenging COX-S2V dataset, where only 25 ROIs of target individual captured during operation are matched against 2500 ROIs of non-target individuals. Results compare the state-of-the-art and baseline systems, and dynamic selection of OC-SVM and e-SVM classifiers. Table 6 presents the average transaction-level performance of systems over the COX-S2V data.

The results observed from Table 6 also confirm that ensembles of e-SVMs yield the more promising performance. Results are convincing even with high ratio of imbalances during operation. The dynamic classifier selection method improves the performance and can provide higher accuracy for both OC-SVM and e-SVM ensembles.

To estimate the performance of different fusion approaches at a certain point of FPR, specific decision thresholds are applied to achieve desired FPR values. As illustrated in Figure 4 (c-e), only one threshold is defined for either feature- or score-level fusion, while decision thresholds dedicated to decision-level fusion (see Figure 4 (e)) are determined separately for each descriptor and the global decision is achieved through majority voting. The average performance of the system considering feature-, score-, and decision-level fusions at an exact point of FPR for both Chokepoint and COX-S2V datasets is presented in Table 7 and Table 8, respectively.

As shown in Table 7, decision-level fusion using 1 block (48x48 pixels) performs better than others in terms of F1 measures. Using 4 blocks (24x24 pixels) provides appropriate performance according to either F1 or TPR, retain FPR less than 1%. FPRs in feature- and score-level fusion are mostly greater than 1% due to inaccurate decision thresholds. Although defining decision thresholds per descriptor and using majority vote may lead to lower FPR, results after applying decision thresholds are generally poor. Defining decision thresholds in such an application may be a challenging task that affects overall system accuracy.

The results shown in Table 8 indicate that defining a dedicated threshold for each face descriptor at decision-level fusion using majority voting can achieve a higher accuracy in terms of F1 measure, mostly due to the lower values of FPR.

In another experiment, the proposed system is evaluated when a subset of background model is used during enrollment. To that end, videos captured from only one camera is considered to generate e-SVM ensembles and the system is tested on videos captured from other cameras. Table 9 presents the average performance of the proposed dynamic ensemble of e-SVMs using score-level fusion of descriptors with 9 blocks

Table 6: Average pAUC(20%) and AUPR performance of different implementations of the proposed framework at the transaction-level over COX-S2V videos. Results are shown using feature-, score-level fusion of patches and descriptors within ensembles against reference state-of-the-art systems.

FR Systems		pAUC(20%)	AUPR
VSkNN [41]		56.80±4.02	26.68±3.58
SVDL [53]		69.93±5.67	44.09±6.29
ESRC-DA [40]		99.00±1.13	63.21±4.56
Ensemble of TMs [6]		84.00±0.86	73.36±9.82
Ensemble of OC-SVMs			
1 block (48x48 pixels)	**Feature-level**	82.98±0.98	71.66±0.96
	Score-level	89.58±1.40	77.76±1.36
4 blocks (24x24 pixels)	**Feature-level**	84.94±1.13	75.84±1.62
	Score-level	90.04±0.88	82.61±0.68
9 blocks (16x16 pixels)	**Feature-level**	88.54±0.60	76.62±1.02
	Score-level	91.10±2.20	80.82±5.94
16 blocks (12x12 pixels)	**Feature-level**	83.91±0.83	74.94±1.26
	Score-level	89.28±1.44	79.96±1.08
Dynamic Ensemble of OC-SVMs			
4 blocks (24x24 pixels)	**Score-level**	94.00±1.78	86.72±1.94
9 blocks (16x16 pixels)	**Score-level**	95.78±0.52	87.48±4.06
16 blocks (12x12 pixels)	**Score-level**	95.58±1.15	87.65±1.72
Ensemble of e-SVMs			
1 block (48x48 pixels)	**Feature-level**	89.94±0.29	84.32±1.30
	Score-level	97.95±0.70	87.54±1.54
4 blocks (24x24 pixels)	**Feature-level**	91.12±1.18	83.24±1.56
	Score-level	99.74±0.06	90.21±0.56
9 blocks (16x16 pixels)	**Feature-level**	99.38±0.26	88.32±1.07
	Score-level	100±0.00	91.20±1.52
16 blocks (12x12 pixels)	**Feature-level**	96.62±0.76	80.60±1.50
	Score-level	98.47±0.32	87.48±1.02
Dynamic Ensemble of e-SVMs			
4 blocks (24x24 pixels)	**Score-level**	100±0.00	92.01±0.92
9 blocks (16x16 pixels)	**Score-level**	100±0.00	92.94±1.96
16 blocks (12x12 pixels)	**Score-level**	99.99±0.01	89.28±2.14

over COX-S2V videos, where for example videos captured from camera1 are used to train e-SVMs and the system is assessed on other videos captured from other cameras (camera2 to camera4).

As shown in Table 9, considering a subset of background model during training e-SVMs can drastically reduce the performance in comparison with the results presented in Table 6. Since video2 and video4 are captured using a higher quality camera, better ensembles can be thus generated and subsequently, the performance of the system for other videos (video1 and video3) are relatively higher.

To analyze the impact of considering different number of unknown persons appearing in the operational scene, the number of unknown persons along with the target individual is varied from 100 to 300 and the AUPR performance is measured as displayed in Figure 8.

Table 7: Average performance of proposed system over Chokepoint videos at a certain point of FPR=1% using feature-, score, and decision-level fusion of descriptors within the ensemble.

Number of blocks	Feature-level			Score-level			Decision-level		
	TPR	FPR	F1	TPR	FPR	F1	TPR	FPR	F1
1 (48x48)	41.49±0.16	0.62±0.06	52.08±0.17	67.85±0.18	8.12±0.05	56.13±0.15	70.39±0.17	8.63±0.07	64.78±0.21
4 (24x24)	48.17±0.21	1.89±0.28	54.35±0.18	56.52±0.27	4.86±0.01	49.23±0.26	44.37±0.26	0.43±0.01	58.29±0.27
9 (16x16)	37.07±0.19	2.24±0.16	34.53±0.17	35.69±0.23	0.06±0.01	40.03±0.24	31.81±0.23	0.35±0.03	41.10±0.24
16 (12x12)	32.43±0.06	1.22±0.15	35.08±0.07	37.83±0.46	3.64±0.08	34.22±0.17	27.42±0.68	2.18±0.15	32.58±0.16

Table 8: Average performance of proposed system over COX-S2V videos at a certain point of FPR=1% using feature-, score, and decision-level fusion of descriptors within the ensemble.

Number of blocks	Feature-level			Score-level			Decision-level		
	TPR	FPR	F1	TPR	FPR	F1	TPR	FPR	F1
1 (48x48)	45.15±4.36	0.52±0.07	54.84±2.55	69.05±2.02	0.28±0.03	67.48±1.25	58.55±1.80	0.00±0.00	69.81±1.54
4 (24x24)	54.45±1.35	0.04±0.01	58.92±1.07	84.95±2.22	0.07±0.02	75.84±1.98	87.85±0.62	1.52±0.06	78.71±0.37
9 (16x16)	67.75±2.28	0.00±0.00	75.91±2.03	86.75±1.70	0.60±0.92	73.88±2.89	82.95±1.56	0.02±0.01	81.70±1.48
16 (12x12)	43.10±2.20	0.70±0.14	45.31±1.46	44.48±3.60	4.42±0.30	35.63±0.10	66.40±2.40	4.36±2.37	53.04±0.32

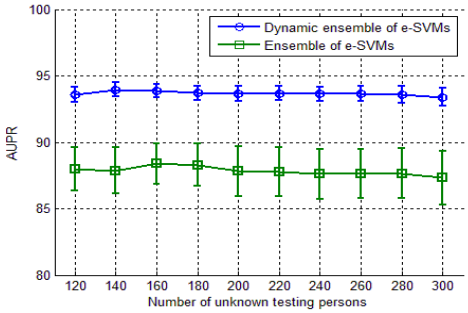


Fig. 8: The analysis of system performance using different number of unknown persons during operation over COX-S2V.

Since the proposed system is comprised of individual-specific ensembles that each one seeks to detect one target individual within the watch-list, as illustrated in Figure 8, it can perform consistently even with observation of severely imbalanced unknown persons during operation.

The proposed ensemble of e-SVMs is also compared against ensemble of TMs as a baseline system at the trajectory-level. In this regards, the scores of individual-specific ensembles are gradually accumulated over a window of consecutive frames using a trajectory defined by the tracker. An example of accumulated scores over the trajectory

Table 9: Average performance of the proposed system over COX-S2V videos, where a subset of background model is used for training.

Background model	**Video1**		**Video2**		**Video3**		**Video4**	
FR Systems	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR
Ensemble of e-SVMs	84.08±1.83	59.40±2.47	74.20±2.22	47.75±2.49	84.18±1.90	59.16±2.84	73.50±2.48	44.04±2.90
Dynamic Ensemble of e-SVMs	92.91±1.16	77.64±2.18	91.05±1.38	75.78±2.09	96.48±1.00	81.66±2.67	91.43±1.18	76.00±1.92

is shown in Figure 9. In this experiment, P1E_S1_C1 video of Chokepoint is employed, where individual ID#03 is considered as watch-list target individual.

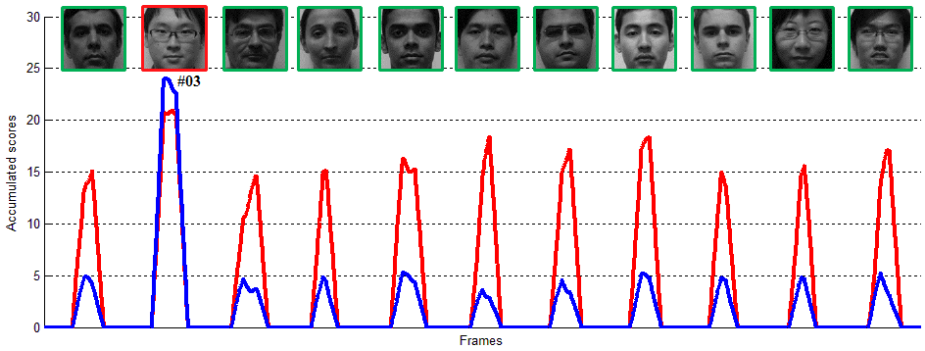


Fig. 9: An example of the scores accumulated over windows of 30 frames with Chokepoint P1E_S1_C1 video using score-level fusion of descriptors with 4 blocks. Ensemble of e-SVMs is the blue curves and ensemble of TMs is the red curves.

As shown in Figure 9, the accumulated scores for target individual (ID#03) is significantly higher than all non-targets individuals for ensemble of e-SVMs, while the accumulated scores of non-targets are greater for ensemble of TMs. It can be observed that the accumulated scores of some non-target individuals are high, due to a higher number of false alarms. To assess the overall performance, the corresponding ROC curve may then plotted for each individual by varying the thresholds from 0 to 30 over accumulated scores, and the AUC are computed as overall performance of ensemble of e-SVMs. The average AUC for each watch-list individual across Chokepoint videos are provided in Table 10.

Table 10 shows the average spatio-temporal recognition performance of the ensemble of e-SVMs is robust and higher than the baseline system.

Another example of trajectory-based analysis over COX-S2V videos is demonstrated in Figure 10, where the individual #001 is the watch-list person.

It can be concluded from Figure 10 that ensemble of e-SVMs can outperform the baseline system under a severe imbalanced operational situation, where the target individual must be detected among more than a hundred people. Thus, the average spatio-

Table 10: AUC accuracy at the trajectory-level for ensemble of e-SVMs and TMs for a random selection of 5 watch-list individuals in the Chokepoint data.

	Individuals of interest					
	ID#03	**ID#05**	**ID#06**	**ID#10**	**ID#24**	**Average**
Ensemble of TMs [6]	93.80±4.80	83.80±8.30	88.80±5.60	86.30±6.60	92.50±6.00	**89.04±6.26**
Ensemble of e-SVMs	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	100.00±0.00	**100.00±0.00**

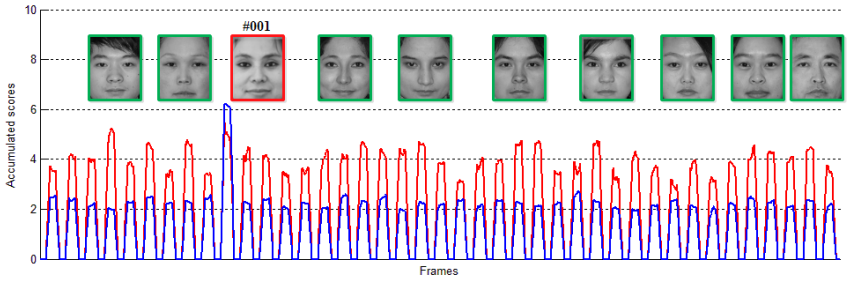


Fig. 10: An example of the scores accumulated over windows of 10 frames with COX-S2V videos. Ensemble of e-SVMs is the blue curves and ensemble of TMs is the red curves.

temporal performance of the proposed and the baseline systems over COX-S2V videos are 100.00±0.00 and 86.01±2.36, respectively.

5 Conclusion

This paper presents a robust multi-classifier system for still-to-video FR that is specialized for watch-list screening applications with a SSPP. Due to the limited number of reference samples per each individual of interest, several feature extraction techniques and patches are employed to generate a diversified ensemble of SVM classifiers per target individual. Feature extraction techniques are chosen based on their robustness against variety of nuisance factors encountered in video surveillance environments. Accordingly, results suggest that using multiple robust face representations for design of facial models favorably allows to build a robust watch-list screening system. Estimating parameters of the classifier with a limited number of target references typically lead to poor generalization and over-fitting. Furthermore, selecting representative non-target samples is needed to optimize performance to overcome the imbalanced data issue, and feature selection. To achieve higher performance, an intuitive dynamic ensemble selection method was proposed to select the most suitable classifiers related to different capture conditions. Extensive experiments show that ensemble with e-SVM classifiers trained using patches isolated in ROI patterns provide solid performance.

In this paper, comprehensive results are presented to observe the performance of the proposed watch-list screening system. Simulation results with the Chokepoint and COX-S2V video datasets indicate that the integration of patches-based approach and face descriptors into an individual-specific ensemble of e-SVM classifiers provides a significantly higher level of performance than either any single representation, or baseline system containing ensemble of OC-SVMs and template matchers. Results also demonstrate that training a separate classifier for each patch and combining their scores outperforms a single classifier trained using a long feature vector of concatenated patches. Since there is only a single reference still per individual of interest during design, videos of unknown non-target individuals in the scene have been used to train the representation-specific classifiers and for normalization to exploit the information of operational phase. Hence, videos of background model are more representative of real scene, contrary to other stills in the cohort model. Since it is difficult to set appropriate thresholds, the scores produced at transaction-level should be considered to obtain the best results. Finally, accumulating ensemble scores over multiple face captures of corresponding individuals using a high-quality track that are provided by the face tracker significantly improves the overall performance.

It was assumed in this paper that reference mugshots are high-quality facial images captured under the same controlled conditions. Future research would involve validating with watch-list individuals having lower quality face images captured with different cameras and conditions.

Acknowledgment

This work was supported by the Fonds de recherche du Québec - Nature et technologies.

References

1. Abate, A.F., Nappi, M., Riccio, D., Sabatino, G.: 2d and 3d face recognition: A survey. Pattern Recognition Letters 28(14), 1885–1906 (2007) 7
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 28(12), 2037–2041 (2006) 3, 8, 14
3. Ahonen, T., Rahtu, E., Ojansivu, V., Heikkila, J.: Recognition of blurred faces using local phase quantization. In: Pattern Recognition, 2008. ICPR 2008. 19th International Conference on. pp. 1–4. IEEE (2008) 3, 8, 14
4. Amira, A., Farrell, P.: An automatic face recognition system based on wavelet transforms. In: Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on. pp. 6252–6255. IEEE (2005) 14
5. Barr, J.R., Bowyer, K.W., Flynn, P.J., Biswas, S.: Face recognition from video: A review. International Journal of Pattern Recognition and Artificial Intelligence 26(05), 1266002 (2012) 2, 6
6. Bashbaghi, S., Granger, E., Sabourin, R., Bilodeau, G.A.: Watch-list screening using ensembles based on multiple face representations. In: Pattern Recognition (ICPR), 2014 22nd International Conference on. pp. 4489–4494 (Aug 2014) 3, 4, 7, 9, 13, 15, 17, 21, 24, 25, 27, 30

7. Batuwita, R., Palade, V.: Fsvm-cil: fuzzy support vector machines for class imbalance learning. Fuzzy Systems, IEEE Transactions on 18(3), 558–571 (2010) 10
8. Bengio, S., Mariéthoz, J.: Biometric person authentication is a multiple classifier problem. In: Multiple Classifier Systems, pp. 513–522. Springer (2007) 5
9. Bereta, M., Pedrycz, W., Reformat, M.: Local descriptors and similarity measures for frontal face recognition: A comparative analysis. Journal of Visual Communication and Image Representation 24(8), 1213–1231 (2013) 3, 9
10. Best-Rowden, L., Klare, B., Klontz, J., Jain, A.K.: Video-to-video face matching: Establishing a baseline for unconstrained face recognition. In: Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on. pp. 1–8. IEEE (2013) 6
11. Beveridge, J.R., Phillips, P.J., Bolme, D.S., Draper, B.A., Givens, G.H., Lui, Y.M., Teli, M.N., Zhang, H., Scruggs, W.T., Bowyer, K.W., Flynn, P.J., Cheng, S.: The challenge of face recognition from digital point-and-shoot cameras. In: Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth International Conference on. pp. 1–8 (2013) 18
12. Chang, C.C., Lin, C.J.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2(3), 27 (2011) 21
13. Chellappa, R., Sinha, P., Phillips, P.J.: Face recognition by computers and humans. Computer 43(2), 46–55 (2010) 2, 5
14. Connaughton, R., Bowyer, K.W., Flynn, P.J.: Fusion of face and iris biometrics. In: Handbook of Iris Recognition, pp. 219–237. Springer (2013) 16
15. Deng, W., Hu, J., Guo, J.: Extended src: Undersampled face recognition via intraclass variant dictionary. PAMI, IEEE Trans on 34(9), 1864–1870 (2012) 6, 7
16. Déniz, O., Bueno, G., Salido, J., De la Torre, F.: Face recognition using histograms of oriented gradients. Pattern Recognition Letters 32(12), 1598–1603 (2011) 3, 8, 14
17. Dreuw, P., Steingrube, P., Hanselmann, H., Ney, H., Aachen, G.: Surf-face: Face recognition under viewpoint consistency constraints. In: BMVC. pp. 1–11 (2009) 8
18. Ekenel, H.K., Stallkamp, J., Stiefelhagen, R.: A video-based door monitoring system using local appearance-based face models. Computer Vision and Image Understanding 114(5), 596–608 (2010) 5
19. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42(4), 463–484 (2012) 3, 9
20. Granger, E., Khreich, W., Sabourin, R., Gorodnichy, D.O.: Fusion of biometric systems using boolean combination: an application to iris–based authentication. International Journal of Biometrics 4(3), 291–315 (2012) 3, 5
21. He, X., Yan, S., Hu, Y., Zhang, H.J.: Learning a locality preserving subspace for visual recognition. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. pp. 385–392. IEEE (2003) 7
22. Huang, Z., Shan, S., Zhang, H., Lao, S., Kuerban, A., Chen, X.: Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset. In: Computer Vision–ACCV 2012, pp. 589–600. Springer (2013) 4, 6, 18
23. Huang, Z., Zhao, X., Shan, S., Wang, R., Chen, X.: Coupling alignments with recognition for still-to-video face recognition. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 3296–3303. IEEE (2013) 6
24. Imam, T., Ting, K.M., Kamruzzaman, J.: z-svm: an svm for improved classification of imbalanced data. In: AI 2006: Advances in Artificial Intelligence, pp. 264–273. Springer (2006) 10, 15
25. Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. Circuits and Systems for Video Technology, IEEE Transactions on 14(1), 4–20 (2004) 2

26. Juneja, M., Vedaldi, A., Jawahar, C., Zisserman, A.: Blocks that shout: Distinctive parts for scene classification. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. pp. 923–930 (June 2013) 11
27. Kamgar-Parsi, B., Lawson, W.: Toward development of a face recognition system for watchlist surveillance. Pattern Analysis and Machine Intelligence, IEEE Transactions on 33(10), 1925–1937 (2011) 6, 7
28. Kan, M., Shan, S., Su, Y., Xu, D., Chen, X.: Adaptive discriminant learning for face recognition. Pattern Recognition 46(9), 2497–2509 (2013) 2, 7
29. Kemmler, M., Rodner, E., Wacker, E.S., Denzler, J.: One-class classification with gaussian processes. Pattern Recognition 46(12), 3507 – 3518 (2013) 9
30. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Burge, M., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1931–1939 (2015) 18
31. Krawczyk, B., Wozniak, M.: Diversity measures for one-class classifier ensembles. Neurocomputing 126(0), 36 – 44 (2014) 10
32. Li, Q., Yang, B., Li, Y., Deng, N., Jing, L.: Constructing support vector machine ensemble with segmentation for imbalanced datasets. Neural Computing and Applications 22(1), 249–256 (2013) 15
33. Li, Y., Shen, W., Shi, X., Zhang, Z.: Ensemble of randomized linear discriminant analysis for face recognition with single sample per person. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. pp. 1–8. IEEE (2013) 3, 9
34. Liao, S., Jain, A.K., Li, S.Z.: Partial face recognition: Alignment-free approach. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(5), 1193–1205 (2013) 3, 8, 9
35. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. Image processing, IEEE Transactions on 11(4), 467–476 (2002) 8
36. Lu, J., Tan, Y.P., Wang, G.: Discriminative multimanifold analysis for face recognition from a single training sample per person. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(1), 39–51 (2013) 8
37. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 89–96. IEEE (2011) 3, 11, 15
38. Misra, I., Shrivastava, A., Hebert, M.: Data-driven exemplar model selection. In: Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on. pp. 339–346 (March 2014) 11
39. Mokhayeri, F., Granger, E., Bilodeau, G.A.: Synthetic face generation under various operatinnal conditions in video surveillance. In: ICIP (2015) 2, 7
40. Nourbakhsh, F., Granger, E., Fumera, G.: An extended sparse classification framework for domain adaptation in video surveillance. In: ACCV, Workshop on Human Identification for Surveillance (2016) 6, 21, 24, 25, 27
41. Pagano, C., Granger, E., Sabourin, R., Marcialis, G., Roli, F.: Adaptive ensembles for face recognition in changing video surveillance environments. Information Sciences 286, 75–101 (2014) 2, 3, 4, 21, 24, 25, 27
42. Pagano, C., Granger, E., Sabourin, R., Gorodnichy, D.O.: Detector ensembles for face recognition in video surveillance. In: Neural Networks (IJCNN), The 2012 International Joint Conference on. pp. 1–8. IEEE (2012) 4, 5
43. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000) 7

44. Scholkopf, B., Smola, A.J.: Learning with kernels: Support vector machines, regularization, optimization, and beyond. MIT press (2002) 10
45. Shaokang, C., Sandra, M., Mehrtash T, H., Conrad, S., Abbas, B., Brian C, L., et al.: Face recognition from still images to video sequences: A local-feature-based framework. EURASIP journal on image and video processing 2011 (2011) 5
46. Tan, X., Chen, S., Zhou, Z.H., Zhang, F.: Face recognition from a single image per person: A survey. Pattern Recognition 39(9), 1725–1745 (2006) 2, 7
47. De la Torre Gomerra, M., Granger, E., Radtke, P.V., Sabourin, R., Gorodnichy, D.O.: Partially-supervised learning from facial trajectories for face recognition in video surveillance. Information Fusion 24(0), 31–53 (2015) 2, 3, 5, 9
48. Veropoulos, K., Campbell, C., Cristianini, N., et al.: Controlling the sensitivity of support vector machines. In: Proceedings of the international joint conference on artificial intelligence. vol. 1999, pp. 55–60 (1999) 10
49. Viola, P., Jones, M.J.: Robust real-time face detection. International journal of computer vision 57(2), 137–154 (2004) 13
50. Wong, Y., Chen, S., Mau, S., Sanderson, C., Lovell, B.C.: Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on. pp. 74–81. IEEE (2011) 4, 18
51. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(2), 210–227 (2009) 6
52. Xie, C., Kumar, B.V., Palanivel, S., Yegnanarayana, B.: A still-to-video face verification system using advanced correlation filters. In: Biometric Authentication, pp. 102–108. Springer (2004) 5
53. Yang, M., Van Gool, L., Zhang, L.: Sparse variation dictionary learning for face recognition with a single training sample per person. In: ICCV. pp. 689–696 (2013) 4, 6, 21, 24, 25, 27
54. Zeng, Z.Q., Gao, J.: Improving svm classification with imbalance data set. In: Neural Information Processing. pp. 389–398. Springer (2009) 9, 15
55. Zhang, J., Yan, Y., Lades, M.: Face recognition: eigenface, elastic matching, and neural nets. Proceedings of the IEEE 85(9), 1423–1435 (1997) 7, 21
56. Zhang, Y., Wang, D.: A cost-sensitive ensemble method for class-imbalanced datasets. In: Abstract and applied analysis. vol. 2013. Hindawi Publishing Corporation (2013) 9, 10, 21
57. Zhang, Y., Martínez, A.M.: From stills to video: Face recognition using a probabilistic approach. In: Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on. pp. 78–78. IEEE (2004) 5
58. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. Acm Computing Surveys (CSUR) 35(4), 399–458 (2003) 8
59. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Computer Vision and Image Understanding 91(1), 214–245 (2003) 5
60. Zong, W., Huang, G.B.: Face recognition based on extreme learning machine. Neurocomputing 74(16), 2541 – 2551 (2011) 9
61. Zou, J., Ji, Q., Nagy, G.: A comparative study of local matching approach for face recognition. Image Processing, IEEE Transactions on 16(10), 2617–2628 (2007) 8