Video Face Recognition From A Single Still Image Using an Adaptive Appearance Model Tracker

M. Ali Akber Dewan
School of Computing and Information Systems, Athabasca University
Edmonton, Canada
adewan@athabascau.ca

E. Granger, R. Sabourin
Department of Automated Production Engineering, École de technologie supérieure, Montreal, Canada
eric.granger@etsmtl.ca,
robert.sabourin@etsmtl.ca

G.-L. Marcialis, F. Roli
Department of Electrical and Electronic Engineering, University of Cagliari
Cagliari, Italy
marcialis@diee.unica.it
roli@diee.unica.it

Abstract: **Systems for still-to-video face recognition (FR) are typically used to detect target individuals in watch-list screening applications. These surveillance applications are challenging because the appearance of faces changes according to capture conditions, and very few reference stills are available a priori for enrollment. To improve performance, an adaptive appearance model tracker (AAMT) is proposed for on-line learning of a track-face-model linked to each individual appearing in the scene. Meanwhile, these models are matched over successive frames against stored gallery-face-models, extracted from reference still images of each target individual (enrolled to the system) for robust spatiotemporal FR. In addition, compared to the gallery-face-models produced by self-updating FR systems, the track-face-models (produced by the AAMT-FR system) are updated from facial captures that are more reliably selected, and can incorporate greater intra-class variations from the operational environment. Track-face-models allow selecting facial captures for modeling more reliably than self-updating FR systems, and can incorporate a greater diversity of intra-class variation from the operational environment. Performance of the proposed approach is compared with several state-of-the-art FR systems on videos from the Chokepoint dataset when a single reference template per target individual is stored in the gallery. Experimental results show that the proposed system can achieve a significantly higher level of FR performance, especially when the diverse facial appearances captured through AAMT correspond to that of reference stills.**

I. INTRODUCTION

Still-to-video FR is an important function in several video surveillance applications, most notably in watch list screening [1]. Given one or few reference facial stills of target individuals enrolled to the system, still-to-video FR seeks to detect their presence in archived or live videos captured with surveillance cameras. Initially, a facial region of interest (ROIs) is isolated within each reference still through segmentation, and discriminant ROI patterns (face descriptors) are extracted to design a gallery-face-model[1] (GFM). These stills are assumed to be high quality mug shots taken under controlled conditions. Then, during operations, ROI patterns of faces captured in videos are matched against each GFM, and the operator is alerted if the matching score surpasses a decision threshold. In video surveillance, persons in the scene may be tracked and matching scores may be accumulated over a facial trajectory (a group of ROIs that correspond to the same high quality track of an individual) for robust spatiotemporal FR [2].

The performance of systems for still-to-video FR typically declines due to variations in capture conditions (e.g., pose, resolution, scale, illumination, blur, and expression) and to camera inter-operability. Moreover, only one or very few reference stills are available during enrollment to design representative the GFM of a target individual. Therefore, GFMs can incorporate limited intra-class variability for face matching.

Single sample per person (SSPP) problems [3] refer to the situation where only one reference pattern is available to design a pattern recognition system. To deal with SSPP problems, methods for adaptation, multiple face representation, or synthetic face generation may provide more representative GFMs. However, these methods may require considerable computational resources, corrupt GFMs if they are incorrectly updated, and assume that reference stills are representative of faces to be captured in videos. They may only incorporate limited information on the variations and uncertainties of faces to be seen in complex operational environment [4].

This paper presents a still-to-video FR system based on Adaptive Appearance Model Tracker (AAMT-FR), where a

[1] A GFM of an individual is defined as a set of reference ROI patterns (for a template matching), or a set of parameters estimated using reference ROI patterns (for a pattern classification) captured through segmentation.

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.38

track-face-model[2] (TFM) is learned during operations for each different person appearing in a camera view point. For online learning of tracked faces, the Sequential Karhunen Loeve method [5] is used within a particle filter-based tracker. At each frame, the TFMs of each person is updated and matched against the GFMs of every individual enrolled to the system. Given that face tracking allows regrouping faces of each person, the matching scores are accumulated over a person's facial track, and compared with an individual-specific decision threshold for robust spatiotemporal recognition.

Though such TFMs have successfully applied to the data association problem in adaptive appearance model tracking [6], to our knowledge these models have never been directly used for FR. TFMs have a number of advantages over GFMs in still-to-video FR applications. They may integrate a greater diversity of information on the intra-class variations of face appearance in a scene than with GFMs, especially when only one or few reference stills are available for face modeling. The facial representation incorporated in a TFM is captured through tracking, from videos in the operational scene, while GFMs are typically produced from a reference still captured under controlled conditions.

In this paper, experimental results were obtained using Chokepoint dataset [7] in which videos are captured under semi- and uncontrolled conditions. Performance of the proposed AAMT-FR system is compared at the transaction and trajectory levels against four other reference systems for still-to-video FR – Template Matching (TM [8]), Template Matching with Self-Update (TMSU [9]), Sparse Variation Dictionary Learning (SVDL [10]), and Multiple Face Representation (MFR [4]).

II. Background – still-to-video FR

Still-to-video FR is performed on video streams that are captured under semi- or controlled conditions across a network of surveillance cameras. In this paper, a generic system for still-to-video FR is comprised with four functional modules: *segmentation*, *tracking*, *classification* and *fusion*.

During enrollment of a target individual, the *segmentation* (or face detection) module isolates ROIs from one or more reference stills. Discriminant features are extracted and concatenated into reference ROI patterns for the design of user-specific GFMs. During operations, each camera captures a video stream that provides a particular viewpoint of individuals populating the scene. For each frame, ROIs are isolated and undergo the same feature extraction process to form input ROI patterns. The *classification* module measures the similarity between input ROI patterns and the GFMs of each individual enrolled to the system. The *face tracking* module initiates a new track once the segmentation module detects a new face in a different location than others. Then, it follows faces captured in successive frames and associates them to the same track. Finally, the *decision fusion* module integrates the matching scores for a face track according to each GFM, and compares with user-specific decision threshold for robust spatiotemporal recognition. It outputs a list of likely target individuals associated with each track.

TM [8] is a reference system for still-to-video FR, where the GFM of each target individual l ($l = 1, \dots, L$) consists of a single reference template $\mathbf{b}^l = (b_1^l, \dots, b_p^l)$. During enrolment, features are extracted from a ROI captured within a single reference still. For each input ROI r ($r = 1, \dots, R$) detected in a video frame $\mathbf{I}_t$, an input ROI pattern $\mathbf{a}_t^k = (a_1^k, \dots, a_p^k)$ linked to face track k is extracted and compared using some similarity measures against all templates $\mathbf{b}^l$ of target individuals in a p-dimensional subspace $\mathbb{R}^p$.

TMSU [9] is another reference system that allows adapting GFMs over time using highly confident operational data in order to increase FR robustness. It performs self-updating by comparing matching scores to a second (usually higher) update-threshold and selecting high confidence input ROI patterns to update the corresponding GFMs. Self-updating is limited on a single trait (e.g., face) to update the GFM of a target individual. To improve the limited representativeness of GFMs, some other techniques have been proposed in SSPP literature, which include multiple face representations, synthetic face generation, and enlarging training set using auxiliary data sets. Using multiple face representations, different face descriptors and patches or sub-images are extracted from a reference still to enhance GFMs for robust FR under various capture conditions [4]. Key issues for FR with multiple face representations are the diversity of representations and their fusion to make a decision, which increases the overall system complexity. Additionally, the original still (from which these representations are extracted) may not be representative of faces captures in videos.

In synthetic generation, multiple virtual face images are generated from a reference still to enhance GFMs. Multiple virtual views are synthesized by linear shape prediction, warping, morphing, symmetry property, partitioning a face into several sub-images, affine transformation, noise perturbation, shifting, and active appearance modeling [3] [11]. By enlarging training set using auxiliary data sets, an auxiliary set containing multiple face appearance per person from the other individuals (called generic set) than the targets in the gallery is exploited to assist in learning the GFM. Sparse Variation Dictionary Learning (SVDL) [10] is an example of a sparse face modeling techniques using auxiliary data sets. A recurring problem with these methods is that they need prior knowledge to guide the generation of virtual views or variations of the face appearances, and the quality and realism of the virtual views are not guaranteed in the operational data. These methods may fail to predict many realistic and unobserved variations in face appearance in a real-world scene.

[2] A TFM of an individual is defined as a set of ROI patterns, or a set of parameters estimated using ROI patters, obtained by tracking an individual's facial appearance in a camera viewpoint over consecutive frames.

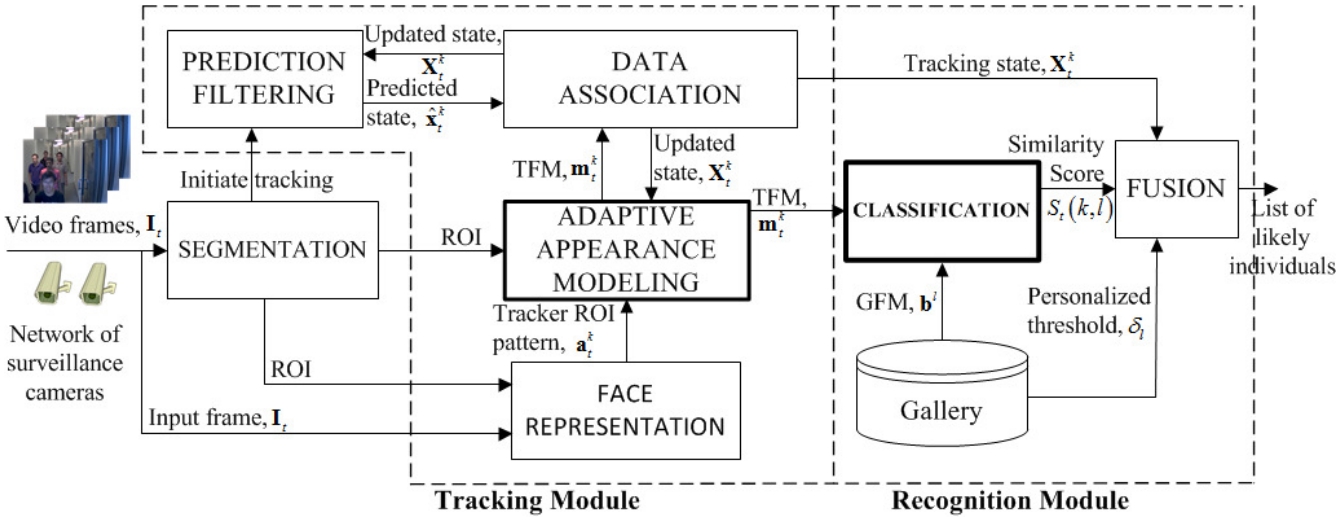


Fig. 1: Block diagram of the proposed AAMT-FR system.

III. Face Recognition System Based on AAMT

A new Adaptive Appearance Model Tracking-based FR system (AAMT-FR) is proposed for still-to-video FR. In the proposed system, a set of GFMs is designed as usual during enrollment, using the reference still images of target individuals. During operations, a TFM is learned online over successive frames for each different person appearing in the scene. These models gradually integrate diverse information on the facial appearance from the operational scene. Meanwhile, for each frame, these TFMs are matched against the GFMs of every target individual enrolled to the system. Matching scores linked to tracking target individual are then accumulated over time and compared with an individual-specific decision threshold for robust recognition. Fig. 1 shows the general block diagram of the proposed AAMT-FR system. It is comprised of *segmentation*, and of *face tracking* and *recognition modules*. Algorithmic description of the proposed system is presented in *Algorithm 1*.

A. Face Tracking

The face tracking module performs four main functions - *face representation*, *prediction filtering*, *adaptive appearance modeling*, and *data association*. The *segmentation* module may capture face region of interests, ROI_t^r, in each frame $\mathbf{I}_t$, where $r = 0, \dots, R$. Given a ROI_t^r captured in a new region of an input frame $\mathbf{I}_t$ during segmentation, the features are extracted in the tracker's *face representation* as a ROI pattern, $\mathbf{a}_t^k$. It allows initiating a TFM, $\mathbf{m}_t^k$, for a new track k. For existing tracks, the ROI pattern is extracted from the candidate region in $\mathbf{I}_t$ through *data association*. In Fig. 1, $\mathbf{a}_t^k$ is a ROI pattern representing the face captured for track k at a candidate region of frame $\mathbf{I}_t$.

During *prediction filtering*, the state of a face in a frame $\mathbf{I}_t$ is predicted based on information in the previous frames, and on some underlying model for state transitions. Given a *ROI* pattern $\mathbf{a}^k$ at frame $\mathbf{I}_t$, the input to the prediction filter is the previous state $\mathbf{X}_{t-1}^k$ and the output is a number of predicted states $\hat{\mathbf{X}}_t^k$ defining the possible new locations and sizes of the face at $\mathbf{I}_t$. A particle filter is used for predicting the new states during tracking [12].

Algorithm 1: Still-to-video FR using AAMT

Input: Input frames $\{\mathbf{I}_t : t = 1, \dots, \infty\}$, templates $\{\mathbf{b}^l : l = 1, \dots, L\}$ of target individuals enlisted in the gallery.

Output: List of likely individuals from watch-list in operational scene

for each frame $\mathbf{I}_t$, for $t = 1, \dots, \infty$, **do**
  – Apply segmentation to detect facial ROIs
  for each ROI_t^r, for $r = 0, \dots, R$, **do**
    Track Initialization: **if** the *ROI* is located in a different location than the existing tracks
      – Increment the number of tracks, $K \leftarrow K + 1$
      – Compute a new FTM $\mathbf{m}_t^k$ with the ROI for the newly initiated face track K
    end if
  end for
  for each FTM $\mathbf{m}_t^k$, for $k = 1, \dots, K$, **do**
    Tracking: – Compute state $\mathbf{X}_t^k$ of the face at frame $\mathbf{I}_t$ using tracking
    – Update the TFM $\mathbf{m}_t^k$ using new state information $\mathbf{X}_t^k$
  end for
  for each input pattern $\mathbf{a}_t^k$ associated with TFM $\{\mathbf{m}_t^k : k = 1, \dots, K\}$ **do**
    for each GFM $\mathbf{b}^l$, for $l = 1, \dots, L$, **do**
      Classification: – Compute score $S_t(k,l) = similarity(\mathbf{a}_t^k, \mathbf{b}^l)$
    end for
  end for
  for k = 1,…, K **do**
    for l = 1,…, L **do**
      – Accumulate scores over W consecutive frames by
      Fusion: $$acc\_S_t(k,l) = \frac{1}{W+1}\sum_{i=t-W}^{t} S_i(k,l)$$
      if $acc\_S_t(k,l) \geq \partial_l$ **then**
        – Detect or predict the appearance of watch-list individual l
      end if
    end for
  end for
end for

Adaptive appearance modeling inside the tracker generates TFMs $\mathbf{m}_t^k$ for newly initiated tracks and updates the models for existing tracks. Once a new track k is initially detected in the scene, the ROI patterns for the first n frames are tracked and captured using template matching. A data-block $\mathbf{A} = \{\mathbf{a}_1^k, \dots, \mathbf{a}_n^k\}$ is thereby defined using the tracked face regions with states $\{\mathbf{X}_1^k, \dots, \mathbf{X}_n^k\}$. Then, the TFM of the target face is generated as $\mathbf{m}_\mathbf{A}^k = \{\cup_\mathbf{A}^k, \bar{\mathbf{a}}_\mathbf{A}^k, \mathbf{\Sigma}_\mathbf{A}^k\}$, where $\cup_\mathbf{A}^k$ is the Eigen vector, $\bar{\mathbf{a}}_\mathbf{A}^k$ is the mean vector, and $\mathbf{\Sigma}_\mathbf{A}^k$ is the covariance matrix computed from the singular value decomposition (SVD) of the centered data matrix of data block, $\mathbf{A}$.

When a new data block $\mathbf{B} = \{\mathbf{a}_{n+1}^k, \dots, \mathbf{a}_{n+q}^k\}$ becomes available after tracking for q additional frames, the updated TFM, $\mathbf{m}_{\mathbf{A+B}}^k = \{\cup_{\mathbf{A+B}}^k, \bar{\mathbf{a}}_{\mathbf{A+B}}^k, \mathbf{\Sigma}_{\mathbf{A+B}}^k\}$ is obtained by using the augmented data matrix $[\mathbf{A}\ \mathbf{B}]$ through the computationally efficient Sequential Karhunen-Loeve (SKL) algorithm [5] as follows:

Step 1: Compute mean vectors, $\bar{\mathbf{a}}_\mathbf{B}^k = 1/m \sum_{i=n+1}^{n+q} \mathbf{a}_i^k$ and $\bar{\mathbf{a}}_{\mathbf{A+B}}^k = \frac{(f.n)}{(f.n)+q} \bar{\mathbf{a}}_\mathbf{A}^k + \frac{q}{(f.n)+q} \bar{\mathbf{a}}_\mathbf{B}^k$

Step 2: Form the matrix $\widehat{\mathbf{B}} = \left[\left(\mathbf{a}_{n+1}^k - \bar{\mathbf{a}}_\mathbf{B}^k\right) \dots \left(\mathbf{a}_{n+q}^k - \bar{\mathbf{a}}_\mathbf{B}^k\right) \sqrt{\{n.q/(n+q)\}} \left(\bar{\mathbf{a}}_\mathbf{B}^k - \bar{\mathbf{a}}_\mathbf{A}^k\right)\right]$

Step 3: Compute $\widetilde{\mathbf{B}} = \mathbf{orth}\left(\widehat{\mathbf{B}} - \mathbf{U}\mathbf{U}^T\widehat{\mathbf{B}}\right)$ and $\mathbf{R} = \begin{bmatrix} f\mathbf{\Sigma} & \mathbf{U}^T\widehat{\mathbf{B}} \\ \mathbf{0} & \widetilde{\mathbf{B}}\left(\widehat{\mathbf{B}} - \mathbf{U}\mathbf{U}^T\widehat{\mathbf{B}}\right) \end{bmatrix}$

Step 4: Compute the SVD of $\mathbf{R}$: $\mathbf{R} \xrightarrow{SVD} \widetilde{\mathbf{U}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{V}}$

Step 5: Finally, $\mathbf{U}_{\mathbf{A+B}} = [\mathbf{U} \quad \widetilde{\mathbf{B}}]\widetilde{\mathbf{U}}$ and $\mathbf{\Sigma}_{\mathbf{A+B}} = \widetilde{\mathbf{\Sigma}}$

Two key parameters – the forgetting factor, f, and batch size, q – determine the plasticity of TFMs over time. Parameter $f \epsilon [0\ 1]$ determines the contribution of older observations to be considered in updating the TFM, where $f = 1$ indicates no forgetting shall occur. Parameter q defines the batch size upon which a TFM is updated during tracking.

Data association compares the TFM $\mathbf{m}_t^k$ and the tracker ROI pattern $\mathbf{a}_t^k$ extracted from a predicted region (or state) defined by particle filter at frame $\mathbf{I}_t$. The region that gives maximum matching score is considered as the new location of the target face k at $\mathbf{I}_t$. Matching score between the face model $\mathbf{m}_t^k$ and the tracker ROI pattern $\mathbf{a}_t^k$ is measured using equation (1) to find face correspondences in consecutive frames.

B. Spatiotemporal Recognition:

Spatiotemporal recognition incorporates two main functions: *classification* and *fusion*. Beside these, the gallery contains the GFMs $\{\mathbf{b}^l : l = 1, \dots, L\}$ of the target individuals. For each frame $\mathbf{I}_t$, *classification* seeks to measure the similarity between each facial model $\mathbf{m}_t^k = \{\mathbf{U}, \bar{\mathbf{a}}_A, \mathbf{\Sigma}\}$ and all templates $\mathbf{b}^l$ in the gallery as follows:

$$S_t(k,l) = exp\{-\|(\mathbf{b}^l - \bar{\mathbf{a}}_t) - \mathbf{U}\mathbf{U}^T(\mathbf{b}^l - \bar{\mathbf{a}}_t)\|^2\} \quad (1)$$

The system's overall decisions are produced at the trajectory level. The fusion module accumulates the scores of a target k over the last W frames for each trajectory using:

$$acc\_S_t(k,l) = \frac{1}{W+1} \sum_{i=t-W}^{t} S_i(k,l) \quad (2)$$

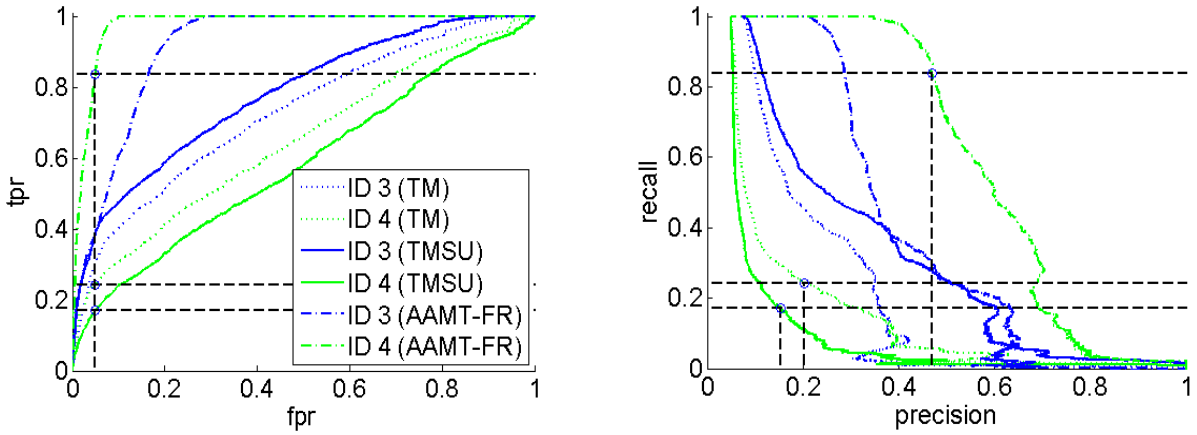
If the accumulated score for a target individual surpasses its decision threshold, ∂_l, the presence of the individual l is detected. The system flags all individuals of interest that are detected in the scene. An individual-specific decision threshold ∂_l is selected using the score distribution obtained by matching the GFM $\mathbf{b}^l$ to ROI patterns extracted from video tracks of non-target individuals at a user defined *fpr* of the cumulative probability density function [4].

IV. Experimental results

To compare the performance of FR systems, videos from the Chokepoint dataset [7] are used. Recorded as a video surveillance scenario, an array of 3 cameras is placed above different portals to capture individuals walking through in a natural way. The dataset contains 54 videos. Each one of the videos captures 29 individuals, where 23 are male and 6 are female. All videos are captured in two portals and 4 sessions, where the recordings of two portals are one month apart. Videos are captured at 30 *fps* and an image resolution is 800×600 pixels. Overall, the dataset contains 64,204 labeled face images each of which are cropped with size 96×96 pixels. This dataset is challenging for FR as the videos are captured under uncontrolled conditions with variations in pose, lighting, scale, and blur.

In experiments, faces are detected (segmentation) using Viola and Jones [13] algorithm. A particle filter based tracker [14] is used to follow the motion of faces, where the number of particles, x and y-translations, rotation, scaling, aspect-ratio, and skew direction changes are set to 9, 9, 0.05, 0.05, 0.005, and 0.001, respectively. The forgetting factor f and batch size q to update TFMs are set to 0.99 and 5, respectively. The facial ROIs are scaled into a common size of 48×48 pixels. For recognition, 81-dimensional Histogram of Oriented Gradient (HOG) features are extracted from each ROI and reduced into 32 using PCA projection.

The proposed AAMT-FR system is compared to four reference systems: TM [8], TMSU [9], SVDL [10], and MFR [4]. In TM, input ROI patterns are extracted from the ROIs detected in a frame and compared with all the GFMs using some similarity measure. The input ROI patterns are linked to tracking trajectory and accumulate the similarity scores over the trajectory for spatiotemporal recognition. Only one reference still is used to design the GFM for each target individual. In TMSU, a FR system similar to TM is employed, where the GFMs is changed adaptively over time. To update the GFMs, only those input ROI patterns are selected for which the similarity scores surpass a second update-threshold for the target individuals. In SVDL, the GFMs are generated from a sparse variation dictionary learned from single training samples per person, as well as an auxiliary dictionary of ROIs captures from non-target UBM individuals appearing in the scene. In MFR, multiple representations of the single sample per person are stored in the gallery as GFM. Multiple feature extraction techniques (LBP, LPQ, HOG, and Haar feature) are applied to patches



(a) ROC curves (b) Inverted PR curves

Fig. 2. ROC and inverted PR curves obtained with TM, TMSU, and AAMT-FR systems for individuals ID 03 and ID 04 with all Chokepoint videos.

isolated from the GFMs to generate diverse face-part representations. Finally, an ensemble of template matchers is applied on multiple face representations for FR.

The performance of systems is evaluated at transaction and trajectory levels. Transaction level analysis show the matching performance of the system based on ROI classification predictions. At trajectory level, information from the facial tracker is used to accumulate classification predictions according to trajectories corresponding to a same individual over a 1 second (W = 30 frames) window. Results are shown in the Receiver Operating Characteristic (ROC) and inverted Precision-Recall (PR) spaces. In ROC space, the partial Area Under Curve (pAUC) is observed for false positive rates (*fpr*) up to 5%. The area under the PR curves (AUPR) is also observed.

Fig. 2 shows an example of the ROC and inverted PR curves obtained at the transaction level when matching ROI patterns extracted from videos GFMs of target IDs 03 and 04. The dotted line in the Figures indicates the operating point at $fpr = 5\%$ related to target ID 04. It is seen from the figures that the AAMT-FR outperforms TM and TMSU. The improved performance can be attributed to the use of

TABLE I. AVERAGE PAUC (5%) AND AUPR PERFORMANCE FOR TM, TMSU, MFR, SVDL, AND AAMT-FR SYSTEMS AT THE TRANSACTION AND TRAJECTORY LEVELS ON ALL CHOKEPOINT VIDEOS.

Systems	Transaction Level		Trajectory Level	
	pAUC	AUPR	pAUC	AUPR
TM [8]	0.24 ± 0.04	0.35 ± 0.02	0.31 ± 0.01	0.36 ± 0.06
TMSU [9]	0.38 ± 0.05	0.39 ± 0.01	0.47 ± 0.03	0.47 ± 0.04
MFR [4]	0.42 ± 0.05	0.41 ± 0.01	0.51 ± 0.03	0.48 ± 0.04
SVDL [10]	0.44 ± 0.07	0.43 ± 0.02	0.54 ± 0.01	0.51 ± 0.05
AAMT-FR	$\mathbf{0.48 \pm 0.04}$	$\mathbf{0.55 \pm 0.03}$	$\mathbf{0.59 \pm 0.02}$	$\mathbf{0.59 \pm 0.07}$

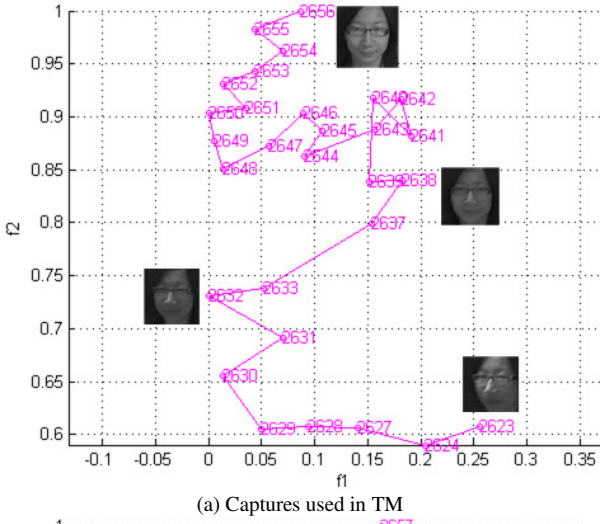
TFMs that incorporates intra-class variability on facial appearances captured during operations. Also, AAMT-FR allows selecting facial captures for face modeling more reliably because it exploits tracking information. TMSU outperforms TM for target IDs 03, when the GFMs for these individuals are updated with the correct operational data. Because of incorrect updates, the performance of TMSU declines for ID 04.

Table I presents the average pAUC (5%) and AUPR performance at the transaction- and trajectory-levels for TM, TMSU, SVDL, MFR and AAMT-FR systems over all the Chokepoint videos. To compare the global performance of systems, the experiments are repeated 10 times, each time randomly selecting five different targets individual (to form the watch list) and 10 other individual as non-targets. The table shows that the proposed AAMT-FR system outperforms others in all the cases.

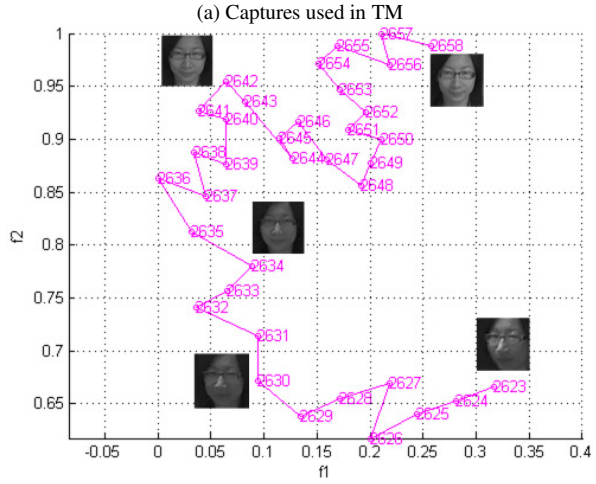
To observe the diversity of information incorporated into the facial captures by different systems, the ROIs for individuals with ID 12 are shown in Fig. 3 for the P1E_S1_C1 sequence. ROI patterns in the 81 feature HOG space of face captures are projected in a 2D space using Sammon mapping. Since, both TM and TMSU systems use ROIs captured through face segmentation, fewer high quality captures are available for modelling than AAMT-FR. Thus, TM and TMSU provide less diversity of appearances.

The computation effort required by the AAMT-FR is mainly found in steps for face model update during tracking. For face model update, the AAMT-FR uses the SKL algorithm [12] whose computational complexity is $O(dm^2)$, where d and m refer to the dimensionality of the input feature vectors and the number of new facial captures considered for face model update, respectively. For tracking, particle filter has been used, whose computational complexity is $O(N)$, where N is the number of particles re-sampled for a time instance by the filter [16].

In SVDL, the complexity of commonly used $l_1 =$ regularized sparse coding is $O(m+d)^{\varepsilon}$, where m is the number of dictionary atoms, d is the dimensionality of the features, and ε is an error constant. In TMSU, the main computation is required for GFM update, where the Eigen space is updated by re-computing the principal components matrix with the increased training set. This operation requires $O(d^3 + d^2 m)$ computations, where d and m refer to the dimensionality and the number of feature vectors used, respectively [17]. The computational complexity for TM is



(a) Captures used in TM



(b) Captures used in AAMT-FR

Fig. 3. Example of Sammon mapping of the facial captures processed by TM and AAMT-FR systems for ID 12 in the video P1E_S1_C1.

TABLE II. AVERAGE COMPUTATION TIME PER FRAME (SEC/FRAME) FOR TM, TMSU, MFR, SVDL, AND AAMT-FR SYSTEMS.

Systems	Time Complexity (sec/frame)	
	Recognition (without tracking)	**Recognition** (with tracking)
TM [8]	0.066 ± 0.06	0.191 ± 0.05
TMSU [9]	0.167 ± 0.05	0.292 ± 0.07
MFR [4]	0.125 ± 0.06	0.250 ± 0.08
SVDL [10]	0.112 ± 0.07	0.237 ± 0.05
AAMT-FR	N/A	0.217 ± 0.06

$O(d)$. It does not update face model and the computational complexity that it requires is mainly for template matching. In MFR, features are extracted from a uniform, non-overlapping patch configuration of 4× 4 (12 × 12 pixels). However, bigger size patches provide much information about the region but increase the complexity of processing.

The computation time per frame for different *FR* systems with/without using tracker are shown in Table II. In AAMT-FR, a tracker is always required to generate TFMs for FR. Thus, the computation time of AAMT-FR cannot be

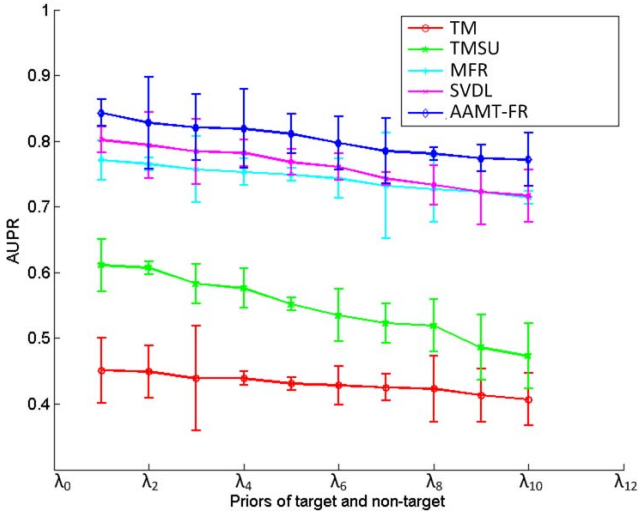
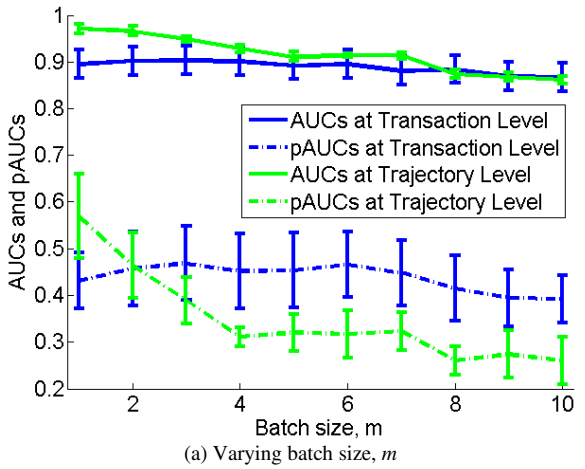


Fig. 4. The AUCs and pAUCs (5%) for different batch sizes and forgetting factors.

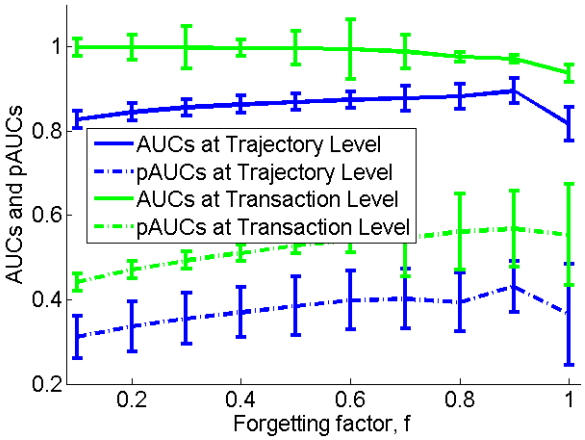
computed without considering tracking. At trajectory level, the performance for each system is improved over the transaction level because of score accumulation (see Table I); however, a tracker must be incorporated with the FR system for accumulating scores from the same individual. Thus, at trajectory level analysis, the total computation times include the times required for the tracking and the recognition.

The performance of the AAMT-FR is compared with the TM, TMSU, SVDL, and MFR systems considering all the entering and leaving sequences captured with Camera 2 (frontal or near frontal view). Average AUPRs for the systems at different priors of targets and non-targets, $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_{10}\} = \{1{:}\,1, 1{:}\,5000, \ldots, 1{:}\,50000\}$, are plotted in Fig. 4. The figure shows that the performance of all the systems declined as the level of target to non-target imbalance grows in the operational data. Performance for TMSU degrades sharply because of incorrect updates of the GFM. The AAMT-FR system outperforms the others because the FTM incorporates diversity information of the facial captures through tracking.

The impact of changing batch size m and forgetting factors f may have a considerable impact on the FR performance using the AAMT-FR. Fig. 5(a) shows the AUC and pAUC produced by AAMT-FR system while varying batch size, m. In this case m is changed from 1 to 10 while keeping the forgetting factor, $f = 0.9$. This figure shows that if m is increased, the performance declines as TFMs are updated after every m frames. Thus, m = 1 gives best performance for the AAMT-FR system, although this may increase the processing time. Fig. 5(b) shows the performance of the AAMT-FR while varying f between 0 and 1, while fixing m = 1. Here, $f = 0$ indicates forget everything whereas with the higher value of f, it allows to remember more past observations. AAMT-FR performance tends to increase with the value of f as it allows incorporating more diverse information of face appearance changes in the TFMs.



(a) Varying batch size, *m*



(b) Varying forgetting factor, *f*

Fig. 5. The AUCs and pAUCs (5%) for different batch sizes and forgetting factors.

V. Conclusion

This paper presents an AAMT-FR system for still-to-video FR. It is specialized for video surveillance applications like watch-list screening, given only one reference still per target person. Inside the system, an AAM tracker is used that gradually learns a track-face-model (TFM) per individual with the facial captures appearing in the scene. Meanwhile these models are matched over time against the gallery-face-models (GFMs), extracted from the reference still images from a gallery of the individuals of interests. Matching scores are accumulated over several frames and multiple cameras for spatiotemporal FR. Simulation results with the Chokepoint videos indicate that the AAMT-FR can provides a significantly higher level of performance than reference TM [8], TMSU [18], SVDL [19] and MFR [4] systems. Indeed, facial models learned with AAMT-FR select captures for face modeling more reliably and incorporates a greater variability of facial representation from the actual environment. However, the current TFM is used just as a proof of concept. The future direction of this research intends to introduce more robust TFMs, matching functions, and GFMs to improve the system's performance in FR.

Acknowledgement

This work was partially supported by the Natural Sciences and Engineering Research Council, the Ministère du développement économique, de l'innovation et de l'exportation du Québec, and Research Incentive Grant (Athabasca University), Canada.

References

[1] F. Matta and J. L. Dugelay, "Person recognition using facial video information," *Journal of Visual Languages and Computing,* vol. 20, no. 3, pp. 180-187, 2009.

[2] M. De la Torre Gomera, E. Granger, R. Sabourin and D. Gorodnichy, "Partially-supervised learning from facial trajectories for face recognition in video surveillance," *Information Fusion,* vol. 24, pp. 31-53, 2014.

[3] X. Tan, S. Chen, Z.-H. Zhou and F. Zhang, "Face recognition from a single image per person: A survey," *Pattern Recognition,* vol. 39, no. 9, pp. 1725-1745, 2006.

[4] S. Bashbagi, E. Granger, R. Sabourin and G.-A. Bilodeau, "Watch-list screening using ensembles based on multiple face representations," in *ICPR*, Stockholm, Sweden, 2014.

[5] A. Levy and M. Lindenbaum, "Sequential karhunen–loeve basis extraction and its application to images," *IEEE Transactions on Image Processing,* vol. 9, no. 8, pp. 1371-1374, 2000.

[6] M. A. A. Dewan, E. Granger, F. Roli, R. Sabourin and G.-L. Marcialis, "A comparison of adaptive appearance methods for tracking faces in video surveillance," in *ICDP*, London, UK, 2013.

[7] 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 *CVPRW*, Colorado, USA, 2011.

[8] R. Brunelli and T. Poggio, "Face recognition: features versus templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 15, no. 10, pp. 1042-1052, 1993.

[9] F. Roli and G. L. Marcialis, "Semi-supervised PCA-based face recognition using self-training," in *SSSPR*, Hong Kong, China, 2006.

[10] M. Yang, L. Van and L. Zhang, "Sparse variation dictionary learning for face recognition with a single training sample per person," in *ICCV*, Washington, USA, 2013.

[11] F. Mokhayeriy, E. Granger and G.-A. Bilodeau, "Synthetic face generation under various operational conditions in video surveillance," in *ICIP*, Quebec, Canada, 2015.

[12] J. Cho, S. Jin, X. Pham, J. Jeon, J. Byun and H. Kang, "A real-time object tracking system using particle filter," in *ICIRS*, Beijing, China, 2006.

[13] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision,* vol. 57, pp. 137-154, 2004.

[14] D. A. Ross, J. Lim, R.-S. Lin and M.-H. Yang, "Incremental learning for robust visual tracking," *International Journal of Computer Vision,* vol. 77, no. 1, pp. 125-141, 2008.

[15] J. D. Hol, T. B. Schon and F. Gustafsson, "On resampling algorithms for particle," in *NSSPW*, Cambridge, UK, 2006.

[16] A. Sharma and K. Paliwal, "Fast principal component analysis using fixed-point algorithm," *Pattern Recognition Letters,* vol. 28, pp. 1151-1155, 2007.

[17] F. Roli and G. L. Marcialis, "Semi-supervised PCA-based face recognition using self-training," in *Structural, Syntactic, and Statistical Pattern Recognition*, Hong Kong, China, 2006.

[18] M. Yang, L. Van and L. Zhang, "Sparse variation dictionary learning for face recognition with a single training sample per person," in *ICCV*, Washington, USA, 2013.