Dynamic Selection of Exemplar-SVMs for Watch-List Screening Through Domain Adaptation

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]*LITIV Lab, Polytechnique Montréal, Canada*
bashbaghi@livia.etsmtl.ca, {eric.granger, robert.sabourin}@etsmtl.ca, gabilodeau@polymtl.ca

Keywords: Face Recognition, Video Surveillance, Multi-Classifier System, Single Sample Per Person, Random Subspace Method, Domain Adaptation, Dynamic Classifier Selection

Abstract: Still-to-video face recognition (FR) plays an important role in video surveillance, allowing to recognize individuals of interest over a network of video cameras. Watch-list screening is a challenging video surveillance application, because faces captured during enrollment (with still camera) may differ significantly from those captured during operations (with surveillance cameras) under uncontrolled capture conditions (with variations in, e.g., pose, scale, illumination, occlusion, and blur). Moreover, the facial models used for matching are typically designed a priori with a limited number of reference stills. In this paper, a multi-classifier system is proposed that exploits domain adaptation and multiple representations of face captures. An individual-specific ensemble of exemplar-SVM (e-SVM) classifiers is designed to model the single reference still of each target individual, where different random subspaces, patches, and face descriptors are employed to generate a diverse pool of classifiers. To improve robustness of face models, e-SVMs are trained using the limited number of labeled faces in reference stills from the enrollment domain, and an abundance of unlabeled faces in calibration videos from the operational domain. Given the availability of a single reference target still, a specialized distance-based criteria is proposed based on properties of e-SVMs for dynamic selection of the most competent classifiers per probe face. The proposed approach has been compared to reference systems for still-to-video FR on videos from the COX-S2V dataset. Results indicate that ensemble of e-SVMs designed using calibration videos for domain adaptation and dynamic ensemble selection yields a high level of FR accuracy and computational efficiency.

1 INTRODUCTION

In decision support systems for video surveillance, face recognition (FR) is increasingly employed to enhance security in public places, such as airports, subways, etc. FR systems are needed to accurately detect the presence of individuals of interest enrolled to the system over a network of surveillance cameras (De la Torre Gomerra et al., 2015), (Pagano et al., 2014). In still-to-video FR, face models generated based on face stills are matched against faces captured in videos under uncontrolled conditions. Thus, face models are composed of one or very few facial regions of interest (ROIs) isolated in reference face stills for template matching, or a neural and statistical classifier, where the parameters are estimated using reference ROIs (De-la Torre Gomerra et al., 2015).

Watch-list screening is among the most challenging application in video surveillance. Face models are typically designed a priori during enrollment using a single reference still (high-quality mugshot or ID photo) under controlled conditions (Bashbaghi et al., 2014). A key issue in still-to-video FR is that the appearance of ROIs captured with still camera differs significantly from ROIs captured with video cameras due to various nuisance factors, e.g., changes in illumination, pose, blur, and occlusion, and camera interoperability (Barr et al., 2012). The single sample per person (SSPP) problem found in these systems has been addressed by different techniques, such as using multiple face representations, synthetic generation of virtual faces, and incorporating auxiliary sets to enlarge the design data (Bashbaghi et al., 2014), (Mokhayeri et al., 2015), (Yang et al., 2013).

Still-to-video FR systems can be viewed as a domain adaptation (DA) problem, where the distribu-

tion of facial ROIs captured from reference stills in the enrollment domain (ED) are different from those video ROIs captured from multiple surveillance cameras, where each one represents a non-stationary operational domain (OD) (Shekhar et al., 2013). Since any distributional change (either domain shift or concept drift) can degrade performance, DA methods may be deployed to design accurate classification systems that will perform well on the OD given knowledge obtained from the ED (Patel et al., 2015).

State-of-the-art systems for FR in video surveillance are typically designed with individual-specific face detectors (one or 2-class classifiers) that can be easily added, removed, and specialized over time (Pagano et al., 2014), (Bashbaghi et al., 2014). Using an ensemble of classifiers per individual with static selection and fusion of diversified set of base classifiers has been shown to enhance the robustness of still-to-video FR (Bashbaghi et al., 2015). Furthermore, dynamic selection (DS) can be also exploited to select the most suitable classifiers for an input video ROI. DS can be considered as an effective approach in ensemble-based systems, when the training data is limited and imbalanced (Britto et al., 2014). To that end, base classifiers can be selected according to their level of competence to classify under specific capture conditions and individual behaviors within an operational environment (Shekhar et al., 2013).

In this paper, a robust dynamic individual-specific ensemble-based system is proposed for still-to-video FR. Multiple feature subspaces corresponding to different face patches and descriptors are employed to generate a diverse pool of classifiers, and to improve robustness against different perturbation factors frequently observed in real-world surveillance environments. During enrollment, an individual-specific ensemble of e-SVM classifiers is designed for each target individual based on the ED data (the limited number of labeled faces in reference stills) and OD data (an abundance of unlabeled faces captured in calibration videos). Thus, an unsupervised DA method is employed to train e-SVMs in the ED, where unlabeled lower-quality videos of unknown persons are considered to transfer the knowledge of the OD. Three different training schemes are proposed using a single labeled target still along with non-target still ROIs from the cohort, as well as, unlabeled non-target video ROIs captured with surveillance camera.

During operations, a novel distance-based criteria is proposed for DS based on the properties of e-SVMs in order to effectively adapt to the changing uncontrolled capture conditions. Thus, the DS approach performs in the feature space to select the most competent e-SVMs for a given probe ROI based on the distance between support vectors of e-SVMs and a target still for each individual of interest. The performance of the proposed system is compared to state-of-the-art systems using the videos from COX-S2V dataset (Huang et al., 2015).

2 SYSTEMS FOR STILL-TO-VIDEO FR

Still-to-video FR systems attempt to accurately match the faces captured from video surveillance cameras against the corresponding facial models of the individuals of interest registered to the system. Due to generation of discriminative facial models, the SSPP problem has been addressed using techniques to provide multiple face representations (Bashbaghi et al., 2014), (Kamgar-Parsi et al., 2011). For instance, face synthesizing through morphology is used in (Kamgar-Parsi et al., 2011), where a specialized neural network is trained for each individual. Multiple face representations (employing patch configurations and different face descriptors) exploited in an ensemble-based system have shown to significantly improve the overall performance of a basic still-to-video FR at the cost of either several template matchers or multiple classifiers (Bashbaghi et al., 2014), (Bashbaghi et al., 2015). As a specialized classification technique considering the SSPP problem, e-SVM classifier is adapted using non-target video ROIs (Bashbaghi et al., 2016), (Malisiewicz et al., 2011).

Spatio-temporal recognition can be also exploited to enhance the robustness, where decisions are produced through a tracker to regroup ROIs of a same person into trajectories (Dewan et al., 2016). Recently, sparse representation based classification (SRC) methods are adopted to increase robustness to intra-class variation using a generic auxiliary training set, such as sparse variation dictionary learning (SVDL) (Yang et al., 2013). Similarly, an extended sparse representation approach through domain adaptation (ESRC-DA) (Nourbakhsh et al., 2016) has been proposed for still-to-video FR incorporating matrix factorization and dictionary learning. According to the availability of labeled data in the OD, unsupervised DA has been proposed, where it does not consider labeled data in the OD as observed in watch-list screening applications (Qiu et al., 2014). Two unsupervised DA approaches are relevant for still-to-video FR based on the knowledge transferred between the enrollment and operational domains (Patel et al., 2015), (Pan and Yang, 2010). Instance transfer methods attempt to exploit parts of the ED data for learning in the OD. In contrast, feature representation transfer

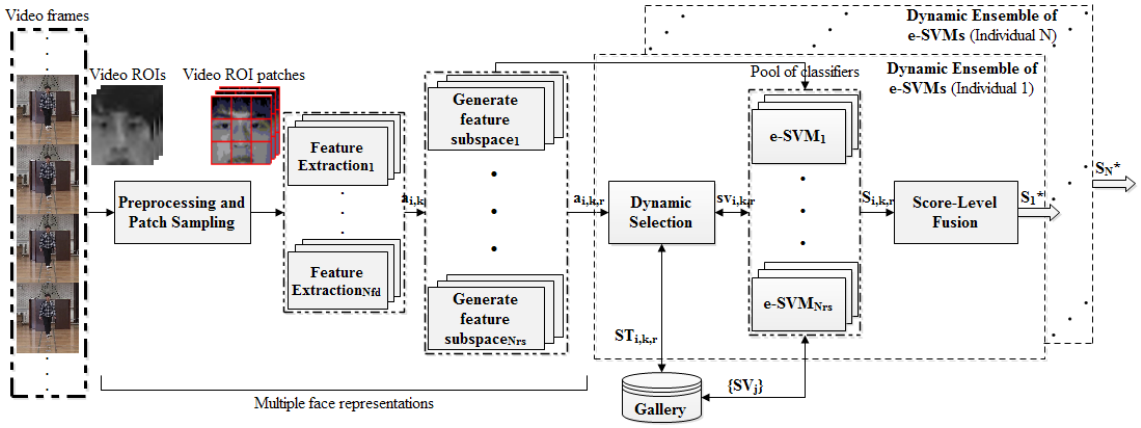


Figure 1: Block diagram of the proposed still-to-video FR system using dynamic ensemble of e-SVMs per target individual.

methods exploit OD data to find a latent feature space that reduces the distribution differences between the ED and the OD (Pan and Yang, 2010).

3 ENSEMBLES OF EXEMPLAR-SVMS THROUGH DOMAIN ADAPTATION

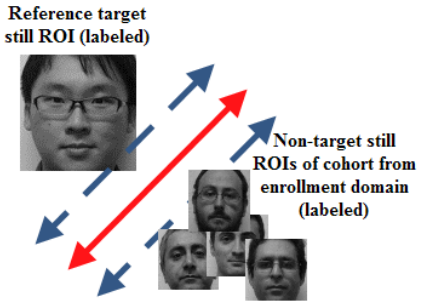
The block diagram of the proposed system is shown in Figure 1. During enrollment, a single reference still of a target individual is employed to train an ensemble of e-SVMs using faces captured in OD and multiple face representations to generate diverse pools of e-SVM classifiers. During operations, the most competent classifiers are selected dynamically for a probe ROI using a new selection criteria according to changes in capture conditions of the OD and combined.

3.1 Enrollment Phase

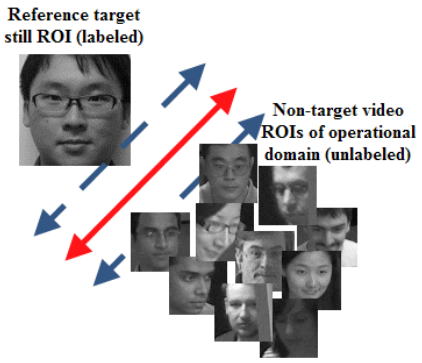
During enrollment of a target individual, a diverse pool of e-SVM classifiers is constructed for each target individual enrolled to the system. In particular, several representations generated from the labeled target still ROI through using different patches, descriptors, and random subspaces. These representations of the target still ROI are used along with the corresponding unlabeled video ROIs of non-target individuals to train e-SVMs. Thus, a pool of $N_p \cdot N_{fd} \cdot N_{rs}$ e-SVMs are trained for each individual of interest and stored in the gallery, where N_p is the number of patches, N_{fd} and N_{rs} are the number of descriptors and random subspaces, respectively.

To generate multiple face representations, random feature subspaces are extracted from patches isolated uniformly without overlapping in each ROI, where patches are represented using several complementary face descriptors, such as LPQ and HOG descriptors (Ahonen et al., 2008), (Deniz et al., 2011). For example, LPQ extract texture features of the face images from frequency domain through Fourier transform and has shown high robustness to motion blur. HOG extract edges using different angles and orientations, where it is more robust to pose and scale changes, as well as, rotation and translation. Random sampling of features extracted from each local patch can provide diversity among classifiers, due to different feature distributions, and exploits information on local structure of faces for FR under changes in pose, illumination, and occlusions.

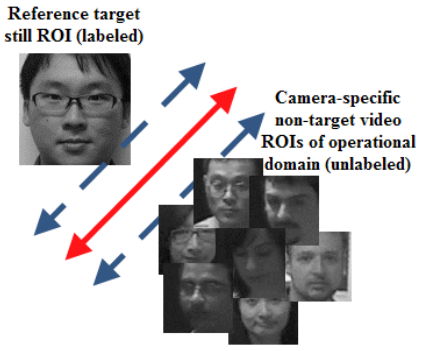
Different training schemes as illustrated in Figure 2 are considered using either the labeled target still ROIs from the cohort or unlabeled non-target video ROIs captured from the calibration videos. To that end, an unsupervised DA approach is considered, where labeled still reference ROIs from the cohort and unlabeled video ROIs captured from the operational environment are employed to train e-SVM classifiers. In the first training scheme (Figure 2 (a)), labeled target still ROIs versus non-target still ROIs from the cohort are employed to train e-SVMs without exploiting unlabeled video ROIs from the OD for DA. The second scheme (Figure 2 (b)) relies on several unlabeled non-target video ROIs from all calibration videos (or background model), while in the third scheme (Figure 2 (c)), unlabeled video ROIs captured from each specific camera are exploited in conjunction with a target still ROI in order to design camera-specific pools of classifiers. Thus, an individual-specific pool of e-SVM classifiers trained with video ROIs of specific camera is employed to recognize individuals from ROIs captured with the corresponding camera.



(a) Training scheme 1



(b) Training scheme 2



(c) Training scheme 3

Figure 2: Illustration of different training schemes for DA with an e-SVM classifier.

To train a classifier under imbalanced data distributions (a single reference labeled still from the ED versus several unlabeled non-target videos from the OD), specialized linear SVM classifiers called e-SVM are adapted (Bashbaghi et al., 2016) for each scheme. Let $\mathbf{a}$ be the labeled target ROI pattern, $\mathbf{x}$ and U are non-target ROI patterns (either labeled still ROIs for scheme 1 or unlabeled video ROIs for schemes 2 and 3) and their number, respectively. The e-SVM is formulated as follows:

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

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. To deal with the imbalanced training data in such a situation and avoid the learning model to bias toward the majority class (unlabeled non-target videos), the regularization term (C_1) of the minority class (a single labeled target reference still) is assigned greater than the regularization term (C_2) of negative samples.

3.2 Operational Phase

During operations, people appear before surveillance cameras (see Figure 1), while each individual-specific ensemble attempts to recognize these faces as an individual of interest. Each frame is segmented to extract facial ROI(s) and then multiple face representations are generated for classification. Then, every ROI is projected into multiple feature subspaces corresponding to classifiers, and those that meet competence criteria are dynamically selected. A given probe ROI is fed to an ensemble of e-SVMs defined through DS. Score-level fusion is adopted to combine the scores of e-SVM classifiers selected from the pool. The operational phase of the proposed system is described in Algorithm 1.

Algorithm 1 Operational phase with DS.

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

As formalized in Algorithm 1, each given probe ROI t is first divided into patches p_i. Feature extraction technique f_k is applied on each patch p_i to form a

ROI pattern $\mathbf{a}_{i,k}$. These patterns are projected into the N_{rs} feature subspaces s_r generated for training e-SVM classifiers and then $\mathbf{a}_{i,k,r}$ is projected into the feature space of the support vectors $\{SV_j\}$ of classifiers C_j and the reference still $\mathbf{ST}_{i,k,r}$ of the target individual j. Finally, those classifiers c_l in C_j that satisfy the levels of competence criteria (line 12) are selected to constitute C_j^* in order to classify the testing sample t, where $\mathbf{sv}_{i,k,r} \in \{SV_j\}$ is the closest support vector to $\mathbf{ST}_{i,k,r}$. Subsequently, the scores of selected classifiers $\mathbf{S}_{i,k,r}$ are combined using score-level fusion to provide final score $\mathbf{S}_\mathbf{j}^*$. However, fusion of all classifiers in C_j is exploited to classify t when none of classifiers fulfill the competence criteria. The calibrated score of e-SVM for the given probe ROI t and the regression parameters (α_a,β_a) is computed as follows (Malisiewicz et al., 2011):

$$f(\mathbf{x}|\mathbf{w},\alpha_a,\beta_a) = \frac{1}{1+e^{-\alpha_a(\mathbf{w}_a^T-\beta_a)}} \quad (2)$$

When a probe ROI is captured, a new DS method is exploited based on e-SVM properties to provide a strong discrimination among probe ROIs. It allows the system to select the subset of classifiers that are the most suitable for the given capture conditions of a given probe ROI. In order to select the most competent classifiers, the proposed internal competence criteria relies on the: (1) distance from the non-target support vectors ROIs, $d(\mathbf{a}_{i,k,r},\mathbf{sv}_{i,k,r})$, and (2) closeness to the target still ROI pattern, $d(\mathbf{a}_{i,k,r},\mathbf{ST}_{i,k,r})$. The key idea is to select the e-SVM classifiers that locate the given probe ROI close to the target support vector, yet far from non-target support vectors. If the distance between the probe and the target still is lower than the distance from support vectors, then those classifiers are dynamically selected as a suitable subset for classifying the probe ROIs.

Classifiers with support vectors that are far from the ROI probes can be also desired candidates, because they may classify them correctly. Distance from non-target support vectors can be defined by considering the closest support vector to the target still ROI in the proposed DS approach (see Figure 3). All the non-target support vectors were sorted a priori based on their distance to the target still (the target support vector) in an offline processing. Then, the closest support vector to the target still is used to compare with the input probe ROIs.

In contrast to the common DS techniques that use local neighborhood accuracy for measuring the level of competence (Britto et al., 2014), it is not mandatory in the proposed DS approach to define neighborhood with a set of validation data, using methods like k-NN. Thus, the proposed criteria exploits the local e-SVM properties, and accounts for the SSPP constraints, where it is efficient in terms of complexity (number of computations to define neighborhood). However, different distance metrics, such as Euclidean can be employed to measure the distances between the probe ROI and either a target still ROI pattern or non-target support vectors.

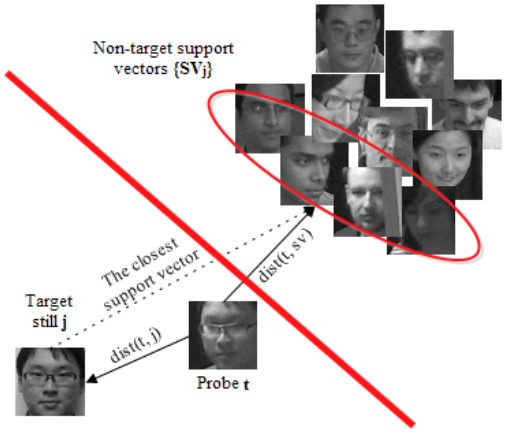


Figure 3: Illustration of the proposed dynamic classifier selection approach in a 2D feature space.

4 EXPERIMENTAL RESULTS

4.1 Methodology for Validation

In this paper, two aspects of the proposed system are assessed experimentally using a real-world video surveillance data. First, different e-SVM training schemes are compared for the proposed individual-specific ensembles. Second, the impact of applying DS is analyzed on the performance. Experiments in this paper are shown at transaction-level to perform face classification[1].

A challenging still-to-video dataset called COX-S2V[2] (Huang et al., 2015) is employed to evaluate performance of the proposed and baseline systems. This dataset consists of 1000 subjects, where each subject has a high-quality still image captured under controlled condition, and four lower-quality facial trajectories captured under uncontrolled conditions using two different off-the-shelf camcorders. Each trajectory has 25 faces (16x20 and 48x60 resolutions), where ROIs taken from these videos encounter changes in illumination, expression, scale, viewpoint,

[1] In still-to-video FR system, operational ROI would be regrouped into trajectories for spatio-temporal recognition

[2] http://vipl.ict.ac.cn/resources/datasets/cox-face-dataset/COX-S2V

and blur. An example of a still ROI belonging to one subject and corresponding video ROIs is shown in Figure 4.

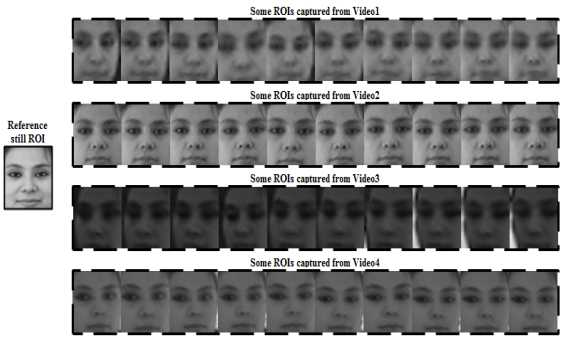


Figure 4: Example of the reference still ROI for enrollment of subject ID #1 and some corresponding ROIs extracted from videos from the 4 OD of the COX-S2V dataset.

In experiments, the high-quality stills for $N_{wl} = 20$ individuals of interest are randomly chosen to populate the watch-list due to assessment of the proposed DS system, as well as, $N_{wl} = 10$ for evaluation of different training schemes. Videos of 100 unlabeled persons from the OD considered as calibration videos are employed during the enrollment for DA. In addition, videos of 100 other unknown people along with videos of the watch-list individuals are merged for testing during the operational phase. Therefore, one target individual at a time and all of the unknown persons within the test videos participate in each testing iteration. In order to achieve statistically significant results, the experiments are replicated 5 times considering different individuals of interest.

The reference still and video ROIs are converted to grayscale and scaled to a common size of 48x48 pixels due to computational efficiency (Huang et al., 2015). Histogram equalization is then utilized to enhance contrast, as well as, to eliminate the effect of illumination changes. Afterwards, each ROI is divided into $N_p = 9$ uniform non-overlapping patch configurations of 16x16 pixels as in (Bashbaghi et al., 2016), (Chen et al., 2015). Libsvm library (Chang and Lin, 2011) is used in order to train e-SVMs differently, where the same regularization parameters $C_1 = 1$ and $C_2 = 0.01$ are considered for all exemplars based on the imbalance ratio (Bashbaghi et al., 2016). Random subspace sampling with replacement is also employed to randomly generate several subspaces $N_{rs} = 20$ from the original feature space. Ensemble of template matchers (TMs) and e-SVMs using multiple face representations (Bashbaghi et al., 2014), (Bashbaghi et al., 2016), ESRC-DA (Nourbakhsh et al., 2016), specialized kNN adapted for video surveillance (VSkNN) (Pagano et al., 2014), and SVDL (Yang et al., 2013) are considered as the baseline and state-of-the-art FR systems to validate the proposed system.

Receiver operating characteristic (ROC) curve is adopted to evaluate performance of the proposed system at transaction-level. Thus, area under ROC curve (AUC) as a global scalar metric of the detection performance is considered, where it may be interpreted as the probability of classification. Another relevant curve that can estimate the system performance under imbalanced data situation is precision-recall (PR), where TPR can be associated as recall and precision (P) is computed as follows: $\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$. System performance are provided using average partial AUC (pAUC) and area under PR (AUPR) along with standard errors. It is worth noting that, the AUPR could be more desirable to represent the global accuracy of the system in skewed imbalanced data conditions.

4.2 Results and Discussion

The average transaction-level performance of different training schemes with considering $N_{wl} = 10$ individuals of interest based on DA are presented in the Table 1 over each video of COX-S2V. Results in Table 1 indicate that the training schemes 1 is greatly outperformed by schemes 2 and 3, where calibration videos from OD are employed for DA to train e-SVMs. However, schemes 2 performs better than the camera-specific training scheme (scheme 3) in terms of both accuracy and computational complexity. In the scheme 2, videos from all of the cameras (global knowledge of the surveillance environment) are employed to generate an e-SVM pool, while 4 camera-specific e-SVM pools are generated for the scheme 3 using videos of each specific camera (partial knowledge of the surveillance environment). For instance, only the classifiers within the pool of camera #1 that are trained using videos captured form camera #1 are employed to classify the probe ROI captured using camera #1 during operations.

Since the capture conditions and camera characteristics are different in COX-S2V dataset, it leads to a significant impact on the system performance. For example, the performance of the proposed system for video3 is lower than other videos. The differences between pAUC(20%) and the corresponding AUPR observed in Table 1 reveal the severely imbalanced operational data, where a large number of e-SVMs can correctly classify the non-target ROIs but some of them can classify the target ROIs correctly. Therefore, the FPR values are very low in all cases and consequently, the pAUC(20%) values obtained from

Table 1: Average pAUC(20%) and AUPR performance of different training schemes and the proposed system with or without DS ($N_{wl} = 10$ for experiments on training schemes and $N_{wl} = 20$ for DS) at transaction-level over COX-S2V videos.

Systems	Video1		Video2		Video3		Video4	
	pAUC(20%)	AUPR	pAUC(20%)	AUPR	pAUC(20%)	AUPR	pAUC(20%)	AUPR
Training Scheme 1	77.62±4.18	57.28±5.08	92.31±1.93	72.90±4.44	69.16±4.32	40.10±5.25	84.63±5.33	58.13±2.89
Training Scheme 2	**100±0.00**	**94.23±0.22**	**99.99±0.00**	**94.06±0.36**	**99.95±0.04**	**94.21±0.33**	**99.99±0.00**	**94.17±0.22**
Training Scheme 3	99.99±0.00	94.13±0.26	99.79±0.13	93.66±0.54	98.27±0.76	89.07±1.98	89.78±0.15	92.68±1.35
Proposed system w.o. DS	100±0.00	94.70±0.19	99.83±0.06	92.49±0.73	95.32±0.87	81.18±1.21	97.04±0.95	84.90±2.01
Proposed system w. DS	**100±0.00**	93.37±0.29	**99.96±0.01**	**92.51±0.41**	**97.68±0.47**	**82.50±1.44**	**98.40±0.44**	**85.23±1.69**

ROC curves are always higher than AUPR values.

Performance of the proposed system either with DS or without DS with $N_{wl} = 20$ are also presented in Table 1 using the second training scheme. As shown in Table 1, applying the proposed DS approach can improve the performance instead of combining all of the classifiers within the pool. It implies that dynamically integrating a subset of competent classifiers leads to a higher level of accuracy over different capture conditions. Since only two distances (distance from the probe to the target still ROI and distance to the closest non-target support vector) are measured in the DS approach, it is efficient and does not significantly increase the computational burden.

The proposed system with DS approach is compared with the state-of-the-art and baseline FR systems in Table 2. It can be seen from Table 2 that the proposed system significantly outperforms ESRC-DA, ensemble of TMs, SVDL, and VSkNN, especially regarding to AUPR values. System using VSkNN and SVDL provide a lower level of performance, mostly because of the considerable differences between the appearance of the target face stills and video faces, as well as, the level of data imbalance of target ROIs versus non-target ROIs observed during operations. It is worth noting that both VSkNN and SVDL are more suitable for close-set FR problems, such as face identification, where each probe face should be assigned to one of the target still in the gallery. However, sparsity concentration index was used as a threshold to reject the probes not appearing in the over-complete dictionaries in SVDL and ESRC-DA. The results observed from Table 2 suggest that the proposed system with DS approach can also achieve a higher or comparable level of performance to (Bashbaghi et al., 2016) with a significant decrease of computational complexity.

Table 2 also presents the complexity in terms of the number of dot products required during operations to process a probe ROI. Computational complexity of the proposed system is mainly affected by the feature extraction, classification, dynamic classifier selection, and fusion for a given probe ROI. In this regard, e-SVM classification is performed with a linear SVM kernel function using a dot product. The complexity to process a probe ROI is $O\left(N_d \cdot N_{sv}\right)$ (Chang and Lin, 2011), where N_d and N_{sv} are the dimensionality of the face descriptors and the number of support vectors, respectively. Thus, the worst case of complexity to process an input ROI can be computed as the product of $N_p \cdot N_{fd} \cdot N_{rs} \cdot N_{sv} \cdot N_d$ according to dot products per e-SVM classifier. For example, the proposed system with DS needs $9 \cdot 2 \cdot 20 \cdot 18 \cdot 71$ dot products for fusion in the worst case, where all of the classifiers are dynamically selected, and $9 \cdot 2 \cdot 20 \cdot 2 \cdot 71$ for performing dynamic selection. Noted that the proposed system in this paper employs two different light-weight face descriptors, whereas ensemble of e-SVMs (Bashbaghi et al., 2016) utilizes four different face descriptors along with applying PCA with the complexity of $O\left(N_d^3\right)$ for feature ranking and selection. Meanwhile, ensemble of TMs and VSkNN employ Euclidean distance with $O\left(N_d^2\right)$ to calculate the similarity among templates. The complexity of ESRC-DA is calculated with $O\left(N_d^2.k\right)$, where k is the number of atoms.

5 CONCLUSION

This paper presents a system specialized for watch-list screening applications that exploits dynamic selection of classifier ensembles trained through multiple face representations and DA. Multiple face representation (different random subspaces, patches, and descriptors) are employed to design the individual-specific ensemble of e-SVMs per target individual, to provide diversity among classifiers, and to overcome the existing nuisance factors in surveillance environments. Unsupervised DA allows to generate diverse pools of e-SVM, where video ROIs of non-target individuals are exploited. Different training schemes were considered using unlabeled non-target video ROIs, and training global e-SVMs on calibration videos from all network cameras performs most efficiently. In addition, a new distance-based criteria of competence is proposed for DS during operations to dynamically select the best subset of classifiers per input probe. Distances of a given probe to the target still and the closest support vector are considered as the competence criteria. Simulation results obtained

Table 2: Average transaction-level performance of the proposed and state-of-the-art FR systems on videos of the COX-S2V.

FR Systems	pAUC(20%)	AUPR	Complexity (number of dot products)
VSkNN (Pagano et al., 2014)	56.80±4.02	26.68±3.58	671,744
SVDL (Yang et al., 2013)	69.93±5.67	44.09±6.29	810,000
Ensemble of TMs (Bashbaghi et al., 2014)	84.00±0.86	73.36±9.82	1,387,200
ESRC-DA (Nourbakhsh et al., 2016)	99.00±0.13	63.21±4.56	432,224,100
Ensemble of e-SVMs (Bashbaghi et al., 2016)	99.02±0.15	88.03±0.85	2,327,552
Proposed system w. DS	**99.02±0.23**	**88.40±0.96**	**504,720**

using videos of the COX-S2V dataset confirm that the proposed system is computationally efficient and outperforms the state-of-the-art systems even when the data is limited and imbalanced.

ACKNOWLEDGMENT

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

REFERENCES

Ahonen, T., Rahtu, E., Ojansivu, V., and Heikkila, J. (2008). Recognition of blurred faces using local phase quantization. In *ICPR*, pages 1–4.

Barr, J. R., Bowyer, K. W., Flynn, P. J., and Biswas, S. (2012). Face recognition from video: A review. *IJPRAI*, 26(05).

Bashbaghi, S., Granger, E., Sabourin, R., and Bilodeau, G.-A. (2014). Watch-list screening using ensembles based on multiple face representations. In *ICPR*, pages 4489–4494.

Bashbaghi, S., Granger, E., Sabourin, R., and Bilodeau, G.-A. (2015). Ensembles of exemplar-svms for video face recognition from a single sample per person. In *AVSS*, pages 1–6.

Bashbaghi, S., Granger, E., Sabourin, R., and Bilodeau, G.-A. (2016). Robust watch-list screening using dynamic ensembles of svms based on multiple face representations. *Machine Vision and Applications*.

Britto, A. S., Sabourin, R., and Oliveira, L. E. (2014). Dynamic selection of classifiers - a comprehensive review. *Pattern Recognition*, 47(11):3665 – 3680.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: A library for support vector machines. *ACM TIST*, 2(3):1–27.

Chen, C., Dantcheva, A., and Ross, A. (2015). An ensemble of patch-based subspaces for makeup-robust face recognition. *Information Fusion*, pages 1–13.

De la Torre Gomerra, M., Granger, E., Radtke, P. V., Sabourin, R., and Gorodnichy, D. O. (2015). Partially-supervised learning from facial trajectories for face recognition in video surveillance. *Information Fusion*, 24:31–53.

De-la Torre Gomerra, M., Granger, E., Sabourin, R., and Gorodnichy, D. O. (2015). Adaptive skew-sensitive ensembles for face recognition in video surveillance. *Pattern Recognition*, 48(11):3385 – 3406.

Deniz, O., Bueno, G., Salido, J., and la Torre, F. D. (2011). Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598 – 1603.

Dewan, M. A. A., Granger, E., Marcialis, G.-L., Sabourin, R., and Roli, F. (2016). Adaptive appearance model tracking for still-to-video face recognition. *Pattern Recognition*, 49:129 – 151.

Huang, Z., Shan, S., Wang, R., Zhang, H., Lao, S., Kuerban, A., and Chen, X. (2015). A benchmark and comparative study of video-based face recognition on cox face database. *IP, IEEE Trans on*, 24(12):5967–5981.

Kamgar-Parsi, B., Lawson, W., and Kamgar-Parsi, B. (2011). Toward development of a face recognition system for watchlist surveillance. *IEEE Trans on PAMI*, 33(10):1925–1937.

Malisiewicz, T., Gupta, A., and Efros, A. (2011). Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, pages 89–96.

Mokhayeri, F., Granger, E., and Bilodeau, G.-A. (2015). Synthetic face generation under various operational conditions in video surveillance. In *ICIP*, pages 4052–4056.

Nourbakhsh, F., Granger, E., and Fumera, G. (2016). An extended sparse classification framework for domain adaptation in video surveillance. In *ACCV, Workshop on Human Identification for Surveillance*.

Pagano, C., Granger, E., Sabourin, R., Marcialis, G., and Roli, F. (2014). Adaptive ensembles for face recognition in changing video surveillance environments. *Information Sciences*, 286:75–101.

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *KDE, IEEE Trans on*, 22(10):1345–1359.

Patel, V., Gopalan, R., Li, R., and Chellappa, R. (2015). Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69.

Qiu, Q., Ni, J., and Chellappa, R. (2014). Dictionary-based domain adaptation for the re-identification of faces. In *Person Re-Identification, Advances in Computer Vision and Pattern Recognition*, pages 269–285.

Shekhar, S., Patel, V., Nguyen, H., and Chellappa, R. (2013). Generalized domain-adaptive dictionaries. In *CVPR*, pages 361–368.

Yang, M., Van Gool, L., and Zhang, L. (2013). Sparse variation dictionary learning for face recognition with a single training sample per person. In *ICCV*, pages 689–696.