Accepted in Computer Vision – ACCV 2016 Workshops, 2017

An Extended Sparse Classification Framework for Domain Adaptation in Video Surveillance

Farshad Nourbakhsh, Eric Granger and Giorgio Fumera

Laboratoire d'imagerie de vision et d'intelligence artificielle
École de technologie supérieure, Université du Québec,
Montréal, Canada
Department of Electrical and Electronic Engineering,
University of Cagliari Piazza d'Armi,
09123 Cagliari, Italy

Abstract. Still-to-video face recognition (FR) systems used in video surveillance applications capture facial trajectories across a network of distributed video cameras and compare them against stored distributed facial models. Currently, the performance of state-of-the-art systems is severely affected by changes in facial appearance caused by variations in, e.g., pose, illumination and scale in different camera viewpoints. Moreover, since an individual is typically enrolled using one or few reference stills captured during enrolment, face models are not robust to intra-class variation. In this paper, the Extended Sparse Representation Classification through Domain Adaptation (ESRC-DA) algorithm is proposed to improve performance of still-to-video FR. The system's facial models are thereby enhanced by integrating variational information from its operational domain. In particular, robustness to intra-class variations is improved by exploiting: (1) an under-sampled dictionary from target reference facial stills captured under controlled conditions; and (2) an auxiliary dictionary from an abundance of unlabelled facial trajectories captured under different conditions, from each camera viewpoint in the surveillance network. Accuracy and efficiency of the proposed technique is compared to state-of-the-art still-to-video FR techniques using videos from the Chokepoint and COX-S2V databases. Results indicate that ESRC-DA with dictionary learning of unlabelled trajectories provides the highest level of accuracy, while maintaining a low complexity.

1 Introduction

With the availability of low-cost video cameras and high capacity memory, technologies for video surveillance (VS) have become more prevalent in recent years. VS networks are increasingly deployed by public security organizations in e.g., airports, train stations and border crossings. Accurate and robust systems are required to recognize individuals and their actions from video feeds.
In VS, decision support systems can rely on facial information (along with other sources, like soft biometrics) to alert an analyst as to the presence of individuals of interest. The ability to automatically recognize faces in videos recorded across

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-54526-4_27

a distributed network of surveillance cameras can greatly enhance security and situational awareness.

Watch-list screening is among the most challenging applications in VS [1, 2]. During enrolment of a target individual, facial regions of interests (ROIs) are isolated from one or few reference still images that were captured under controlled conditions. Then, during operations, each ROI captured in videos are matched against the facial models of each individual enrolled to the system. Robust spatio-temporal FR is typically performed using a person tracker (based on head, face and other information). This allows to accumulate the matching scores for each enrolled individual over a trajectory, e.g., a set of facial ROIs corresponding to a same person tracked in the scene. Thus, a spatio-temporal fusion module compares these accumulated scores with decisions thresholds in order to detect target individuals associated with each trajectory [3].

Currently, the performance of state-of-the-art systems for still-to-video FR is severely affected by variations in, e.g., pose, scale, blur, illumination and camera viewpoint [4]. Systems for FR in VS are typically implemented with individual-specific face detectors (e.g., two-class classification systems)[5]. During enrolment, it is assumed that a detector is designed to encode a facial model, using labelled ROIs extracted from target reference stills versus cohort and other non-target ROIs, all of which are captured under the same controlled conditions (in the enrolment domain, ED). In watchlist screening each individual is typically enrolled using one or few reference stills captured during enrolment, face models are often poor representatives of the faces to be recognized during operations. Moreover, during operations, video ROIs are captured in a camera field of view (FoV) (in an operational domain, OD) under uncontrolled conditions. Capture conditions may vary dynamically within an OD according to environmental conditions and individual behaviours. Therefore, the face model of target individuals are not robust to the intra-class variations of ROIs in an OD, and many yield poor FR performance.

Still-to-video FR can be addressed using techniques proposed in literature for single sample per person (SSPP) problems [6, 7]. In order to improve robustness to intra-class variability, FR techniques specialized for SSPP problems must often rely on adaptation, multiple face representations, synthetic face generation, and enlarged auxiliary reference datasets. However, multiple representation and synthetic generation techniques alone are only effective to the extent where reference target ROIs captured in the environmental domain (ED) are representative of an OD [8].

An important issue in still-to-video FR is that probe ROIs are captured over multiple distributed surveillance cameras, and each one represents a non-stationary OD. Their data distribution differs significantly from ROIs captured with a still camera in the ED [9]. Any distributional change (either domain shift or concept drift) can degrade system performance [10]. Context-aware systems could efficiently adapt to different and changing capture conditions [11]. However, in the most common approach, prior expert knowledge of the expected OD is employed to define typical contexts and to design specialized individual detectors. Then,

a suitable detector is selected dynamically among the pool for a given OD. In practice, however, this approach would only provide coarse adaptation because still-to-video FR systems are deployed in diverse and unknown capture condition. It is difficult to predict, and to collect adequate labelled data for each context. Given an adequate number of labelled reference samples, numerous adaptive classifiers [12] and ensemble methods could also be employed for adaptation in non-stationary environments, some of which are specialized for video-to video FR.

Several transfer learning methods have recently been proposed to design accurate recognition systems that will perform well on OD data given knowledge from the ED. Since the learning tasks and feature spaces between ED and OD are the same, but their data probability distributions are different, our transfer learning scenario is related to domain adaptation (DA). According to the information transferred between an ED and an OD, two unsupervised DA approaches from literature are relevant for still-to-video FR [9], [13], [14]. Instance transfer methods attempt to exploit parts of the OD data for learning in the ED. In contrast, feature representation transfer methods exploit OD data to find a good common feature representation space that reduces the difference between ED and OD spaces and the classification error. Note that most methods in literature initially require an adequate number of labelled reference samples from the ED.

In this paper, we focus on sparse modelling techniques that are suitable for SSPP problems, and allow for instance-based DA. In particular, a framework based on the Extended Sparse Representation Classification (ESRC)[15] algorithm is proposed for the design of still-to-video FR systems, as needed in watchlist screening applications that is called ESRC-DA. Assume that each individual of interest is enrolled to the system using one reference still image capture under controlled condition. Facial models based on one reference still conditions and that facial diverge from the faces captured with video surveillance cameras. Therefore with the ESRC-DA algorithm, unsupervised DA is exploited for accurate recognition of individuals. A large auxiliary dictionary of unlabelled ROIs is extracted from an abundance of facial trajectories captured under different OD conditions, from each camera FoV in the surveillance network. This is combined with an under-sampled dictionary extracted from few reference labelled facial ROIs captured from reference target stills under controlled ED conditions. Apart from improving robustness to intra-class variations in ODs, an advantage of the new ESRC-DA algorithms is the ease for managing information from multi-source environments by organizing and combining different domains. Indeed, a learned dictionary that combines ROIs from ED and OD data for SRC may not be optimal if they have different data distributions. Finally, dictionary learning methods are recommended to compactly represent the variational information from potentially large sets of unlabelled video ROIs.

The rest of the manuscript is organized as follows. Section 2 introduces some notation of sparse modelling in still-to-video FR. Section 3 introduces the proposed method based on ESRC with domain adaptation. Section 4 is devoted to showing the effectiveness of ESRC-DA on two public video-surveillance datasets

(Chokepoint and COX-S2V), and to its comparison with state of the art reference methods. Finally, some concluding discussion are provided.

2 Sparse Modelling in Still-to-Video FR

Watch-list screening is among the most challenging applications in of video surveillance. In VS, FR attempts systems to detect the presence of individuals of interest enrolled to the system. During enrolment some target individual, facial regions of interests (ROIs) are isolated in one or few reference still images that were captured under controlled conditions in the ED. It is assumed that discriminant features are extracted into ROI patterns using a state-of-the-art face descriptor to design a facial model. During operations, ROIs isolated from videos are matched against the facial models where these ROIs are captured under uncontrolled conditions in the OD. Therefore, the performance of state-of-the-art FR systems is severely affected by occlusion, variations in capture conditions (pose, scale, expression, illumination, blur, etc.) [8]. In addition, watchlist screening systems must employ a limited reference samples in order to design the facial models. To overcome the aforementioned challenges of SSPP problems and enhance robustness to intra-class variability, FR systems may exploit auxiliary data captured from unknown individuals or actors during some calibration process.
SRC recently has received much attention in FR literature due to its potential to handle intra class variability. The main assumption in sparse modelling is that as any test sample (ROI pattern) from a class (individual) enrolled with a sufficient number of training samples will be approximately represented in the linear span of training set of corresponding class. The sparse representation of a sample can be formulated as

$$\mathbf{y} = \mathbf{D}\mathbf{x_0}, \tag{1}$$

where matrix $\mathbf{D}$ is defined as over-complete dictionary with N distinct classes that has enough samples for each class. The column of matrix $\mathbf{D}$ contains training samples, and $\mathbf{x_0}$ is a sparse coefficient vector. Ideally, the entries of $\mathbf{x_0}$ are always zero, except for the same class as $\mathbf{y}$. By increasing the number of classes that increases the sparsity of coefficient vector, Eq (1) is able to handle some degree of noise, and provides a unique solution. The Eq (1) can be solved with a l_1 minimization as follows:

$$\mathbf{x_0} = \min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{2}$$

where $\lambda > 0$, is the scalar regularization parameter that controls reconstruction error and sparsity. According to the Eq (2) that is the main part of SRC algorithm, the test sample $\mathbf{y}$ can be reconstructed linearly with a relevant basis of dictionary $\mathbf{D}$ respect into the sparse coefficient with non zero value for the corresponding class and almost zero for non classes. [16] has proposed a sparsity concentration index (SCI) that is defined as below

$$SCI(\mathbf{x}) = \frac{N.\max_i |\delta_i(\mathbf{x})|_1/|\mathbf{x}|_1 - 1}{N-1} \tag{3}$$

For the solution found by Eq (2), SCI is 1 when the test sample belongs to the one of the N classes and it is zero when the sparse coefficient are spread over all classes. SRC method has been applied successfully on FR application [16] with a large amount of training data that has a direct effect on classification result. It is clear that this technique is not fit with an application that provides few number of training samples that is called under-sampled dictionary, like still-to-video video surveillance.

A few FR systems based on the SRC algorithm have been proposed in the literature for VS applications. Naseem et. al. [17] extended the SRC for video-based FR framework and compared the performance of proposed systems with state of art SIFT methods. They have improved the video-to-video FR system by fusing the scores obtained with SRC and SIFT. Nagendra et al. [18] applied their method on video-based FR to identify a video trajectory of a person while rejecting unknown individuals. Optimized combination of Gabor and HOG features are obtained to produce a robust descriptor for video-to-video FR system that is followed by regularized SRC algorithm. Cui et al. [19] proposed a video-to-video FR system that measured the similarity between two image sets that uses joint sparse representation.

There are very few systems specialized for still-to-video FR. Jianquan et al. [20] have extended the problem of FR with SSPP to patch based dictionary learning to improve robustness to variations. The literature mainly focuses on closed-set classification in video-to-video FR or still-to-still FR for a fixed number of individuals. Moreover, the problem of still-to-video FR surveillance become a very challenging when ROIs captured in the operational domain video are different from ROIs captured during enrolment. Transfer learning methods provide techniques to cope with the domain difference between ED and OD.

Transductive transfer learning requires the source and target task be the same with different domain. DA is defined as a relaxed notion of transductive transfer learning by availability of part of unlabelled data in target domains that divide to supervised and unsupervised categories and sparse coding and dictionary based methods have been proposed for DA. Fore example, [21] modelled dictionaries across different domains with a parametric mapping function. Ni et al. [22, 23] proposed an unsupervised DA dictionary learning framework by generating a set of intermediate dictionaries, which smoothly connect the ED and OD. It allows the synthesis of data associated with the intermediate domains that can then be used to build classifiers for domain shifts. A semi-supervised DA dictionary learning framework was proposed for learning a dictionary to optimally represent both ED and OD data [24]. Both the domains are projected into a common low-dimensional space, allowing disregarding irrelevant information. Finally, a view independent representation is achieved by storing all the intermediate dictionaries. Another technique is introduced by [25, 26] with learning a target classifier from classifiers trained on the source domain(s). Zheng et al. [27] propose a dictionary learning approach based on DA between videos from two

domains by forcing the two videos from a same frame in different domain have the same sparse representation to learn two separate dictionaries. In this way, the source view video can be directly applied on the target view video. Shekhar et al. [28] proposed a DA method by mapping source and target domain to a low dimensional subspace and learn a common dictionary.
ESRC [15] is sparse representation modelling method that is adapted into SSPP where external data can be used to learn intra-class variabilities. Given a limited set of labelled reference data for target individuals during enrolment, and an abundance of unlabelled non target videos that can captured from different camera domains, ESRC can perform DA by representing the variations of faces appearing during operations.
The issue of undersampled data is addressed in [15] with ESRC that extends Eq (2) to the following

$$\mathbf{x_0} = \min_{\mathbf{x}} \|\mathbf{y} - [\mathbf{D}, \mathbf{E}] \begin{bmatrix} \mathbf{x_d} \\ \mathbf{x_e} \end{bmatrix} \|_2^2 + \lambda \|\mathbf{x}\|_1 \tag{4}$$

Where $\mathbf{D}$ is an under-sampled dictionary populated with target data, $\mathbf{E}$ is external dictionary with non target data and $\mathbf{x_e}$ is corresponds to the variant bases. After obtaining sparse coefficient $\mathbf{x}$, a test image $\mathbf{y}$ will be assigned to the class with the minimum class-wise reconstruction error $r(\mathbf{y})$, defined as:

$$r(\mathbf{y}) = \|\mathbf{y} - [\mathbf{D}, \mathbf{E}] \begin{bmatrix} \delta(\mathbf{x_d}) \\ \mathbf{x_e} \end{bmatrix} \|_2^2 \tag{5}$$

where $r_k(\mathbf{y})$ and $\delta_{\mathbf{k}}(\mathbf{x_d})$ refer to the k-th class, $k = 1, ..., N$ that is a vector whose only nonzero entries are the entries in $\mathbf{x_0}$ that are associated with class k and the test image $\mathbf{y}$ will be assigned to the class with the minimum class-wise reconstruction error that is $r_k(\mathbf{y})$.
One of the solutions to handle large amount of external data with different domains is to select data randomly from external data to reduce the time complexity. The drawback of this method is that the performance of ESRC affected by amount of data exploited to cover all the common variations in all domains. The second option is to apply DL methods allow for a compressed representation of all domains from external data. For instance, Shafiee et al. [29] have investigated the impact on performance of three different DL methods for SRC. They used Metaface DL, Fisher Discriminative DL (FDDL), Sparse Modelling Representative Selection (SMRS) to obtain compact representation of training data. They showed that the FDDL method provides a high recognition accuracy compare to other methods. K-Means Singular Value Decomposition (KSVD) [30] and Method of Optimal Directions (MOD) [31] are two popular unsupervised DL techniques which have been used in the literature. These EM style methods alternate between dictionary and sparse coding. The difference between these two methods are in dictionary updating – KSVD updates atom by atom, and MOD updates all atoms simultaneously. Graph compression is another option to learn a compact representation of different domains. Depending on the type of encoding, these methods produce a lossy or lossless compression. Data can be presented as a collection of feature vectors or representation of the similarity/dissimilarity relations among data samples. Nourbakhsh et al. [32] have

proposed a graph compression method based on matrix factorization that focuses on structural information of the input data and the extension of the proposed method on DL is presented on [33]. Finally, several methods that combine DL and classification have recently been proposed like SVDL [34] and [35] to produce a compact auxiliary dictionary and classification in the same time.

3 ESRC with Domain Adaptation

In this paper, a still-to-video FR system for watchlist screening is proposed. It can efficiently adapt to multiple non-stationary sources of ROIs based on contextual information from the specific operational VS environment. Despite the very limited number of labelled still ROIs from enrolment (ED), the framework relies on a bank of unlabelled non-target video ROIs captured (during a prior calibration phase) from each different camera FoVs (OD) in a distributed surveillance network. The proposed system is robust to variations of an individual's facial appearance caused by changes in, e.g., pose, illumination, scale and camera viewpoint in the VS environment by extracting a compact auxiliary dictionary from different ODs. ESRC-DA benefits from the abundance of external facial data that is readily available in VS applications, followed by DL to compress the external (variational) dictionary and improve accuracy. Figure 1 illustrates the processing pipeline of the proposed ESRC-DA method for still-to-video FR. It is divided into design and operational phases.

3.1 Design Phase:

The system designed consist in learning external dictionary $\mathbf{E}$ from an abundance of facial trajectories captured under difference conditions and camera FoVs during the calibration process and and under-sampled dictionary $\mathbf{D}$ of labelled reference ROIs captured with a high quality still camera in the **ED** (see Algo 1). Video trajectories are gathered from several cameras with different ODs. The Faces of unknown people or actors are captured during data collection therefore data are not labelled. For each frame, (**line 1**) face detection and preprocessing module performs a segmentation on each frame I_i^v to isolate a region of interest (ROI) to each different face (**line 2**). Then features are extracted from each normalized ROI_r and stored as a pattern $\mathbf{a}_{r,i}^v$ (**line 4**) to produce an overcomple dictionary $\mathbf{M} = [\mathbf{m}_1, ..., \mathbf{m}_r]$ that passes to dictionary learning module which finds a compact representation of ROI patterns from video calibration (**line 5**). DL reduces the size of overcomplete dictionary $\mathbf{M}$ to a compressed auxiliary dictionary $\mathbf{E} = [\mathbf{e}_1, ..., \mathbf{e}_i]$ (**line 7**). The still ROIs captured during the enrolment process are stored in dictionary $\mathbf{D}$ as a gallery faces model by extracting faces from each ROI of still image $\mathbf{I}_j^s$ and preprocessing each ROI(**line 9**). In the next step, feature extraction is used on each ROI_k to produce a pattern $\mathbf{a}_{k,j}^s$ (**line 11**) stored in the undersampled dictionary $\mathbf{D}$,(**line 12**).

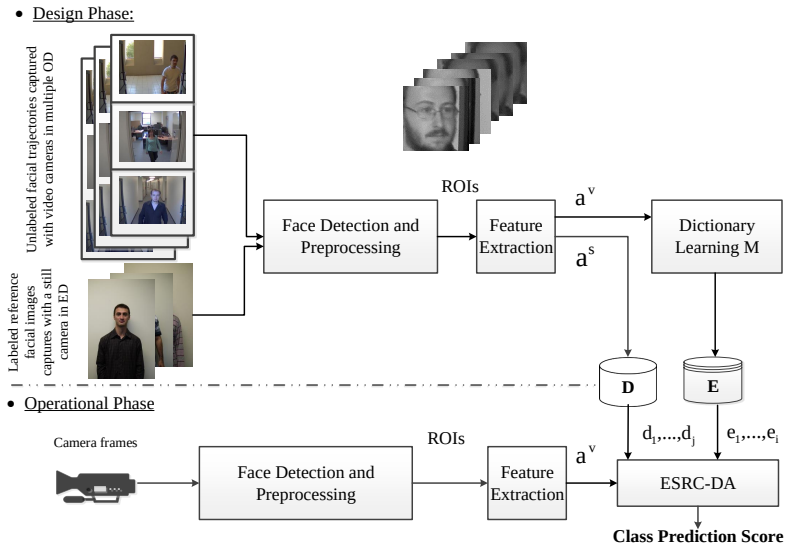


Fig. 1. Block diagram of ESRC-DA method that generates matching scores for an input ROI captured with one surveillance camera.

DL is presented as a method to reduce the size of overcomplete dictionary to produce a compressed dictionary that obtain almost the same performance as uncompressed data. The large amount of external data makes inefficient to process it as a dictionary that covers intra class variability. DL provides a compact representation of data by partitioning the input data into coherent cluster. A cluster can be seen as a set of elements being mutually similar and dissimilar to elements belonging to other cluster that the compressed graph is equal to identity matrix in those cases. All gathered data from different domains have similarity to other elements so they can be grouped as cluster to reduce the size of original data. Therefore DL, clusters data from different ODs to produce a compact dictionary that is representative of domain variability.

3.2 Operational Phase:

During operation, the proposed system is designed to obtain a score that depicts the similarity of a given input probe ROI to each to a watchlist individual enrolled to the system (see Algo 2). Faces are captured using a face detector that extracts ROIs in each frame I_i^v (**line 1** & **2**). Pattern is $\mathbf{a}_{n,i}^v$ calculated by applying feature extraction module (**line 4**). For ESRC, the main assumption is intraclass variation of any watchlist face can be approximated by a sparse linear combination of the intraclass differences from sufficient number of generic faces. ESRC assign $\mathbf{a}_{n,i}^v$ to the corresponding gallery watchlist dictionary $\mathbf{D}$ with respect to the external dictionary $\mathbf{E}$ so column of matrix $\mathbf{D}$ and $\mathbf{E}$ are normalized

to have unit l_2-norm (**line 5**) and the coefficients $\mathbf{x}_0$ is obtained by minimizing the following equation $\min_x \|\mathbf{a}^{\mathbf{v}}_{\mathbf{n,i}} - [\mathbf{D}, \mathbf{E}] \left[\begin{smallmatrix}\mathbf{x_d}\\ \mathbf{x_e}\end{smallmatrix}\right] \|_2^2 + \lambda\|\mathbf{x}\|_1$ (**line 6**). Finally, $r_{k,i,n}(\mathbf{y}) = \|\mathbf{a}^{\mathbf{v}}_{\mathbf{n,i}} - [\mathbf{d_k}, \mathbf{E}] \left[\begin{smallmatrix}\delta(\mathbf{x_d})\\ \mathbf{x_e}\end{smallmatrix}\right] \|_2^2$ computes the similarity of the probe input ROI to each facial model (**line 8**).

Algorithm 1 Design Phase for all Cameras FoVs

*** Input:** Unlabeled facial trajectories captured with video cameras in multiple OD and labelled reference facial images captures with a still camera in ED

*** Output:** Undersampled dictionary $\mathbf{D} = [\mathbf{d_1}, ..., \mathbf{d_K}]$ of still images, $\mathbf{E} = [\mathbf{e_1}, ..., \mathbf{e_i}]$ as an external dictionary

// Unlabelled facial trajectories captured with video cameras in multiple OD

1: For each frame I_i in the video sequence, $i = 1, ..., \infty$
2: ROI $\leftarrow$ face detector and preprocessing on frame I_i^v
3: **for** r=1 to number of ROIs in frames **do**
4: $\quad \mathbf{a}^{\mathbf{v}}_{\mathbf{r,i}} \leftarrow$ feature extraction on ROI$_r$
5: $\quad \mathbf{M} \leftarrow [\mathbf{M}; \mathbf{a}^{\mathbf{v}}_{\mathbf{r,i}}]$
6: **end for**
7: $\mathbf{E} = [\mathbf{e_1}, ..., \mathbf{e_i}] \leftarrow$ dictionary learning (See section 3 on DL)

// Labelled reference facial images captures with a still camera in Ed

8: For each still images I_j in the sequence, $j = 1, ..., \infty$
9: ROI $\leftarrow$ face detector and preprocessing on image I_j^s
10: **for** k=1 to number of ROIs in image **do**
11: $\quad \mathbf{a}^{\mathbf{s}}_{\mathbf{k,j}} \leftarrow$ feature extraction on ROI$_j$
12: $\quad \mathbf{D} = [\mathbf{d_1}, ..., \mathbf{d_{k,j}}] \leftarrow [\mathbf{D}; \mathbf{a}^{\mathbf{s}}_{\mathbf{k,j}}]$
13: **end for**

4 Experimental Results

4.1 Methodology for Validation

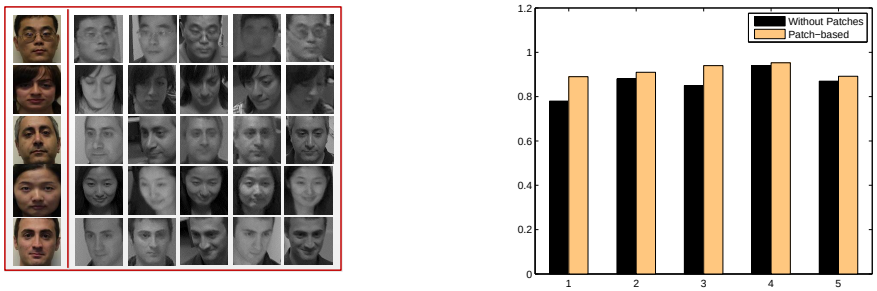
The proposed system is validated and compared experimentally using videos from two challenging real-world data sets: Chokepoint and COX-S2V. They are video surveillance datasets that emulate watchlist screening applications. The main characteristics of these two datasets with respect to others is that they contain a high-quality still face image and surveillance videos for each subject. Videos are captured over a distributed networks of cameras that covers a range of variations changes in, e.g., pose, illumination, blur, scale. These are presently among the most representative public data sets for watchlist screening applications.

Algorithm 2 Operational Phase for Frames Captured with one Camera

*** Input:** Undersampled dictionary $\mathbf{D} = [\mathbf{d_1}, ..., \mathbf{d_K}]$ of still images, external dictionary $\mathbf{E} = [\mathbf{e_1}, ..., \mathbf{e_i}]$ of videos from different domains,
*** Output:** Matching score for a probe transaction (ROI) for each facial model in dictionary $\mathbf{D}$, $r_{[k=1,...,K]}(y)$,

// Transaction Level Processing
for each frame I_i in the video sequence, $i = 1, ..., \infty$ **do**
  ROI $\leftarrow$ Face Detector and Preprocessing on frame I_i^v
  for n=1 to number of ROIs in frame **do**
    $\mathbf{a^v_{n,i}} \leftarrow$ Feature Extraction on ROI$_n$
    Normalize the columns of $\mathbf{D}$ and $\mathbf{E}$ to have unit l_2-norm
    Solve the l_1-minimization problem
      $\min_x \|\mathbf{a^v_{n,i}} - [\mathbf{D}, \mathbf{E}] \begin{bmatrix} \mathbf{x_d} \\ \mathbf{x_e} \end{bmatrix} \|_2^2 + \lambda \|\mathbf{x}\|_1$
    for k=1 to number of facial model in Dictionary $\mathbf{D}$ **do**
      Compute the residuals
      $S = 1 - r_{k,i,n}(\mathbf{y}) = \|\mathbf{a^v_{n,i}} - [\mathbf{d_k}, \mathbf{E}] \begin{bmatrix} \delta(\mathbf{x_d}) \\ \mathbf{x_e} \end{bmatrix} \|_2^2$ decision
    end for
  end for
end for

Chokepoint dataset [36] contains high quality frontal still image of each individuals and videos captured from those same people passing through different portal with 3 surveillance cameras. It is composed of videos of 25 individual with 19 subjects are male and 6 are female. The 48 out of 54 video sequences are one person at a time and the rest remainder contain a mixture of people. Videos are captured from 3 cameras positioned over two portals during 4 sessions (1 month intervals), at 30 fps, and with an image resolution is 800×600 pixels. Each ROI or face images are scaled to a common size of 96×96 pixels. This dataset features variations in pose illumination, lighting, scale, and blur that makes it challenging for still-to-video FR. Fig 2(a) shows ROIs of 5 selected target individuals and their test video correspondence is recorded with 3 cameras above different portals.
COX-S2V [37] consists of 1000 subjects, where each subject has a high quality still images under controlled conditions, and four lower-quality facial trajectories captured under uncontrolled conditions. Each trajectory has 25 faces, where ROIs taken from these videos encounter changes in illumination, expression, scale, viewpoint, and blur.
In all experiments with Chokepoint dataset, 5 target individuals among 25 are selected randomly to design a watch-list that includes a high quality frontal captured images of selected people. The background model for the operational phase is designed based on selecting 10 unknown individual video sequences along 5 video sequences of already selected people in the watch-list. Moreover, the remaining 10 individual's video sequences are used as external data to build auxiliary dictionary. For the experiment with COX-S2V, three times randomly 20 individuals among 1000 subjects are selected to build a watch-list from high



(a) (b)

Fig. 2. (a)An example of ROIs extracted from mugshots of 5 redundancy selected target individuals of interest and some of the ROIs from corresponding operational video; (b) Comparison of ESRC-DA w/ DL MFA with and without patch-based extraction from ROIs.

quality data and their corresponding low quality video trajectories are applied for testing. Background model is made of 100 video sequence of individuals to be applied on operational phase.

As described in Section 2, there are many methods in literature for DL like KSVD, MOD and SMRS. Nourbakhsh et. al [32] proposed a method to compress a graph based on matrix factorization approach (MFA) that needs an efficient time to construct the reduced graph and it is parameter free. Let M is a $N \times N$ similarity matrix and $K \leq N$ a constant and the compression rate is K/N. The goal of DL is to produce a reduced matrix R with order $K \times K$ and a many to one mapping function $\psi : [n] \rightarrow [k]$ between vertices of the original graph and reduced graph. The mapping function is expressed in terms of a left stochastic matrix. A least squares approximation is applied on following minimizer by dropping a left-stochastic constrain to a real matrix X to calculate the reduced matrix and mapping function.

The optimization can be addressed as a EM method which alternates updates of the variable R and updates of the variable X. The minimization approach converges to a stationary point by updating a decrease of the objective function in every iteration.

$$min \quad f(X,R) = ||M - X^T R X||_2^2 \tag{6}$$

where $s.t. \quad X \in S, \quad R \in R^{k \times k}$

and

$$f(X,R) = \sum_{(i,j)\in\{1,2,\ldots,N\}} \sum_{(k,h)\in\{1,2,3,\ldots K\}} \delta_{(k,i)\neq(h,j)} X_{ki} X_{hj}$$

$$\times (M_{ij} - R_{kh})^2 + \sum_{i\in\{1,2,\ldots,N\}} \sum_{k\in\{1,2,3,\ldots K\}} X_{ki}(M_{ii} - R_{kk})^2$$

It has been shown that DL algorithm reduces the complexity of many algorithms from n^2 to $(k^2 + n)$ by replacing the original data with its corresponding factorization.
Receiver Operating Characteristic (ROC) is applied to evaluate the performance of the proposed system that is defined as true positive rate (TPR) versus false positive rate (FPR). TPR is the proportion of correctly detected as individual of interest over the total number of target ROIs and FPR is the proportion of non-target detected as individual of interest over the total number of non-target ROIs. Area Under the ROC (AUC) is a global measure of performance which is defined as the probability of classification over the range of TPR and FPR. Finally Precision-Recall Operating Characteristic (PROC)curve constitutes a graphical representation of detector performance where the impact of data imbalance is considered. The precision between positive predictions (precision $PR = \frac{TP}{TP+FP}$) is combined with the TPR (or recall) to draw a PROC curve. To measure the efficiency of different algorithms, this paper also shows the time complexity. It is estimated as the total number of dot products (DPs) needed by an algorithms to produce a matching score in response to one probe ROI captured in operational videos.

4.2 Results and Discussion

Several experiments were conducted on the Chokepoint dataset to characterize the effectiveness of the proposed method. AUROC and AUPR curves were produced using ESRC-DA *w/* DL MFA on the test videos of five randomly selected watchlist individuals. These curves show the impact of exploiting DA. ESRC-DA can achieve a higher level of performance, even when the proportion of target to non-target is imbalanced (as seen in PR curves). Figure 3 shows the ROC and P-R curves obtained with the proposed system (ESRC-DA w/ DL MFA) on Chokepoint data when each ROI is represented with HOG features. The faces captured in video and stills were scaled and normalized to 48×48 pixels ROIs and represented using HOG descriptors. In the next experiment, these ROIs were divided to 9 fixed size patches (forming 9 different external dictionaries), where the scores computed from local matching were combined using the average score-level fusion rule.
Figure 2 (b) shows the AUC performance with and without patch and patch-based extraction using ESR-DA *w/* DL MFA. It measure the effectiveness of local patch-based matching to handle variations in pixel features. Results show that that patch based method improve the performance processing ROIs distorted by, e.g., pose variations but are computationally slower.
Table 1 shows the average AUC and P-R accuracy, and time complexity of ESRC-DA w/ DL versus state of the art methods which are NN, SRC, SRC with SCI, RSC, ESRC w/o DL, RADL w/o DL, RADL w/ DL and SVDL w/ DL for still-to-video FR. The time complexity $O(N_d \times N_v)$ is calculated in terms of the overall number of dot products (DPs) to process a ROI in operational phase, where N_d is dimensionality and N_v is number of vector. The worse case scenario is considered for each method. Fore example the time complexity of SRC

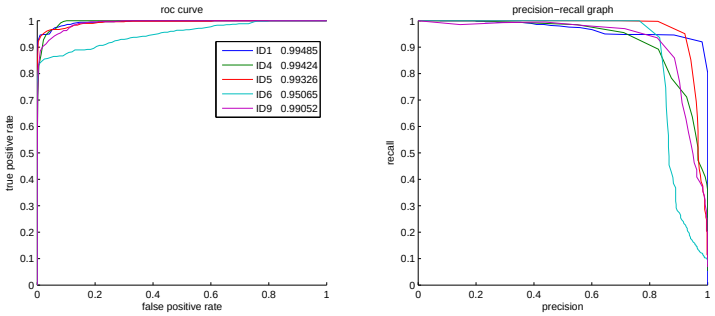


Fig. 3. ROC and P-R curves for the ESRC-DA methods on Chokepoint videos. Area under the curve of ROC is reported for five randomly selected individuals in the legend.

Table 1. Average performance of ESRCA-DA and references methods on Chokepoint dataset.

Classifiers	**AUC**	**AUPR**	**Time Comp (DPs)**
NN (Template Matching)	0.6991±0.0411	0.2635±0.0663	1,890
SRC [17]	0.8866±0.0523	0.5447±0.1427	3,572,100
SRC with Reject Criteria (SCI)	0.9041±0.0509	0.5576±0.1539	3,572,100
RSC [38]	0.8007±0.0600	0.2478±0.0759	3,572,100
ESRC w/o DL (Random) [15]	0.9360±0.0311	0.6317±0.1418	228,614,400
RADL w/o DL (Random) [35]	0.8313±0.0701	0.3000±0.1164	228,614,400
RADL $w/$ DL [35]	0.8211±0.0601	0.2106±0.0762	230,748,921
SVDL $w/$ DL [34]	0.7123±0.0141	0.3112±0.0543	230,748,921
ESRC-DA $w/$ DL KSVD [30]	0.9450±0.0455	0.6432±0.1783	228,614,400
ESRC-DA $w/$ DL MOD [31]	0.9408±0.0404	0.6553±0.1655	228,614,400
ESRC-DA $w/$ DL SMRS [39]	0.9342±0.0399	0.6316±0.1719	228,614,400
ESRC-DA $w/$ DL MFA [32]	**0.9716±0.0128**	**0.7697±0.0673**	228,614,400

and ESRC is considered quadratic to the number of training samples [40]. The table shows the average performance over 5 replications with random selection of individuals. The reference methods can be divided to three main categories: methods without DL that are NN, SRC, RCS, ESRC w/o DL and RADL w/o DL where SRC and NN are used as a base line methods, algorithms with different DL and classification that are RADL $w/$ DL and SVDL. The last category is DA methods that are ESRC followed with KSVD, MOD, SMRS and MFA as DL. The ESR-DA $w/$ DL MFA shows a significantly higher level of AUC and AUPR accuracy, with lower time complexity to reconstruct DL $O(k^2 + n)$ compare to state of the art methods and it is parameter free. The proposed system (ESRC-DA $w/$ DL) is also compared using COX-S2V data with the state-of the art methods in Table 2.

The proposed system (ESRC-DA $w/$ DL) is also compared using Chokepoint dataset with still-to-video state-of the art methods in Table 3. The result shows

Table 2. Average performance of ESRCA-DA and references methods on COXS2V dataset.

Classifiers	**AUC**	**AUPR**	**Time Comp (DPs)**
SVDL *w/* DL [34]	0.6993±0.5671	0.4409±0.6291	432,364,725
ESRC-DA *w/* DL KSVD [30]	0.9729±0.0313	0.5280±0.2813	432,224,100
ESRC-DA *w/* DL MOD [31]	0.9781±0.0312	0.5612±0.2213	432,224,100
ESRC-DA *w/* DL SMRS [39]	0.9743±0.0405	0.5712±0.1581	432,224,100
ESRC-DA *w/* DL MFA [32]	0.9900 ±0.0113	0.6321±0.0456	432,224,100

that ESRC-DA */w* DL MFA provide a high level of performance with a lower time complexity. For the time complexity, the codes are implemented in MATLAB, using a 3.40 GHz and 8 GB RAM computer.

Table 3. Average performance of ESRCA-DA and references methods on Chokepoint.

Classifiers	**AUC**	**AUPR**	**Time Complexity**
AAMT-FR [3]	0.6490±0.070	0.793±0.03	0.217±0.06 Sec
Ensemble of e-SVMs (1 block) [8]	0.9228±0.54	0.909±0.284	0.435±0.07 Sec
ESRC-DA *w/* DL MFA [32]	0.9643 ±0.051	0.7151±0.022	0.119±0.04 Sec

5 Conclusion

In this paper, an Extended SRC framework for still-to-video FR is proposed to accurately recognize individuals of interest across a distributed network of video surveillance cameras. To overcome the limitations of labelled reference still ROIs that are captured during enrolment for face modelling, this algorithm exploits are abundance of external video ROIs that can typically be captured with different cameras in the operational domain. The Extended SRC through Domain Adaptation (ESRC-DA) algorithm enhanced the robustness of facial models by integrating information from its operational an under-sampled dictionary from target reference stills with an auxiliary dictionary learned using facial trajectories captured under different operational capture conditions. The proposed matrix factorization approach is well adapted for this context, and it is shown that ESRC-DA *w/* DL MFA outperforms state-of-the-art methods. Results were obtained on two real-world video surveillance data sets (Chokepoint and COX-S2V) and using various face representations and matching schemes. Exploiting unlabelled facial ROIs captured in videos of the operational environment allow for the ESRC-DA algorithm to achieve a higher level of FR accuracy than state-of-the-art systems. Furthermore, it represents a cost-effective solution for still-to-video FR, as required in several VS applications.

References

1. Chen, S., Mau, S., Harandi, M.T., Sanderson, C., Bigdeli, A., Lovell, B.C.: Face recognition from still images to video sequences: A local-feature-based framework. EURASIP J. Image and Video Processing (2011)
2. Kamgar-Parsi, B., Lawson, W., Kamgar-Parsi, B.: Toward development of a face recognition system for watchlist surveillance. IEEE Trans. PAMI (2011) 1925–1937
3. Dewan, M.A.A., Granger, E., Marcialis, G.L., Sabourin, R., Roli, F.: Adaptive appearance model tracking for still-to-video face recognition. Pattern Recognition **49** (2016) 129–151
4. Pato, J., Millett, L.: Biometric recognition: Challenges and opportunities. Whither Biometrics Committee, National Research Council of the NSA (2010)
5. Pagano, C.C., Granger, E., Sabourin, R., Gorodnichy, D.O.: Detector ensembles for face recognition in video surveillance. In: IJCNN,. (2012) 1–8
6. Tan, X., Chen, S., Zhou, Z.H., Zhang, F.: Face recognition from a single image per person: A survey. Pattern Rec. (2006) 1725–1745
7. Kan, M., Shan, S., Su, Y., Xu, D., Chen, X.: Adaptive discriminant learning for face recognition. Pattern Recognition **46** (2013)
8. Bashbaghi, S., Granger, E., Sabourin, R., Bilodeau, G.: Ensembles of exemplar-svms for video face recognition from a single sample per person. In: IEEE Int. Conf. on Advanced Video and Signal Based Surveillance,. (2015) 1–6
9. Patel, V.M., Gopalan, R., Li, R., Chellappa, R.: Visual domain adaptation: A survey of recent advances. IEEE Signal Process. Mag. **32** (2015) 53–69
10. Minku, L.L., White, A.P., Yao, X.: The impact of diversity on online ensemble learning in the presence of concept drift. IEEE Trans. on Knowledge and Data Engineering (2010)
11. Snidaro, L., Garca, J., Llinas, J.: Context-based information fusion: A survey and discussion. Information Fusion **25** (2015)
12. Huang, Z., Wang, R., Shan, S., Li, X., Chen, X.: Log-euclidean metric learning on symmetric positive definite manifold with application to image set classification. In: ICML. JMLR Workshop and Conf. Proc. (2015)
13. Margolis, A.: Automatic annotation of spoken language using out-of-domain resources and domain adaptation. IEEE Trans. on Knowledge and Data Eng. (2011)
14. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. on Knowl. and Data Eng. (2010)
15. Deng, W., Hu, J., Guo, J.: Extended src: Undersampled face recognition via intraclass variant dictionary. IEEE Trans. on PAMI (2012)
16. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. IEEE Trans. Pattern Anal. Mach. Intell. (2009)
17. Naseem, I., Togneri, R., Bennamoun, M.: Sparse representation for video-based face recognition. In: Third Int. Conf. Advances in Biometrics. (2009)
18. Nagendra, S., Baskaran, R., Abirami, S.: Video-based face recognition and face-tracking using sparse representation based categorization. Procedia Computer Science (2015) Int. Conf. on Data Mining and Warehousing.
19. Cui, Z., Chang, H., Shan, S., Ma, B., Chen, X.: Joint sparse representation for video-based face recognition. Neurocomputing (2014)
20. Gu, J., Liu, L., Hu, H.: Patch-based Sparse Dictionary Representation for Face Recognition with Single Sample per Person. In: Biometric Recognition: Chinese Conf.,. (2015)

21. Qiu, Q., Patel, V.M., Turaga, P.K., Chellappa, R.: Domain adaptive dictionary learning. In: European Conf. on Computer Vision. (2012)
22. Ni, J., Qiu, Q., Chellappa, R.: Subspace interpolation via dictionary learning for unsupervised domain adaptation. In: IEEE Conf. on Computer Vision and Pattern Recognition,. (2013)
23. Qiu, Q., Ni, J., Chellappa, R.: Dictionary-based domain adaptation methods for the re-identification of faces. In: Person Re-Identification. (2014)
24. Shekhar, S., Patel, V.M., Nguyen, H.V., Chellappa, R.: Generalized domain-adaptive dictionaries. In: IEEE Conf. on Computer Vision and Pattern Recognition,. (2013)
25. Duan, L., Tsang, I.W., Xu, D., Chua, T.: Domain adaptation from multiple sources via auxiliary classifiers. In: Proc. of Int. Conf. on Machine Learning. (2009)
26. Duan, L., Tsang, I.W., Xu, D.: Domain transfer multiple kernel learning. IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012)
27. Guo, H., Jiang, Z., Davis, L.S.: Discriminative dictionary learning with pairwise constraints. In: Asian Conf. on Computer Vision Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers,. (2012)
28. Shekhar, S., Patel, V.M., Nguyen, H.V., Chellappa, R.: Generalized domain-adaptive dictionaries. In: IEEE Conf. on Computer Vision and Pattern Recognition,. (2013)
29. Shafiee, S., Kamangar, F., Athitsos, V., Huang, J.: The role of dictionary learning on sparse representation-based classification. In: Int. Conf. on PErvasive Technologies Related to Assistive Environments. (2013)
30. Aharon, M., Elad, M., Bruckstein, A.: K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. Trans. Sig. Proc. **54** (2006)
31. Engan, K., Aase, S.O., Hakon Husoy, J.: Method of optimal directions for frame design. In: Int. Conf. of Acoustics, Speech, and Signal Processing. (1999)
32. Nourbakhsh, F., Bulò, S.R., Pelillo, M.: A matrix factorization approach to graph compression with partial information. Int. Journal of Machine Learning & Cybernetics (2015)
33. Nourbakhsh, F., Granger, E.: Learning of graph compressed dictionaries for sparse representation classification. In: Int. Conf. on Pattern Recognition Applications and Methods, ICPRAM,. (2016) 309–316
34. Yang, M., Van Gool, L., Zhang, L.: Sparse variation dictionary learning for face recognition with a single training sample per person. In: The IEEE Int. Conf. on Computer Vision (ICCV). (2013)
35. Wei, C., Wang, Y.F.: Undersampled face recognition via robust auxiliary dictionary learning. IEEE Trans. Image Processing **24** (2015)
36. 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: (CVPR) Workshops on Biometrics, IEEE (2011)
37. Huang, Z., Shan, S., Wang, R., Zhang, H., Lao, S., Kuerban, A., Chen, X.: A benchmark and comparative study of video-based face recognition on COX face database. IEEE Trans. Image Proc. (2015)
38. Yang, M., Zhang, L., Yang, J., Zhang, D.: Robust sparse coding for face recognition. In: Int. Conf. on Computer Vision and Pattern Recognition. (2011)
39. Elhamifar, E., Sapiro, G., Vidal, R.: See all by looking at a few: Sparse modeling for finding representative objects. In: IEEE Conf. on Computer Vision and Pattern Recognition,. (2012)
40. Donoho, D.L., Tsaig, Y.: Fast solution of l_1-norm minimization problems when the solution may be sparse. IEEE Trans. Information Theory **54** (2008) 4789–4812