Watch-List Screening Using Ensembles Based on Multiple Face Representations

Saman Bashbaghi, Eric Granger, Robert Sabourin
Laboratoire d'imagerie de vision et d'intelligence artificielle
École de technologie supérieure, Université du Québec, Montréal, Canada
bashbaghi@livia.etsmtl.ca, {eric.granger,robert.rabourin}@etsmtl.ca

Guillaume-Alexandre Bilodeau
LITIV Lab
École Polytechnique de Montréal, Canada
gabilodeau@polymtl.ca

Abstract*—Still-to-video face recognition (FR) is an important function in watchlist screening, where faces captured over a network of video surveillance cameras are matched against reference stills of target individuals. Recognizing faces in a watchlist is a challenging problem in semi- and unconstrained surveillance environments due to the lack of control over capture and operational conditions, and to the limited number of reference stills. This paper provides a performance baseline and guidelines for ensemble-based systems using a single high-quality reference still per individual, as found in many watchlist screening applications. In particular, modular systems are considered, where an ensemble of template matchers based on multiple face representations is assigned to each individual of interest. During enrollment, multiple feature extraction (FE) techniques are applied to patches isolated in the reference still to generate diverse face-part representations that are robust to various nuisance factors (e.g., illumination and pose) encountered in video surveillance. The selection of relevant feature subsets, decision thresholds, and fusion functions of ensembles are achieved using faces of non-target individuals selected from reference videos (forming a universal background model). During operations, a face tracker gradually regroups faces captured from different people appearing in a scene, while each user-specific ensemble generates a decision per face capture. This leads to robust spatio-temporal FR when accumulated ensemble predictions surpass a detection threshold. Simulation results obtained with the Chokepoint video dataset show a significant improvement to accuracy, (1) when performing score-level fusion of matchers, where patches-based and FE techniques generate ensemble diversity; (2) when defining feature subsets and decision thresholds for each individual matcher of an ensemble using non-target videos; and (3) when accumulating positive detections over multiple frames.

I. Introduction

The growing need for enhanced public security has driven the interest to integrate face recognition into decision support systems for video surveillance. Systems for FR in video surveillance (FRiVS) attempt to detect the presence of individuals of interest that are enrolled to the system. Accurate and timely responses are required to recognize faces captured under semi-controlled or uncontrolled conditions, as found at various checkpoint entries, inspection lanes, portals, etc. Faces captured under theses conditions incorporate variations in illumination, pose, scale, expressions, occlusion and blur [1]. Despite these challenges, it is generally possible to exploit spatiotemporal information extracted from video sequences to improve robustness and accuracy of FRiVS [2]. By tracking different faces in a scene, evidence from individual frames can be integrated over a video streams to reduce ambiguity.

Watch-list screening is an important application in video surveillance that involves still-to-video FR. During enrollment, facial regions of interests (ROIs) are extracted from reference still images that were captured under controlled condition to design facial models[1]. During operations, faces captured in videos are matched against the facial models of individuals enrolled to the watchlist, and an alarm is triggered if any matching score surpasses a individual-specific threshold [3]. Still-to-video FR is particularly challenging because very few reference samples are typically available for system design, and because ROIs captured with still cameras (during enrollment) have different properties than those captured with video cameras (during operations).

In pattern recognition literature, the situation where only one reference sample is available for system design are often referred to as a "single sample per person" (SSPP) or "one sample training" problem. Techniques specialized for SSPP in FR include multiple face representations, synthetic face generation, and enlarging the training set using auxiliary set [4]. In this paper, the SSPP problem found in still-to-video FR is addressed by exploiting multiple face representations, and in particular patch-based and FE techniques.

Systems for FRiVS are typically modeled in terms of independent detection problems, each one implemented using template matchers or using one- or two-class classifiers per person. These individual-specific detectors are designed with a mixture of references face samples from target and non-target individuals (from a cohort or the background model). The advantages of modular architectures with individual-specific detectors include the ease with which face models may be added, updated and removed from the systems, and the possibility of specializing feature subsets and decision thresholds to each specific individual [5], [6], [7]. Finally, given the limited and imbalanced number of reference samples, and the complexity of environments for FRiVS, individual-specific detectors have also been implemented using ensemble methods. The combination of a diversified pool of classifiers, has been shown to improve the overall system accuracy [7], [8]. However, designing discriminative ensembles based on multiple diverse face representations of a single target ROI sample may have a significant impact on the overall accuracy and robustness of still-to-video FR [4].

[1] A *facial model* of an individual is defined as a set of one or more reference face samples (used for a template matching system), or parameters estimated from reference samples (for a classification system).

This paper provides a performance baseline and guidelines to design individual-specific (multiple) classifier ensemble for still-to-video FR. During enrollment of an individual, the facial model is encoded into an ensemble of template matchers using a ROI extracted from a single high-quality reference still. A pool of matchers is generated from multiple diverse face representations extracted from the reference still. These representations are robust to various nuisance factors commonly found in video surveillance applications. Diversity among base template matchers is created by using different feature types and face patches. In this paper, the Local Binary Pattern (LBP), Local Phase Quantization (LPQ), Histogram of Oriented Gradient (HOG), and Haar FE techniques are applied to uniform non-overlapping patches isolated in the reference ROI [9], [10], [11].

Given the SSPP, the selection of relevant feature subsets, decision thresholds and fusion functions over multiple matchers and frames is achieved using ROI samples captured from faces of non-target individuals in reference videos. In particular, the design of ensemble-based systems for still-to-video FR can benefit considerably from the abundant reference samples of non-target individuals in the cohort (from a still camera) or from the background (from an operational video camera). During operations, a face tracker gradually regroups faces captured for different persons in a scene, while each individual-specific ensemble generates a prediction per face capture. This leads to robust spatio-temporal FR when the ensemble predictions accumulated over some window of time surpass a decision threshold.

In this paper, the impact on performance of using individual-specific ensembles based on multiple face representations, of defining feature subsets and decision thresholds and fusion functions using non-target videos, and of accumulating ensemble predictions over multiple frames are assessed using videos from the Chokepoint dataset [12].

II. Background on Still-to-Video FR

Few specialized techniques have been proposed for still-to-video FR [13]. A framework based on local facial features has been proposed to match stills against video frames with different features (e.g., manifold to manifold distance, affine hull method, and multi-region histogram) [13]. These features are extracted from a set of stills utilizing spatial and temporal video information. More recently, partial and local linear discriminant analyses have been proposed using a high quality still and a set of low resolution video sequences of each individual [14]. Finally, a specialized feed-forward neural network is trained for each individual of interest in a watch-list to identify the decision regions of individual faces in the feature space, where morphology is employed to synthetically generate variations of a reference still [6].

Multiple face representations of a single ROI reference sample may provide diversity of opinion. Patch-based techniques also provide multiple representations, and are typically used to recognize partially occluded faces. With patch-based methods, facial ROIs are divided into several overlapping or non-overlapping regions called patches, and then features are extracted locally from each patch for recognition purposes. Some specialized decision fusion techniques have been also introduced in [15], [16] for patch-based FR. It is worth noting that the still-to-video FR systems from literature assume that the single face reference is consistent and representative of the individuals in operational conditions.

III. Ensembles with Multiple Representations

In this paper, the system proposed for still-to-video FR is comprised of an ensemble of matchers dedicated to each individual of interest in the watch-list. Face models are produced during ensemble design phase. Various FE techniques are applied to face patches isolated in the single reference ROI sample to generate multiple face representations. During operations, template matching is performed to detect the presence of watch-list individuals, using faces captured over time from a network of video cameras.

As illustrated in Figure 1, frames captured by a video camera may include several people. For each frame, gray-scale conversion is first performed and then segmentation is applied in order to capture face(s). Then, the resulting ROI is scaled into a common size and aligned based on the location of eyes. Afterward, multiple feature vectors are extracted from the either entire ROI, for i = 1, 2, ..., M number of FE techniques, or each patch p = 1, 2, ..., P. Each matcher provides a similarity score $S_{i,p}(\mathbf{a}_{i,p}, \mathbf{m}_{i,p})$ between every patch of the input vector $\mathbf{a}_{i,p}$ and the corresponding patch template $\mathbf{m}_{i,p^j}$ in the gallery, where j=1, 2, ..., N indicates the number of individuals of interest, using suitable similarity metrics as reported in [11]. Scores output from matchers 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 $d_{i,p}$. The face tracker allows to regroup faces from each different person, and accumulate positive predictions over time for robust spatio-temporal recognition. In particular, decision fusion accumulates the number the decisions and provides the final decision d_j^* ($d_{i,p} = 1$ if fusion of scores surpasses $\gamma_{i,p}$, and $d_{i,p} = 0$ otherwise) of each ensemble over a fixed size window W according to:

$$d_j^* = \sum_{w=0}^{W-1} d_{i,p}[\mathbf{a}_{i,p(W-w)}] \quad \in [0, W] \tag{1}$$

In still-to-video FR, there is only one reference still per target individual to design a facial model, and that still has been captured using a still camera in a scene that is different from the operational environment. In addition to these inter-operability issues, the system must recognize faces captured under semi- or uncontrolled conditions, where faces vary due to pose, illumination, resolution. etc. However, it is possible to exploit an abundance of reference videos with non-target individuals (seen here as the generator of a universal background model) to optimize individual-specific feature subset, decision thresholds, and fusion functions. Since ROI extracted from those videos are close to ROIs seen during operations, they are exploited to optimize system parameters.

A. Feature Extraction and Selection

Employing multiple FE and selection methods can compensate the limited number of target samples. This is true to the extent that FE techniques are uncorrelated and robust to at least

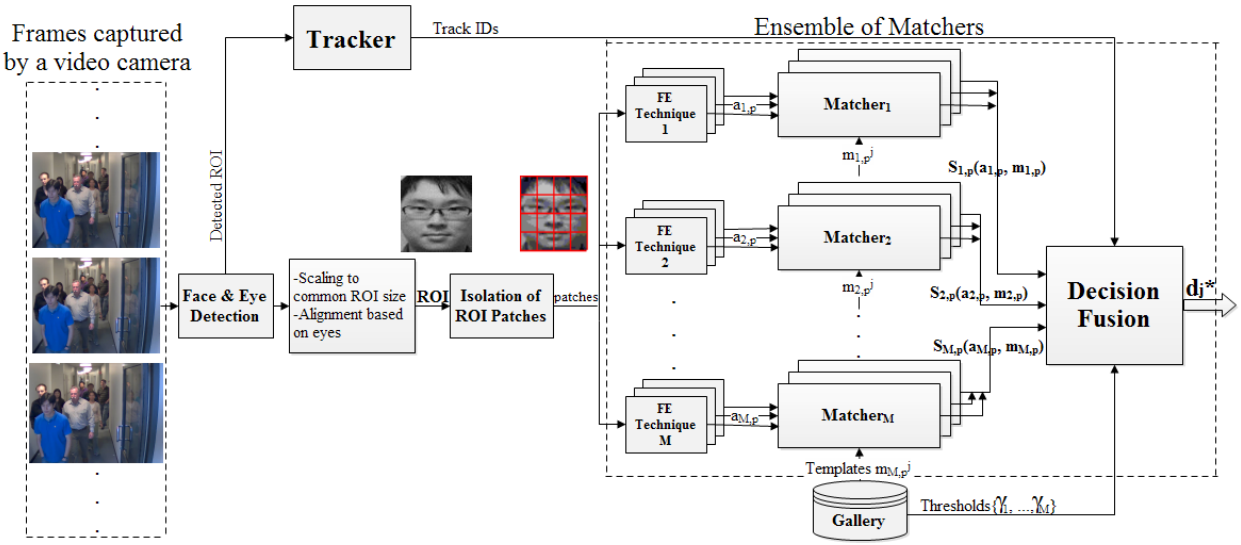

Fig. 1: Block diagram of still-to-video FR: the ensemble of matchers for one individual j in the watch-list.

one of the nuisance factors that may occur in VS environments such as changes in illumination, pose, etc. In Table I, different common FE techniques from literature are shown along with their robustness to various nuisance factors.

FE techniques have been chosen precisely based on their capability to overcome the nuisance factors.

Nuisance factors	Feature extraction techniques
Illumination	LDA, PCA, **LBP**, Gabor, LPP, **Haar**, SIFT, SURF, **HOG**
Pose	**Haar**, **HOG**
Scale	SIFT, SURF, Gabor, **HOG**
Motion Blur	**LPQ**
Occlusion	**Haar**, SURF (partial occlusion)

TABLE I: FE techniques selected in this paper to provide robust face representations under various conditions.

The rational for picking out at least one robust FE technique against each nuisance factor is that it may lead to a robust system in real-world VS environments (that is comprised of a mixture of nuisance factors). Hence, the LBP, LPQ, HOG, and Haar techniques are selected. Both LBP and LPQ extract textures of face in different ways. 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 edges information, which remains almost the same regardless of illumination change. HOG and Haar are selected to extract the information more related to shape. HOG 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. In particular, the Haar transform performs well under pose changes and partial occlusion.

These four techniques extract features from either the entire ROI or patches of the ROI. To improve discrimination and reduce feature subsets, the PCA technique is used to project and select features. Since PCA needs many representative samples to compute a projection matrix, either samples captured with a still camera of other individuals (in the cohort) or samples captured with a video camera (people captured in the operational background) must be employed.

B. Selection of Decision Thresholds

Since there are several representations for each individual, it is reasonable to determine specific thresholds for each representation, instead of defining a global threshold for each individual. It typically leads to higher accuracy and allows for decision-level fusion and moreover, accumulation over time. In this paper, using score distribution computed by comparing the single target representation (from a stills), and for a feature space, against all non-targets in the corresponding background (from videos), thresholds can be selected at a desired false positive rate (FPR) value depending on the application. The cumulative probability density function is utilized to determine appropriate thresholds. In this approach, the scores obtained for a given feature space are divided into several bins, where the number of bins corresponds to the number of unique scores. The number of samples with scores greater than the bin values is counted and then divided by the total number of samples. Once the cumulative curve is plotted, the thresholds can be defined.

C. Fusion over multiple matchers and frames

Fusion can be performed at a: (1) feature-level (concatenated all the extracted features into one discriminative feature vector), (2) score-level (combine the scores generated by multiple matchers to provide an overall score, and (3) decision-level (integrates the decisions of matchers after applying thresholds to produce the final Boolean output. Fusion at frame-level is also feasible using tracker for spatio-temporal recognition. ROI captures for different individuals are regrouped through face tracking. Predictions for each individual may be accumulated over time and if positive predictions surpass a detection threshold, then an individual of interest is detected.

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

Fig. 2: ROIs captured from the 'neutral' mugshot of 5 target individuals of interest.

IV. Experiments

A. Dataset and Experimental Protocol

In this paper the Chokepoint video dataset [12] is employed for validation of ensemble-based systems for watch-list applications. An array of three cameras is placed above two portals to capture videos of persons walking through them simultaneously, where videos incorporate changes in illumination, pose, scale, blur and occlusion.

To analyze the performance of the proposed systems, all video sequences (in both entering and leaving cases) from the Chokepoint dataset have been considered. Each video sequence views 25 subjects walking through a portal. In experiments, 5 persons are randomly selected to be in the watch-list, where there is just a single high quality frontal still (mug-shot) for design, along with video sequences of 10 unknown people that are assumed to reflect the background model. The remaining video sequences include 10 other unknown persons and 5 persons who are already enrolled in the watch-list. Furthermore, the size of mug-shots and captured ROIs are converted to gray-scale and scaled as 48x48 pixels due to limit processing time. However, validation data contain neutral mug-shots of 5 individuals of interest (see Figure 2) and videos of 10 individuals from the background model that are not neither in watch-list and test videos. This imitates the case in practice for real-world still-to-video FR systems. During design, ROIs of these individuals in the background are used as non-targets to define thresholds for individuals in the watch-list. The histogram intersection similarity measure is utilized for LBP, LPQ, and HOG, and Euclidean distance is used as a dissimilarity measure for Haar. The methodology for the proposed face screening system is formalized in Algorithm 1. In this protocol, it is assumed that each matcher provides a score and final decisions are produced through fusion.

To assess the transaction-level performance of a watch-list system, *Receiver Operating Characteristic (ROC)* space is considered. The proportion of target (genuine) 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, the proportion of non-target (imposter) ROI detected as individual of interest over the total number of non-target ROIs, is computed as false positive rate (FPR). A global scalar measure of detection performance is the area under the ROC curve (AUC). AUC can be interpreted as the probability of classification (matching) over the range of TPR and FPR. In order to estimate the performance of the system based on target ROIs, the precision-recall (PROC) space is also considered. To measure the performance in the imbalanced data situation, recall is the TPR and precision (PR) is computed as follows $PR = TP/(TP + FP)$. The AUPROC is suitable scalar measure to illustrate the accuracy of the system in the skewed imbalanced data circumstances.

Algorithm 1 Testing protocol of face screening system.

Repeat for each frame in the video sequence
for $n = 1$ to $NumberOfFrames$ **do**
  Repeat for each subject in the video sequence
  for $j = 1$ to $NumberOfIndividuals$ **do**
    Set watch-list = individual of interest, j, with all
    corresponding template $\mathbf{m}_j$ from Gallery
    Apply Viola-Jones face detector [17] to frame n
    for $r = 1$ to $NumberOfROIs$ **do**
      Extract uniform patches
      for $p = 1$ to $NumberOfPatches$ **do**
        for $i = 1$ to $NumberOfFE$ **do**
          Perform FE to patch p of ROI r
          Set $\mathbf{a}_{i,p}$ = representation i of patch p
          for feature-, score-level fusion **do**
            PCA projection
            Score $\mathrm{S}_{i,p} \leftarrow$ matching between
            input $\mathbf{a}_{i,p}$ and templates $\mathbf{m}_{i,p}$
            Normalize scores $\mathrm{S}_{i,p}$
          end for
          for trajectory-level analysis **do**
            if $\mathrm{S}_{i,p} \geq \gamma_{i,p}$ **then**
              Decision $\mathrm{d}_{i,p} \leftarrow$ True
            end if
          end for
          Final Decision $\mathrm{d}_j^* \leftarrow$ trajectory-level
          fusion of $\mathrm{d}_{i,p}$
        end for
      end for
    end for
  end for
end for

In transaction-level analysis, the average performance of FE techniques and feature selection using PCA, as well as, patch-based are analyzed as presented in Table II, along with standard errors. It shows the pAUC(20%) and AUPROC accuracy for each FE technique with all features, features selected after PCA, and patch-based with 16 non-overlapping patches of size is 12x12. It should be noted that matching scores are normalized using average cohort normalization method employing universal background model. Performance is provided for each individual of interest in the watch-list for all video sequences. After PCA, the first 32, 64, and 128 ranked features are selected. In the patch-based method, the output scores obtained from each patch are combined using mean function to produce global score based on different FE techniques.

B. Experimental Results

As shown in Table II, HOG and Haar feature types significantly outperform the other FE techniques. However, it can be concluded that the number of features used by HOG and Haar would have a direct impact on the time complexity, since the number of features in LBP and LPQ is less than with HOG and Haar. By applying PCA after FE techniques and selecting feature subsets, PCA improves the performance slightly in most cases, since it may generate more discrim-

Individual of Interest	Individual Face Representations							
	LBP (max: 256 feature)		LPQ (max: 256 feature)		HOG (max: 500 feature)		Haar (max: 2300 feature)	
	pAUC(20%)	AUPROC	pAUC(20%)	AUPROC	pAUC(20%)	AUPROC	pAUC(20%)	AUPROC
Person #1:								
PCA (32)	31.5±1.1	34.9±2.1	34.4±0.8	33.8±2.2	68.3±0.9	61.1±6.3	58.5±1	57.6±5.2
PCA (64)	38.5±1.2	37.6±.2	41.2±0.9	37.4±2.4	68.7±0.6	61.3±6.7	61.5±1.1	61.7±4.3
PCA (128)	39.2±1	40.4±2.1	49±0.8	40.4±2.1	69.3±0.8	61.5±5.4	59.5±0.9	59.3±4.4
All Features	33.5±1	29.8±2.2	39.5±0.6	35.9±1.3	72±1.1	71.4±4.3	52±0.8	51.5±4.3
Patch-based	51.5±0.9	50±1.9	55.9±1.3	56.8±2.2	82.9±0.3	69.7±0.1	97.9±0.4	70.9±1.1
Person #2:								
PCA (32)	41.5±0.7	44.9±1.8	33.5±0.8	33.4±4.6	51.5±1.2	48.2±5.2	62.5±1.3	58.2±4.3
PCA (64)	50±0.5	50.2±2	36.5±0.6	36.7±4.2	60±1.3	55.4±5.6	61.6±1.1	55.2±4.5
PCA (128)	52±0.4	50.1±1.8	39.5±0.6	37.4±4	59.6±1.3	56.4±5.2	64.5±1.1	59.7±4.2
All Features	48±0.6	43.6±1.5	40±0.5	36.5±1.1	68±1.9	55.6±6.1	59.5±0.7	47.8±4
Patch-based	54.1±0.6	51.5±2.7	41.3±1.2	45.4±2.1	83.1±0.2	62.8±0.8	99±0.9	63.9±1.3
Person #3:								
PCA (32)	39±0.5	37.2±3.1	34±0.3	29.4±2.3	59.1±0.4	44.3±5.4	54.7±0.8	55.6±5.1
PCA (64)	37.5±0.7	37.3±3.4	35.3±0.5	29.6±2.4	59.8±0.5	44.9±5.6	56.9±1	56.4±4.8
PCA (128)	40.1±0.7	38.3±3.3	35±0.4	34.8±2.3	62±0.5	50.9±5.6	58.8±0.9	55.7±5
All Features	32.5±0.4	30.4±0.7	28.5±0.5	26.3±0.7	63±1.5	50.5±7	56.5±0.8	53.4±4.3
Patch-based	40.5±0.8	39.7±2.5	37.9±1.1	35.3±1.9	73.5±0.3	52.6±1.2	98.9±0.2	65.9±1.7
Person #4:								
PCA (32)	34±0.3	33.4±1.2	36±0.5	33.8±3.1	64.5±0.9	49.5±3.5	58.3±1.3	52.2±4.3
PCA (64)	36.9±0.4	35.5±1.4	38.5±0.4	35.8±3	69.5±0.8	50.4±3.7	59±1.1	57.2±4.5
PCA (128)	37.5±0.5	37±1.3	39.8±0.5	37.9±3.4	71.2±0.8	52.4±3.5	66.1±0.9	62.4±4.5
All Features	33.5±0.4	32.6±0.5	36.5±0.4	33.9±0.7	66.5±1.6	52.9±4	59.5±1.4	53.7±3.2
Patch-based	37.5±0.7	38.9±2.2	44.7±1.2	47.6±2.7	78.7±0.2	56.5±1.1	99.8±0.2	66.1±1.1
Person #5:								
PCA (32)	38.5±0.8	38.3±2.8	44.5±0.7	44.8±3.3	64.5±0.9	63.7±6.2	56.1±0.8	53.3±4
PCA (64)	39±1.1	39.5±3.5	45.4±0.6	45.3±3.1	70.5±1.1	68.2±5.8	60.4±1.2	57.6±3.3
PCA (128)	39.8±0.9	40.2±3.6	46.5±0.6	46.8±3.4	69.3±1	68.5±5.4	66.5±1.1	63.1±3.6
All Features	35.5±0.5	34.8±1.2	45.6±0.5	45.6±0.8	62.5±2	61.2±8.3	58.5±1.4	54.7±4.3
Patch-based	38.5±0.7	40.4±3	45.9±1.2	49.5±2.8	83.3±0.2	70.4±1.1	97.7±0.3	78.9±1.6

TABLE II: Average transaction-level analysis for individuals of interest using 4 different FE techniques with all features and those selected features by PCA and Patch-based.

Individual of Interest	Multiple Face Representations					
	Feature-level (concatenation)		Score-level (mean fusion)		Patch-based score-level (mean fusion)	
	pAUC(20%)	AUPROC	pAUC(20%)	AUPROC	pAUC(20%)	AUPROC
#1	78.5±1.1	76.9±4.2	92.5±0.9	90.3±5.1	99.5±0.1	99.7±0.9
#2	76.3±1.6	75.1±4.3	92.3±1.1	90.7±6.7	99.4±0.2	95.2±1.8
#3	69.8±1.5	68.7±4.2	90.5±1.3	89.4±6.6	99.2±0.4	98.9±1.6
#4	76.3±1.2	74.4±3.8	91.6±1.2	89.9±6.5	99.4±0.1	98.3±1.4
#5	72.4±1.7	73.3±4.6	90.1±1.5	90.2±7.1	99.1±0.1	99.5±2.6

TABLE III: Average transaction-level analysis for individuals of interest using feature and score-level fusion.

inative representations. Applying PCA on the concatenated feature vectors provides a higher level of performance in contrast to their simple feature vectors $\mathbf{a}_{i,p}$ from individual FE techniques. The results achieved with the patch-based method greatly outperforms others in all cases. Since the features are extracted from each patch, it generates matching relies on more discriminant information.

Fusion at feature- and score-level for FE techniques are presented in Table III along with patch-based score-level, respectively. As presented in Table III, fusion at score-level using simple mean function greatly outperforms each FE individually, as well as, feature-level fusion. It improves the performance for with or without the use of patches. Using concatenated features from different FE techniques, applying PCA, and selecting the first 128 features (feature-level fusion) is outperformed by score-level fusion. The performance achieved by FE techniques solely (Table II), is always lower than with score-level fusion. It also confirms that patch-based method outperforms FE techniques at the score-level.

In trajectory-based analysis, ROIs are captured for 10 unknown individuals and 1 individual of interest during tracking, where for each individual has about 35 ROIs are captured in the sequence. The proposed system accumulates the positive detections (decisions) for each individual-specific ensemble over a window of 30 frames along a trajectory using fusion at decision-level. This experiment is repeated for 5 individuals of interests in the watch-list employing P1E_S2_C2 video sequence. The ROC curves produced by varying detection thresholds from 0 to 30 are plotted and the AUROCs are computed as presented in Table IV using both FE techniques and patch-based method.

	Individual of Interest				
	#1	#2	#3	#4	#5
Without patches	87.3%	73.6%	72.7%	73.1%	76.4%
Patch-based	92.2%	79.4%	80.2%	81.3%	83.8%

TABLE IV: AUROC based on accumulative positive detections for each individual of interest.

As an example, the accumulation over time of detections for individual #1 (blue curve) and all non-targets in the scene are illustrated in the Figure 3 (left). This video sequence

captures ROIs from individual #1, and 10 other unknown individuals. The corresponding ROC performances of the system is shown by the blue curve in Figure 3 (right). Using patch-based (the green curve represented in this figure) outperforms the proposed system without patches. As shown in Figure 3, some individuals in the scene are also detected falsely, specially in frames between 1650 and 1685. It can be seen that these ROIs correspond to an unknown person that is highly similar to the individual of interest.

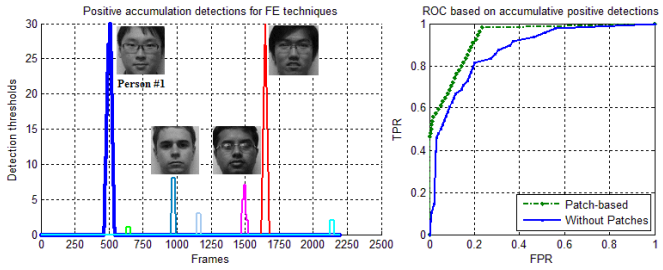

Fig. 3: An example of the accumulated ensemble decisions over a video sequence, and the corresponding ROCs for individual #1 using patch-based and without patches.

V. Conclusions

This paper presents a 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, different FE techniques and local patches (uniform patching) are employed to generate a diversified ensemble of matchers per individual. Results suggest that using multiple face representations for design of facial models are robust to the variety of nuisance factors encountered in video surveillance environments.

In this paper, low-level results are presented to observe the performance of the watch-list screening system. Simulation results with the Chokepoint video data indicate that the integration of different patches-based and feature type representations into an individual-specific ensemble provides a significantly higher level of performance than any single representation. Results also indicate that score-level fusion of patches outperforms score- and feature-level fusion of FE techniques defined by multiple face representations because of extracting features from local parts instead of entire face region. Since there is one reference still per individual of interest, videos of unknown non-target individuals in the scene have been used to define individual-specific decision thresholds and feature subsets. Hence, videos of background model are more representative of real scene, contrary to other stills in the cohort model. Finally, accumulating ensemble predictions over multiple face captures of corresponding individuals using a high quality track that are provided by the face tracker significantly improves the overall performance.

In order to achieve higher performance, future research can utilize random subspace methods to perform matching based on subsets of each representation, selecting the best matchers/ensemble of matchers dynamically, such as exploiting dynamic classifier/ensemble selection methods. Contextual quality of captured ROIs may also be incorporated to weight the output of matchers due to dynamically selecting different fusion strategy.

VI. Acknowledgement

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

References

[1] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas, "Face recognition from video: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 26, no. 05, 2012.

[2] F. Matta and J.-L. Dugelay, "Person recognition using facial video information: A state of the art," *Journal of Visual Languages & Computing*, vol. 20, no. 3, pp. 180–187, 2009.

[3] R. Chellappa, P. Sinha, and P. J. Phillips, "Face recognition by computers and humans," *Computer*, vol. 43, no. 2, pp. 46–55, 2010.

[4] M. Kan, S. Shan, Y. Su, D. Xu, and X. Chen, "Adaptive discriminant learning for face recognition," *Pattern Recognition*, vol. 46, no. 9, pp. 2497–2509, 2013.

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

[6] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi, "Toward development of a face recognition system for watchlist surveillance," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 10, pp. 1925–1937, 2011.

[7] C. Pagano, E. Granger, R. Sabourin, and D. O. Gorodnichy, "Detector ensembles for face recognition in video surveillance," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, 2012, pp. 1–8.

[8] E. Granger, W. Khreich, R. Sabourin, and D. O. Gorodnichy, "Fusion of biometric systems using boolean combination: an application to iris–based authentication," *International Journal of Biometrics*, vol. 4, no. 3, pp. 291–315, 2012.

[9] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 28, no. 12, pp. 2037–2041, 2006.

[10] O. Deniz, G. Bueno, J. Salido, and F. De la Torre, "Face recognition using histograms of oriented gradients," *Pattern recognition Letter*, vol. 32, no. 12, pp. 1598–1603, 2011.

[11] M. Bereta, W. Pedrycz, and M. Reformat, "Local descriptors and similarity measures for frontal face recognition: A comparative analysis," *Journal of Visual Communication and Image Representation*, vol. 24, no. 8, pp. 1213–1231, 2013.

[12] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "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*. IEEE, 2011, pp. 74–81.

[13] C. Shaokang, M. Sandra, H. Mehrtash T, S. Conrad, B. Abbas, L. Brian C *et al.*, "Face recognition from still images to video sequences: A local-feature-based framework," *EURASIP journal on image and video processing*, vol. 2011, 2011.

[14] Z. Huang, S. Shan, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset," in *Computer Vision–ACCV 2012*. Springer, 2013, pp. 589–600.

[15] B. Topcu and H. Erdogan, "Decision fusion for patch-based face recognition," in *Pattern Recognition (ICPR), 2010 20th International Conference on*. IEEE, 2010, pp. 1348–1351.

[16] S. Nikan and M. Ahmadi, "Human face recognition under occlusion using lbp and entropy weighted voting," in *"Pattern Recognition (ICPR), 2012 21st International Conference on"*. IEEE, 2012, pp. 1699–1702.

[17] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.