Pattern Recognition 49 (2016) 129–151

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Adaptive appearance model tracking for still-to-video face recognition

M. Ali Akber Dewan [a], E. Granger [b,*], G.-L. Marcialis [c], R. Sabourin [b], F. Roli [c]

[a] School of Computing and Information Systems, Athabasca University, Edmonton, Canada
[b] Laboratoire d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure Université du Québec, Montreal, Canada
[c] Department of Electrical and Electronic Engineering, University of Cagliari, Piazza d'Armi, Cagliari, Italy

ARTICLE INFO

Article history:
Received 20 December 2014
Received in revised form
23 June 2015
Accepted 5 August 2015
Available online 15 August 2015

Keywords:
Biometrics
Video surveillance
Face recognition
Watch-list screening
Single sample per person
Face tracking
Online and incremental learning
Adaptive appearance modeling

ABSTRACT

Systems for still-to-video face recognition (FR) seek to detect the presence of target individuals based on reference facial still images or mug-shots. These systems encounter several challenges in video surveillance applications due to variations in capture conditions (e.g., pose, scale, illumination, blur and expression) and to camera inter-operability. Beyond these issues, few reference stills are available during enrollment to design representative facial models of target individuals. Systems for still-to-video FR must therefore rely on adaptation, multiple face representation, or synthetic generation of reference stills to enhance the intra-class variability of face models. Moreover, many FR systems only match high quality faces captured in video, which further reduces the probability of detecting target individuals. Instead of matching faces captured through segmentation to reference stills, this paper exploits Adaptive Appearance Model Tracking (AAMT) to gradually learn a *track-face-model* for each individual appearing in the scene. The Sequential Karhunen–Loeve technique is used for online learning of these track-face-models within a particle filter-based face tracker. Meanwhile, these models are matched over successive frames against the reference still images of each target individual enrolled to the system, and then matching scores are accumulated over several frames for robust spatiotemporal recognition. A target individual is recognized if scores accumulated for a track-face-model over a fixed time surpass some decision threshold. The main advantage of AAMT over traditional still-to-video FR systems is the greater diversity of facial representation that may be captured during operations, and this can lead to better discrimination for spatiotemporal recognition. Compared to state-of-the-art adaptive biometric systems, the proposed method selects facial captures to update an individual's face model more reliably because it relies on information from tracking. Simulation results obtained with the Chokepoint video dataset indicate that the proposed method provides a significantly higher level of performance compared state-of-the-art systems when a single reference still per individual is available for matching. This higher level of performance is achieved when the diverse facial appearances that are captured in video through AAMT correspond to that of reference stills.

1. Introduction

Automatic face recognition (FR) is increasingly employed by public safety organizations to detect individuals of interest for enhanced security and situational awareness [1]. In decision support systems for video surveillance (VS), the human operator may rely on FR to detect the presence of target individuals captured over a network of surveillance cameras. Accurate and timely responses are required to recognize faces captured under semi-controlled and uncontrolled conditions, as found at various security checkpoints, inspection lanes, portals, etc. Faces captured under these conditions are subject to a variety of nuisance factors, including changes in illumination, pose, scale, expression, occlusion, and blur [2], and to camera interoperability issues. Despite these challenges, it is generally possible to exploit spatiotemporal information (e.g., tracking and multi-frame fusion) and camera arrays to improve robustness and accuracy in VS applications [1].

Face recognition in video surveillance is employed in a range of still-to-video and video-to-video applications. The still-to-video FR applications typically need to match faces in low-quality videos captured under unconstrained conditions against high quality still face images, whereas in video-to-video query video sequences are matched against a set of target video sequences [3]. Watch list (WL)

* Corresponding author.
E-mail addresses: adewan@athabascau.ca (M.A.A. Dewan), eric.granger@etsmtl.ca (E. Granger), marcialis@diee.unica.it (G.-L. Marcialis), robert.sabourin@etsmtl.ca (R. Sabourin), roli@diee.unica.it (F. Roli).

http://dx.doi.org/10.1016/j.patcog.2015.08.002
0031-3203/

screening is an important still-to-video FR application [4], where given one or few reference still images, FR is applied to WL screening seek to detect the presence of target individuals enrolled to the system. It is assumed that facial regions of interests (ROIs) are extracted from reference still images (high quality mug shots or ID photos) that were taken under controlled condition to design gallery-face-models. The gallery-face-model of an individual is defined as a set of one or more reference ROI patterns (used for a template matching system), or a set of parameters estimated using reference ROI patterns (for a pattern classification system). Then, during operations, ROI patterns of faces captured in videos are matched against the gallery-face-model of each individual enrolled to the WL gallery. The operator is alerted if any matching score surpasses an individual-specific threshold [1].

Systems for still-to-video FR applied to VS are typically modeled in terms of independent detection problems [5], each one implemented using template matching or a one- or two-class classifier per individual followed by thresholding. These individual-specific detectors are designed with reference ROI patterns from target, and possibly non-target individuals (from the cohort or universal background). The advantages of such modular architectures 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].

Still-to-video FR is particularly challenging because few reference stills are available for system design (face modeling), and because ROIs captured with still cameras (during enrollment) have different properties than those captured with video cameras (during operations) [1]. In pattern recognition literature, the situation where only one reference pattern is available for system design is referred to as a single sample per person (SSPP) problem [6]. This paper seeks to address the SSPP problem found in still-to-video FR with WL screening applications in mind.

Given a limited number of reference images, it is difficult to design representative gallery-face-models. For instance, when applying a common template matching (TM) system to WL screening, discriminant and compact features are extracted from reference facial ROIs to form template ROI patterns. Then the same features are extracted from faces captured in video frames, and matched against these templates. The performance of this still-to-video FR system may be poor, since templates provide a limited representation of faces to be recognized during operations [7,8]. To enhance gallery-face-models, techniques for adaptation, multiple-representation, synthetic generation, and enlarging the training data (using some auxiliary set) may be used to represent different face capture conditions [7–9]. These techniques however may fail to provide more representative gallery-face-models since they incorporate limited information on the intra-class variations and uncertainties of a face in the complex operational environment. The update and management of template galleries with faces captured during operations may improve intra-class variability of gallery-face-models and the FR performance, though these adaptive methods may corrupt the gallery-face-model if incorrectly updated [7,9].

Spatiotemporal FR systems rely on tracking to capture temporal information, and have been shown to improve performance over the traditional FR systems in VS [1,2]. Face tracking (FT) can play two important functions in video FR – (1) regroup ROIs of a person and integrate evidence (e.g., matching scores) from each frame and from multiple cameras of a video stream in order to reduce ambiguity of predictions [1,2]; (2) confirm the detection of highly confident facial regions in a frame for the segmentation process [10]. Though many algorithms have been proposed for object tracking in general, ones based on adaptive appearance modeling are well suited for FT. They learn internal *track-face-models* that adapt with the facial changes in the environment for enhanced data association [11,12].

Though track-face-models have been exploited for accurate data association in FT, to our knowledge these models have not been used for matching in video-based FR. Track-face-models have several potential advantages over gallery-face-models in still-to-video FR. A track-face-model may integrate a greater diversity of information on the variations of face appearance in a scene than gallery-face-model produced with one or few reference stills. The facial information incorporated in a track-face-model is captured from the specific operational scene (i.e., camera viewpoint) via tracking, which cannot be induced in a gallery-face-model that is produced from a reference still captured under controlled conditions, even if the model is enhanced through adaptation, synthetic generation, multi-face representations, or by enlarging the training set using non-target ROI patterns. Furthermore, by matching track-face-models (instead of a single ROIs from segmentation) with gallery-face-models, FR performance can be improved even if a limited number of reference stills are used to generate gallery-face-models. Since the track-face-model is updated within a tracker, it is more likely to update that model with faces from the same person in a scene without employing any additional gallery management technique. This is a challenging problem within adaptive biometric systems by themselves.

This paper presents a still-to-video FR system called the Adaptive Appearance Model Tracker-based Face Recognition (AAMT-FR), where a track-face-model is learned online (during operations) for each different person appearing in a camera view point. For online learning, Sequential Karhunen–Loeve method [13] is used within a particle filter-based tracker. At each frame, the track-face-model for each different person in the scene is updated and matched against the gallery-face-model of every individual enrolled to the system. Given that face tracking allows us to regroup faces of each person, the matching scores for a person are accumulated over a facial *trajectory*,[1] and compared with an individual-specific decision threshold for robust spatiotemporal recognition. During operations, track-face-models are updated incrementally, and improve their representativeness by incorporating diverse information on the facial appearance from the scene. Concurrently, the tracking information used to accumulate the matching scores over time also increases intra-class variability of face-track-models and improves FR discrimination.

Performance of the proposed system is evaluated with a generic still-to-video FR system for WL screening applications, where each gallery-face-model corresponds to a template (ROI pattern) extracted a priori from a high quality reference face still. Simulation results were obtained using video form the Chokepoint dataset [15], where an array of three cameras was placed above several portals to capture individuals walking through. These videos capture faces of individuals under semi- and uncontrolled conditions. Experiments compare the transaction- and trajectory-level performance of the AAMT-FR with respect to several state-of-the-art FR systems.

The organization of this paper is as follows. Section 2 presents a generic still-to-video FR system as needed for WL screening applications. Given the limitation of using a single reference still for designing gallery-face-model, a brief review on state-of-the-art adaptive biometric systems, face-modeling techniques developed for the SSPP problem, and spatiotemporal FR techniques are also presented in this section. The proposed AAMT-FR system for still-to-video FR is described in Section 3. In Section 4, the experimental methodology (dataset, protocol, and performance metrics) for validation of FR systems is described. Benchmarking results are presented and discussed in Section 5 with WL screening applications in mind.

[1] A trajectory is defined as a set of facial ROIs that correspond to a same high quality track of an individual across consecutive frames [14].

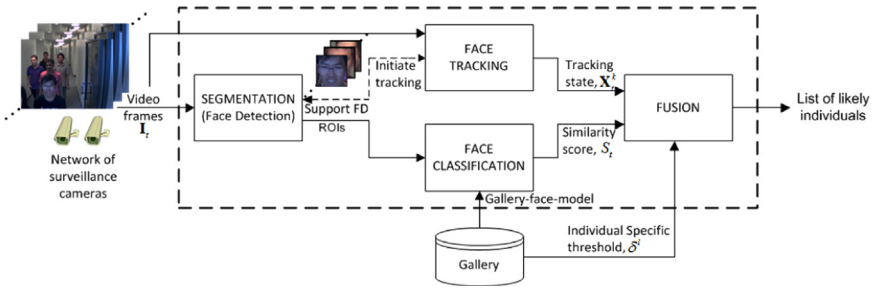

Fig. 1. A generic system for still-to-video face recognition.

2. Still-to-video face recognition

2.1. A generic system for spatiotemporal face recognition

The problem addressed in this paper is the design of a robust still-to-video FR system for WL screening, where individuals of interest must be detected under semi- or controlled conditions across a network of surveillance cameras. Fig. 1 shows a generic system for still-to-video FR according to a *track-and-classify* approach.

The system first performs segmentation to isolate the regions of interest (ROIs) corresponding to faces in each frame. During enrollment process, from one or more reference ROIs of a target individual, features are extracted and combined into reference ROI patterns to design user-specific gallery-face-models stored in the gallery. A gallery-face-model consists of a set of one or more templates or parameters of a classifier designed using reference ROI patterns. During operations, each camera captures video streams of 2D images or frames $\mathbf{I}_t$, and provides a particular viewpoint of individuals populating the scene. For each ROI, the same features are extracted and combined into an input ROI pattern for classification and to initiate tracking for new persons in the scene.

The *face tracking* module associates faces captured in successive frames to a same person (to define trajectories) and can also confirm ROIs found through segmentation. The classification module computes similarity of input ROI patterns against gallery-face-models. In spatiotemporal recognition, the fusion module accumulates the matching scores for a trajectory according to each gallery-face-model, and compares these accumulations with decision threshold for robust recognition. It outputs a list of likely target individuals associated with each trajectory. The main functions performed by the still-to-video FR system shown in Fig. 1 (*segmentation*, *tracking*, *classification*, and *fusion*) are described in more detail below.

The *segmentation* module (also referred to as the *face detection module*) allows isolating ROIs containing a face in each frame. At a discrete time t, the input for the segmentation module is a video frame $\mathbf{I}_t$ and output is the set of ROIs delimiting faces in $\mathbf{I}_t$, if any. The segmentation module may detect new faces at different locations than others, and corresponding ROIs are needed to initiate new tracks. Some pioneering work on face detection are presented in [16–18]. A common limitation of segmentation algorithms is their complexity for the video surveillance applications, which may be overcome with the Viola and Jones face detector [16].

The *face tracking* module initiates a new track once the segmentation module detects a new face in a different location from others. After the first detection at frame $\mathbf{I}_t$, the tracker generates an internal track-face-model for an individual k, and seeks to estimate its states $\mathbf{X}_{t+1}^k, \ldots, \mathbf{X}_{\infty}^k$ in future frames. A state $\mathbf{X}_t^k$ is typically represented as the set $\{x_t^k, y_t^k, w_t^k, h_t^k\}$, where the symbols x_t^k, y_t^k, w_t^k and h_t^k represent the x center location, y center location, width, and height of the bounding box of face k, respectively, for frame $\mathbf{I}_t$. During tracking, several states for a face may be predicted, each one representing a candidate region of a face in a frame. Some features used for face modeling are extracted from each of the candidate regions and are assembled into input patterns for data association with the track-face-model. The state $\mathbf{X}_t^k$ that achieves the highest matching score is considered as the actual state for face k on frame $\mathbf{I}_t$. The state information $\mathbf{X}_t^k$ is then used to predict the possible states of the face in the next frame $\mathbf{I}_{t+1}$. The states selected for face k provide temporal information about that face, such as the position and size in the frames, speed, and acceleration.

The tracker assigns a unique identifier for each face track so that the tracking module can follow the movement of a face in a sequence. Facial trajectories are formed along these tracks. When the track quality falls below some threshold, the track is dropped automatically. Track quality is often measured by comparing the predicted and new locations of a face in successive frame, or by measuring the difference between the appearances of a face at its track-face-model. Decisions to drop a track can be improved by cumulating the estimations of the tracking quality over several consecutive frames.

Adaptive appearance model-based trackers (AAMT) are shown to be efficient in FT [11,12]. AAMTs in the literature are categorized into two main types: generative and discriminative. Generative types [19–21] learn the appearance of faces and track them by searching for the region most similar to the target face appearances in each frame. They do not exploit any background information for data association. With discriminative types [22–24], trackers localize the target using a classifier that learns a decision boundary between the appearance of target and that of background and non-target ROI patterns. The limitations of AAMT are that they may computationally expensive and carry the risk of adapting models to other targets or background scenes. However, AAMT has been shown to efficiently address current challenges of FT because of using track-face-model that is updated incrementally.

The *face classification* module seeks to measure the similarity between input ROI patterns and the gallery-face-models of individuals enrolled to the system. During enrollment of a target individual, ROIs are extracted from still images captured under controlled environment. For each one, discriminant features are extracted and assembled into reference ROI patterns. ROI patterns are extracted by using some holistic or local face descriptors such as Local Binary Pattern (LBP), Principle Component Analysis (PCA), and Histogram of Oriented Gradients (HOG) [2]. During operations, ROIs are detected in successive frames $\mathbf{I}_t$ of a video, and then the same features are extracted into an input ROI pattern. Matching is typically performed using a template matcher or a pattern classifier. The resulting score $\mathbf{S}_t(k,l)$ indicates the likelihood that an input ROI captured in frame $\mathbf{I}_t$ corresponds to the gallery-face-model of target individual l enrolled to the system, for $l = 1, 2, \ldots, L$

along face track k. In TM, the gallery-face-model is defined as one or more templates extracted from the reference ROI patterns, while in pattern classification, it is defined as set of parameters estimated from the training ROI patterns of the gallery.

The *fusion* module integrates the responses from the tracking and classification over several frames, and possibly multiple-cameras, and outputs the predicted list of target individuals in the scene. If an accumulated score for a trajectory surpasses a individual-specific decision threshold, the corresponding individual of interest is considered to be detected. This trajectory-based analysis is typically used by the surveillance system to reduce ambiguity in video-based FR [14].

For simplicity, this paper considers a common still-to-video FR system based on template matching (TM), where only one reference still is available to design the gallery-face-model [8]. In addition, it is assumed that the gallery-face-model of each target individual $l = 1, \ldots, L$ consists of a single reference pattern $\mathbf{b}_l$ with features extracted from the single ROI captured in the reference still. For each input ROI r ($r = 1, \ldots, R$) detected in a frame $\mathbf{I}_t$, an input ROI pattern $\mathbf{a}_t^k = (a_1^k, \ldots, a_p^k)$ linked to face track k is extracted and compared using some similarity measure against all reference patterns $\mathbf{b}^l = (b_1, \ldots, b_p)^l$ of target individual l in a p-dimensional subspace $\Re^p$. Algorithmic description for a track-and-classify approach for still-to-video FR using TM is given in Algorithm 1. It exploits FT to regroup input ROIs of each person, and accumulates their respective scores for robust spatiotemporal recognition. This algorithm implements the system termed as TM-FR in this paper.

Algorithm 1. A track-and-classify approach to still-to-video FR using TM.

Input: Input frames $\{\mathbf{I}_t : t = 1, \ldots, \infty\}$. Gallery-face-models $\{\mathbf{b}^l : l = 1, \ldots, L\}$ of target individuals enlisted in the gallery
Output: List of likely individuals from watch list in the scene

for *each* $\boldsymbol{I}_t$ *with* $t = 1, \ldots, \infty$ *frames* **do**
  Apply *segmentation* to detect ROIs corresponding to faces in a frame
  // Track Initialization
  for *each* $\{ROI_t^r : r = 0, \ldots, R\}$ **do**
    if *the ROI is located in a different place than the existing tracks* **then**
      Increment the number of tracks, $K \leftarrow K + 1$
      Compute a new track-face-model $\mathbf{m}_t^k$ with the ROI for the newly initiated face track K
  // Tracking
  for *each track-face-model* $\boldsymbol{m}_t^k$ *with* $k = 1, \ldots, K$ **do**
    Compute the state $\mathbf{X}_t^k$ of the face in frame $\mathbf{I}_t$ using tracking
    Update the track-face-model $\mathbf{m}_t^k$ using the new state information $\mathbf{X}_t^k$
  // Classification
  for *each each input pattern* $\boldsymbol{a}_t^k$ *associated with* $\boldsymbol{X}_t^k$ *with* $k = 1, \ldots, K$ **do**
    for *each gallery-face-model* $\boldsymbol{b}^l$ *with* $l = 1, \ldots, L$ **do**
      Compute similarity (score) between $\mathbf{a}_t^k$ and $\mathbf{b}^l$ by $S_t(k,l) = similarity(\mathbf{a}_t^k, \mathbf{b}^l)$.
  // Fusion
  for $k = 1$ *to* K **do**
    for $l = 1$ *to* L **do**
      Accumulate scores over W consecutive frames by $acc\_S_t(k,l) = \sum_{i=t-W}^{t} S_i(k,l)$.
      if $acc\_S_t(k,l) \geq \gamma_l$ **then**
        Detect or predict the presence of watch list individual l

2.2. Adaptive face modeling

Adaptive face modeling techniques seek to improve and maintain FR performance under complex capture conditions by updating facial models over time using operational data [7]. These techniques are divided into two categories: *self-update* and *co-update* [7]. *Self-update* techniques apply a second (usually higher) update-threshold to each matching score to select ROI patterns to update the corresponding gallery-face-model [9,14]. The *co-update* techniques seek corroboration of scores from two or more matchers, typically on multiple traits (e.g., face and finger prints) for cross-updating [25,26].

The main limitation of *self-update* method is that it carries the risk of incorrectly updating, and recognition of target individual with the update-threshold selection. Moreover, an overly conservative threshold may allow us to limit false updates, but the system will never adapt to changes in the environment. Conversely, a less conservative update threshold may lead to an increase in the number of false updates and deteriorates system performance. Both the situations adversely affect the performance of the FR system. For example, to update templates with reliable ROI patterns, the ensemble-based system proposed by De la Torrea et al. [14] exploits face trajectories. If accumulated matching scores within the trajectory surpass a predefined threshold, all the ROI patterns related to that trajectory are selected for updating the gallery-face-model. However, threshold selection is still sensitive. *Co-update* methods perform better in capturing large intra-class variations of the input data without relaxing acceptance threshold by using multiple traits. However, collecting information for multiple traits, such as finger print, along with the faces, may not be feasible in WL screening, where capture is performed without user cooperation.

2.3. SSPP techniques for still-to-video face recognition

To improve the limited representativeness of gallery-face-models, several techniques have been proposed in SSPP literature. They are divided into three categories: *multiple face representations*, *synthetic face generation*, and *enlarging the training sets* using auxiliary data [6].

In *multiple face representations*, different discriminant features may be extracted from a reference still image to enhance the gallery-face-models. Each ROI may be extracted using different patches or sub-images to provide multiple representations of a face [27]. Several feature extraction techniques have been proposed in computer vision and in pattern recognition literature for FR, such as LBP, HOG, and PCA, used for FR [27,28]. Key issues for FR with multiple representations is the fusion of representations and system complexity.

In *synthetic generation*, multiple virtual face images are generated from each single reference still to enhance gallery-face-models. Multiple virtual views are synthesized by linear shape prediction [29], mesh warping [30], morphing [31], symmetry property [32], partitioning a face in several sub-images [33], affine transformation [34], noise perturbation [35], shifting [36], and active appearance model [37]. A recurring problem with the synthetic generation is that they need to locate facial components reliably to determine the pose angle for pose compensation. Moreover, these methods need prior knowledge to guide the generation of virtual views, though the quality and realism of the virtual views may not be guaranteed. Synthetic generation may fail to predict many realistic and unobserved variations in face appearance in a real-world scene.

When *enlarging training set using auxiliary data sets*, a generic data set containing multiple ROI patterns from other individuals than target individuals (in the gallery), and possibly under different capture conditions, is exploited to assist in learning the gallery-face-models. Artificial neural network (ANN) [38], Fisher's Linear Discriminant [39], 3D modeling [40], and sparse variation dictionary learning [41] are examples of techniques applied for face modeling using auxiliary data sets. All these methods assume that the intra-personal variation of faces in all humans is similar. The intra-personal variation for a face of an individual can therefore be approximated by using a generic set containing multiple ROIs per person. However, variation of face appearance in a generic set is often quite different from that of the single ROI set. Therefore, the discriminant model learned from the generic set is more suitable to distinguish the persons in the generic set, rather than those in the single ROI pattern set.

2.4. Spatiotemporal fusion

Spatiotemporal FR approaches merge spatial information (e.g., face appearance) with the sequential variations presented over time [1]. These systems exploit tracking to follow the movement of person or faces in a scene, and ideally define a track per different person. A trajectory is defined as a set of ROIs that correspond to a high quality track of same individual appearing in the camera viewpoint. Spatiotemporal recognition often accumulates matching scores for ROIs across trajectories to improve accuracy and robustness.

Several recent methods have been proposed for spatiotemporal face recognition [1,42]. In [43], the spatial and temporal information are merged within a Hidden Markov Model (HMM) by modeling the probability distributions of the motions, and by selecting the highest likelihood score provided by the HMM to decide the identity in the test video sequences. In [44], facial ROIs are divided into several sub-regions, and use an estimation of optical flow to weight the importance of each of the sub-regions when estimating posterior probabilities. This technique considers the motion between each pair of frames, including information from changes of expression. In [45], an appearance-based approach is proposed that estimates the joint posterior distribution of the motion vector and identity variable by combining adaptive observation model and adaptive state transition model within a particle filter-based framework. In [46], a what-and-where fusion neural network is used to classify faces in each frame, where a bank of Kalman filters is used to track the movement of faces in a scene. Finally, an evidence accumulation module accumulates the classifier's responses over time according to face tracks [47].

In [48], a door monitoring system is designed where confidence scores are computed by a local appearance-based FR algorithm. The confidence scores are progressively combined to provide the identity estimate for an entire sequence. Three different measures are used to weight the contribution of individual frames in the identity estimation: distance-to-model, distance-to-second-closest, and their combination. In [49], score driven and quality driven methods are analyzed for spatiotemporal FR. In score driven method, facial regions are continuously matched against facial models until a matching score is above a predefined threshold. In quality driven method, facial images are processed until a quality intrinsic to the considered image is above a predefined threshold. In both of the cases, the matching score over the predefined threshold indicates a positive identification of the sequence.

In [14], a learn-and-combine strategy is employed for spatiotemporal FR in VS. When the number of positive ensemble predictions surpasses a higher update threshold, all target samples from the trajectory are combined with non-target samples (selected from the cohort and universal models) to update the corresponding face model. In addition, a memory management strategy based on Kullback–Leibler divergence is proposed to rank and select the most relevant target and non-target reference samples to be stored in memory as ensembles evolve.

2.5. Challenges

In still-to-video FR, the limited number of reference stills leads to gallery-face-models that are poor representations of target individuals as they would appear in a video camera viewpoint. It is difficult to design robust gallery-face-models that can deal the uncertainties and intra-class variations that may appear in VS environments. During operations, faces are typically captured in low quality videos under semi- and uncontrolled conditions which may be very different from the gallery-face-models. Beyond issues with camera interoperability, these face captures incorporate variations in scale, blur, pose, expression, and lighting [3]. Moreover, fewer face ROIs are typically detected for matching in VS environments using the segmentation module, which also reduces performance. Fig. 2 shows a 2D projection of faces captured by segmentation (face detection) versus those obtained by tracking an individual in a scene. To observe the variation of appearances, the facial captures are extracted into an 81-dimension the case with HOG patterns and with projected in a two dimensional space using Sammon mapping. It is clear from the figure that the face detection (Fig. 2(b)) with face plus eye detection (Fig. 2(a)) detects few facial captures (33 and 7, respectively) out of 67 frames where the individual appears. Using tracker, 60 out of 67 facial captures are observed in Fig. 2(c). As an indication of diversity, the intra-class variations of facial captures in the original HOG space obtained using face detection only, and face and eye detection yield an average standard error of 0.07 and 0.03 with respect to the mean location, respectively, whereas with tracking, it is 0.12.

To incorporate greater diversity of facial representation, and improve FR with limited reference stills, adaptive biometric and SSPP techniques may be used. The limitation of adaptive systems is that they may corrupt the gallery-face-model if updated with incorrectly classified ROI patterns.

Among the SSPP techniques, multiple face representations are complex because they require multiple face descriptors and patch schemes within a common framework [27]. Though these techniques achieve some level of robustness against illumination variation or partial occlusion by using different features, they do not address pose or expression changes during operation. Synthetic face generation and enlarging training set using auxiliary set may not provide representative samples to enlarge gallery-face-models, as they cannot predict many unobserved variations of faces in a real-world scene. The derived transformation function (for the synthetic generation) or the learned discriminant model (by the enlarging training set using auxiliary data set) may not be suitable to distinguish persons from the single reference ROI pattern.

3. Still-to-video FR using adaptive appearance model tracking

A new Adaptive Appearance Model Tracking-based FR (AAMT-FR) system is proposed for still-to-video FR. In the proposed system, a set of gallery-face-models is designed as usual during enrollment, using the reference still images of target individuals. During operations, a track-face-model is learned online over successive frames for each different person appearing in the scene. These models gradually incorporate intra-class information on the facial appearance from the operational scene. Meanwhile, for each frame, these track-face-models are matched against the gallery-face-models of every target individual enrolled to the system. Matching scores linked to target individual are then accumulated over time and compared with an individual-specific decision threshold for robust spatio-temporal recognition.

The proposed system (shown in Fig. 3) is comprised of the same general modules as in Fig. 1, but FR is performed using an AAMT. Algorithm 2 presents an algorithmic description of a track-and-classify system for still-to-video FR using AAMT. The main advantage of AAMT-FR is that a greater diversity of facial representation may be captured from face in the scene during operations with AAMT, and this can lead to a higher level of discrimination for spatiotemporal recognition. The rest of this section presents additional details about the track-face-modeling and spatiotemporal recognition of this system, making the same simplifying assumption as in Section 2 (regarding the gallery-face-model and template matching system).

Algorithm 2. A track-and-classify approach to still-to-video FR using AAMT.

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

Output: List of likely individuals from WL in the scene

for *each* $\boldsymbol{I}_t$ *with* $t = 1, ..., \infty$ **do**
- Apply *segmentation* to detect ROIs corresponding to faces in a frame.
- // Track Initialization
- **for** *each* $\{ROI_t^r : r = 0, ..., R\}$ **do**
  - **if** *the ROI is located in a different place than the existing tracks* **then**
    - Increment the number of tracks, $K \leftarrow K + 1$
    - Initiate a new track-face-model $\mathbf{m}_t^k$ with the ROI for the newly initiated face track K
- // Tracking
- **for** *each track-face-model* $\boldsymbol{m}_t^k$ *with* $k = 1, ..., K$ **do**
  - Compute the state $\mathbf{X}_t^k$ of the face in frame $\mathbf{I}_t$ using tracking
  - Given a data block $\mathbf{A}_k$, compute a face-track-model $\mathbf{m}_t^k$ using Singular Value Decomposition
  - Given a new data block $\mathbf{B}_k$, update the track-face-model $\mathbf{m}_t^k$ using the newly computed state information $\mathbf{X}_t^k$ by SKL algorithm (*seeAlgorithm*3)
- // Classification
- **for** *each track-face-model* $\boldsymbol{m}_t^k$ *with* $k = 1, ..., K$ **do**
  - **for** *each template* $\boldsymbol{b}^l$ *with* $l = 1, ..., L$ **do**
    - Compute similarity score, $S(k, l)$ between $\mathbf{m}_t^k$ and $\mathbf{b}^l$ using *Equation* 2
- // Fusion
- **for** $k = 1$ *to* K **do**
  - **for** $l = 1$ *to* L **do**
    - Accumulate scores over W consecutive frames by $acc\_S_t(k, l) = \sum_{i=t-W}^{t} S_i(k, l)$.
  - **if** $acc\_S_t(k, l) \geq \gamma_l$ **then**
    - Detect or predict the presence of WL individual l

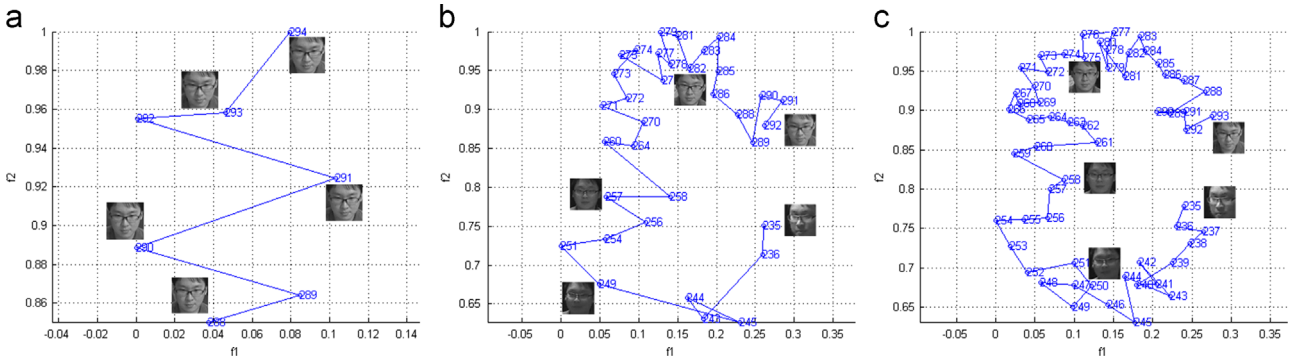

Fig. 2. Sammon mapping of facial captures obtained using (a) combined face and eye detection FD and ED (7 captures), (b) face detection FD (33 captures), and (c) face tracking from video of individual ID 03 in *P1E_S1_C1* of Chokepoint dataset FT (60 captures).

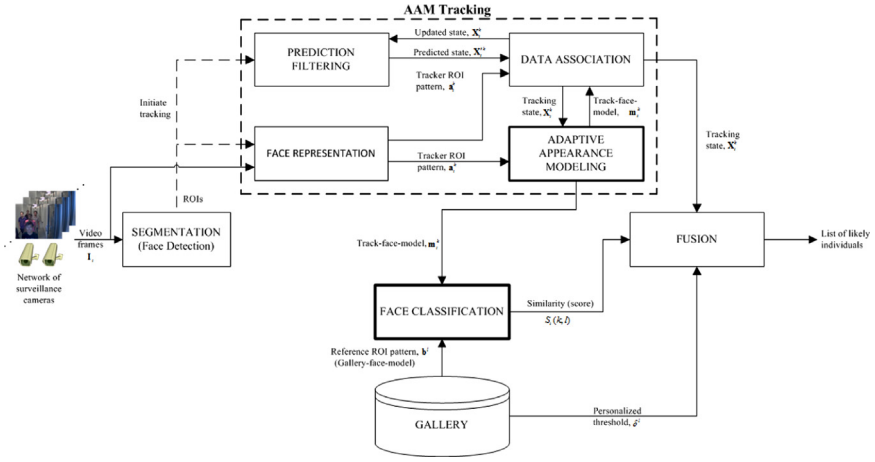

Fig. 3. Framework for the proposed AAMT-FR.

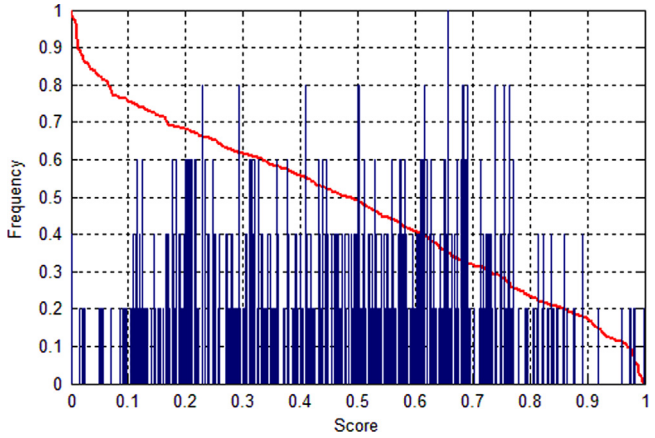

Fig. 4. Example of the cumulative probability distribution function for δ^3 (target ID 03) versus ROI patterns of non-target individuals over all non-target videos (UBM). In this example, individual IDs 18–27 are captured in all the non-target videos. (For interpretation of the references to color in the text, the reader is referred to the web version of this paper.)

3.1. Track face modeling

The face tracking module incorporates four main functions – *face representation*, *prediction filtering*, *data association*, and *adaptive appearance modeling* of faces in tracking. The segmentation module captures face ROIs in each frame $\mathbf{I}_t$, ROI_t^r, where $r = 0, \ldots, R$. Given a ROI_t^r captured in a new region of an input frame $\mathbf{I}_t$ by the segmentation module, the features are extracted into a ROI pattern $\mathbf{a}_t^k$. It allows initiating a track face model $\mathbf{m}_t^k$ for a new track. For existing tracks, the ROI pattern is extracted from the candidate region in $\mathbf{I}_t$ itself through data association. In Fig. 2, $\mathbf{a}_t^k$ is a ROI pattern representing the face captured for track k at a candidate region of frame $\mathbf{I}_t$.

The *adaptive appearance modeling* module inside the tracker generates track-face-models $\mathbf{m}_t^k$ for the newly initiated tracks, and updates the models for the existing tracks. Once a new face track k is initially detected in the scene, the ROI pattern for the first n frames are

Fig. 5. Illustration of reference ROI patterns extracted from high quality mug shots represented in the gallery and a selection of ROI patterns from videos of the target individuals in the scene for replication one.

Table 1
Number of ROIs captured per camera for each individual in Chokepoint videos.

ID	Entering			Leaving			Total
	Camera 1	Camera 2	Camera 3	Camera 1	Camera 2	Camera 3	
01	425	434	291	481	499	334	2464
03	421	448	313	456	561	360	2559
04	314	337	215	362	359	199	1786
05	359	411	281	466	483	307	2307
06	394	459	303	523	525	312	2516
07	488	487	324	448	501	337	2585
09	330	358	269	451	454	311	2173
10	366	394	289	454	478	289	2270
11	465	484	314	521	510	378	2672
12	280	339	258	468	447	269	2061
13	490	534	371	634	648	431	3108
14	406	431	280	484	488	324	2413
15	383	417	296	459	476	270	2301
16	269	320	222	400	396	253	1860
17	369	380	261	447	477	322	2256
18	354	379	276	513	508	341	2371
19	320	418	283	431	517	340	2309
20	443	441	312	491	497	329	2513
21	404	391	265	469	466	303	2298
22	368	371	256	425	441	276	2137
23	359	343	226	400	408	250	1986
24	447	485	326	554	582	391	2785
25	484	486	355	559	562	370	2816
26	307	320	234	359	371	243	1834
27	351	354	252	467	436	308	2168
Total	9596	10,221	7072	11,722	12,090	7846	58,548

tracked and captured. A data block $\mathbf{A}^k = \{\mathbf{a}_1^k, \ldots, \mathbf{a}_n^k\}$ is thereby defined using the tracked face regions with states $\{\mathbf{X}_1^k, \ldots, \mathbf{X}_n^k\}$. Then, the track-face-model of the target face is generated with three components, $\mathbf{m}_A^k = \{\mathbf{U}_A^k, \overline{\mathbf{a}^k}_A, \Sigma_A^k\}$, where $\mathbf{U}_A^k$ is the eigen vector, $\overline{\mathbf{a}}_A^k$ is the mean vector, and Σ_A^k is the covariance matrix computed from the singular value decomposition (SVD) of the centered data matrix of data block $\mathbf{A}^k$.

When a new data block, $\mathbf{B}^k = \{\mathbf{a}^k_{n+1}, \ldots, \mathbf{a}^k_{n+q}\}$, becomes available after tracking for q additional frames, the updated face model $\mathbf{m}^k_{A+B} = \{\mathbf{U}_{A+B}, \overline{\mathbf{a}}_{A+B}, \Sigma_{A+B}\}$ is obtained by using the augmented data matrix [**AB**] through the computationally efficient Sequential Karhunen–Loeve (SKL) algorithm [13]. The input to *adaptive appearance modeling* module is the ROI pattern $\mathbf{a}^k_t$ extracted from the region defined by the tracking state $\mathbf{X}^k_t$ for face k at time t, and the output is the updated face model $\mathbf{m}^k_t$. Two key parameters – the forgetting factor f, and batch size q – determine the plasticity of the track face models. The parameter $f \in [0, 1]$ determines the contribution from older observations to be considered in updating the track-face-model, where $f = 1$ indicates no forgetting is to occur at all. The parameter q defines the batch size at which the face model is updated during tracking. An algorithmic description of the SKL method is given in Algorithm 3.

Algorithm 3. Update of the track-face-model using Sequential Karhunen–Loeve (SKL) method.

Input: Initial track-face-model $\mathbf{m}^k_A = \{\mathbf{U}_A, \overline{\mathbf{a}}_A, \Sigma_A\}$ obtained using data block $\mathbf{A}^k = \{\mathbf{a}^k_1, \ldots, \mathbf{a}^k_n\}$; new data block obtained after tracking $\mathbf{B}^k = \{\mathbf{a}^k_{n+1}, \ldots, \mathbf{a}^k_{n+q}\}$; Batch size ($q$); forgetting factor ($f$).

Output: Updated track-face-model, $\mathbf{m}^k_{A+B} = \{\mathbf{U}_{A+B}, \overline{\mathbf{a}}_{A+B}, \Sigma_{A+B}\}$.

1 Compute mean vectors $\overline{\mathbf{a}}_B = 1/q \sum_{i=n+1}^{n+q} \mathbf{a}_i$, $\overline{\mathbf{a}}_{A+B} = \frac{f.n}{f.n+q} + \frac{q}{f.n+q}\overline{\mathbf{a}}_B$

2 Form the matrix $\hat{\mathbf{B}} = \left[(\mathbf{a}_{n+1} - \overline{\mathbf{a}}_B) \ldots (\mathbf{a}_{n+q} - \overline{\mathbf{a}}_B) \sqrt{\{(n.q)/(n+q)\}} (\overline{\mathbf{a}}_\mathbf{B} - \overline{\mathbf{a}}_A)\right]$

3 Compute $\tilde{\mathbf{B}} = \mathbf{orth}(\hat{\mathbf{B}} - UU^T\hat{\mathbf{B}})$ and $\mathbf{R} = \begin{bmatrix} f\Sigma & \mathbf{U}^T\hat{\mathbf{B}} \\ 0 & \tilde{\mathbf{B}}(\hat{\mathbf{B}} - \mathbf{U}\mathbf{U}^T\hat{\mathbf{B}}) \end{bmatrix}$

4 Compute the SVD of $\mathbf{R}$: $\mathbf{R} \overset{SVD}{\rightarrow} \tilde{\mathbf{U}}\tilde{\Sigma}\tilde{\mathbf{V}}$

5 Update eigen vector, $\mathbf{U}_{A+B} = \left[\mathbf{U}\tilde{\mathbf{B}}\right]\tilde{\mathbf{U}}$ and $\Sigma_{A+B} = \tilde{\Sigma}$

The SKL algorithm has space and time complexity that is constant within n, the number of tracker ROI patterns capture so far. Specifically, each update makes use of only the k largest singular values and basis vectors from the previous model update stage. Along with the storage required for the q new ROI patterns, each one extracted from a d dimensional facial captures, this reduces the space complexity to $O(d(k+q))$, down from $O(d(k+q)^2)$ with the traditional PCA approach, where d is the dimension of the facial captures. Similarly, the time complexity is also reduced to $O(dq^2)$, versus $O(d(n+q)^2)$ for recomputing the entire SVD. More details and complexity analysis of the SKL algorithm are described in [13].

In *Prediction filtering*, the state of a face in a frame is predicted based on information in the previous frames, and on some underlying model for state transitions. The objective of prediction filtering is to achieve improved data association while reducing search space. Given a ROI pattern a^k_t at frame $\mathbf{I}_t$, the input to the prediction filter is the previous state $\mathbf{X}^k_{t-1}$ and the output is a number of predicted states $'\mathbf{X}^k_t$ defining the new locations and sizes of the face at $\mathbf{I}_t$. A particle filter is used [50] for predicting the new states during tracking.

Data association compares the face model $\mathbf{m}^k_t$ and the tracker ROI pattern $\mathbf{a}^k_t$ 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 $\mathbf{I}_t$. Inputs to the data association module are $\mathbf{m}^k_t$ and a tracker ROI pattern $\mathbf{a}^k_t$ extracted from the region defined by a predicted state $'\mathbf{X}^k_t$, and the output is the estimated state $\mathbf{X}^k_t$ of the face in $\mathbf{I}_t$. The state vector $\mathbf{X}^k_t$ is used to predict the location of the target face k in next frame $\mathbf{I}_{t+1}$ for the further

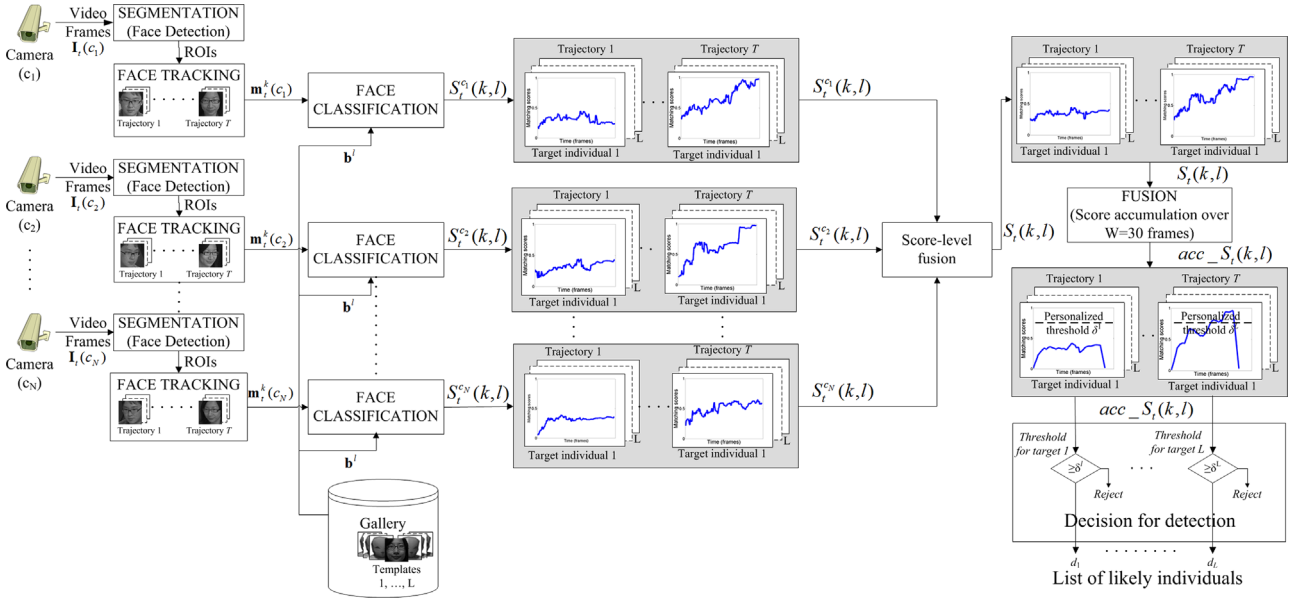

Fig. 6. A block diagram of AAMT-FR method using score-level fusion over the 3 Chokepoint cameras.

Table 2
List of Experiments

Experiments	Description	Performance Measures
Transaction level analysis (Section 5.1)	Face matching performance without camera fusion Face matching performance with camera fusion	ROC, precision, recall, AUC, AUPR
Trajectory level analysis (Section 5.2)	Overall performance based on trajectories without camera fusion Overall performance based on trajectories with camera fusion	ROC, AUC
Performance analysis at different priors (Section 5.3)	Overall performance at different priors of target and non-target	AUPR
Time complexity analysis (Section 5.4)	Time needed to process one frame	sec/frame
Detailed analysis(Section 5.5)	Best and worst case scenarios Sensitivity of different parameters	Accumulated scores and scatter plots of the facial captures AUC varying forgetting factors and batch-sizes

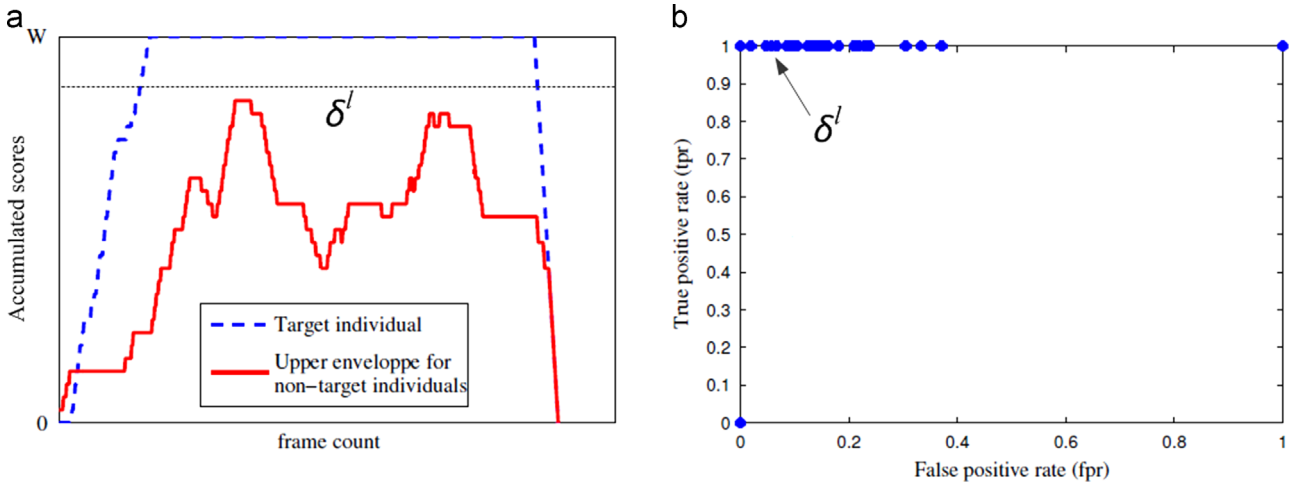

Fig. 7. Illustration of trajectory-level analysis used to evaluate the overall performance of a system for still-to-video FR over one video [14]. (a) Score. (b) ROC space.

tracking. Similarity between the face model $\mathbf{m}_t^k$ and the tracker ROI pattern $\mathbf{a}_t^k$ to find face correspondences in consecutive frames is measured using

$$P(\mathbf{a}_t \mid \mathbf{X}_t^k) = P_{d_t}(\mathbf{a}_t \mid \mathbf{X}_t^k) P_{d_w}(\mathbf{a}_t \mid \mathbf{X}_t^k) \tag{1}$$

where $P_{d_t} = N(\mathbf{a}_t; \boldsymbol{\mu}, \mathbf{U}\mathbf{U}^T + \epsilon\mathbf{I})$ refers to the probability of a ROI pattern generated from a subspace which is governed by a Gaussian distribution and $P_{d_w} = N(\mathbf{a}_t; \boldsymbol{\mu}, \mathbf{U}\Sigma^{-2}\mathbf{U}^T)$ refers to the probability of the projected ROI pattern which is modeled by the distance from the mean [21].

3.2. Spatiotemporal recognition

The fusion module incorporates information from both AAMT and classification for spatiotemporal recognition. Assume a gallery that contains gallery-face-model $\{\mathbf{b}^l : l = 1, \ldots, L\}$ of the reference stills, and individual-specific decision thresholds $\{\delta^l : l = 1, \ldots, L\}$. During operations, the similarity between a face-track-model $\mathbf{m}_t^k = \{\mathbf{U}, \overline{\mathbf{a}}_t, \Sigma\}$ and a gallery-face-model $\mathbf{b}^l$ is computed. The input to the classification step are $\mathbf{m}_t^k$ and $\mathbf{b}^l$, and the output is the similarity score:

$$S_t(k, l) = \exp\left\{ -\left\| (\mathbf{b}^l - \overline{\mathbf{a}}_t) - \mathbf{U}\mathbf{U}^T(\mathbf{b}^l - \overline{\mathbf{a}}_t) \right\|^2 \right\} \tag{2}$$

At a time instance t, $S_t(k, l)$ indicates the likelihood that the facial model $\mathbf{m}_t^k$ corresponds to the template for individual l enrolled to the system.

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

$$acc\_S_t(k, l) = \sum_{i=t-W}^{t} S_i(k, l) \tag{3}$$

If the accumulated score for a target individual surpasses its decision threshold δ^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 thresholds δ^l may be selected using the score distribution obtained by matching the gallery-face-model $\mathbf{b}^l$ to ROI patterns of non-target individuals from the UBM at a desired *fpr* of the

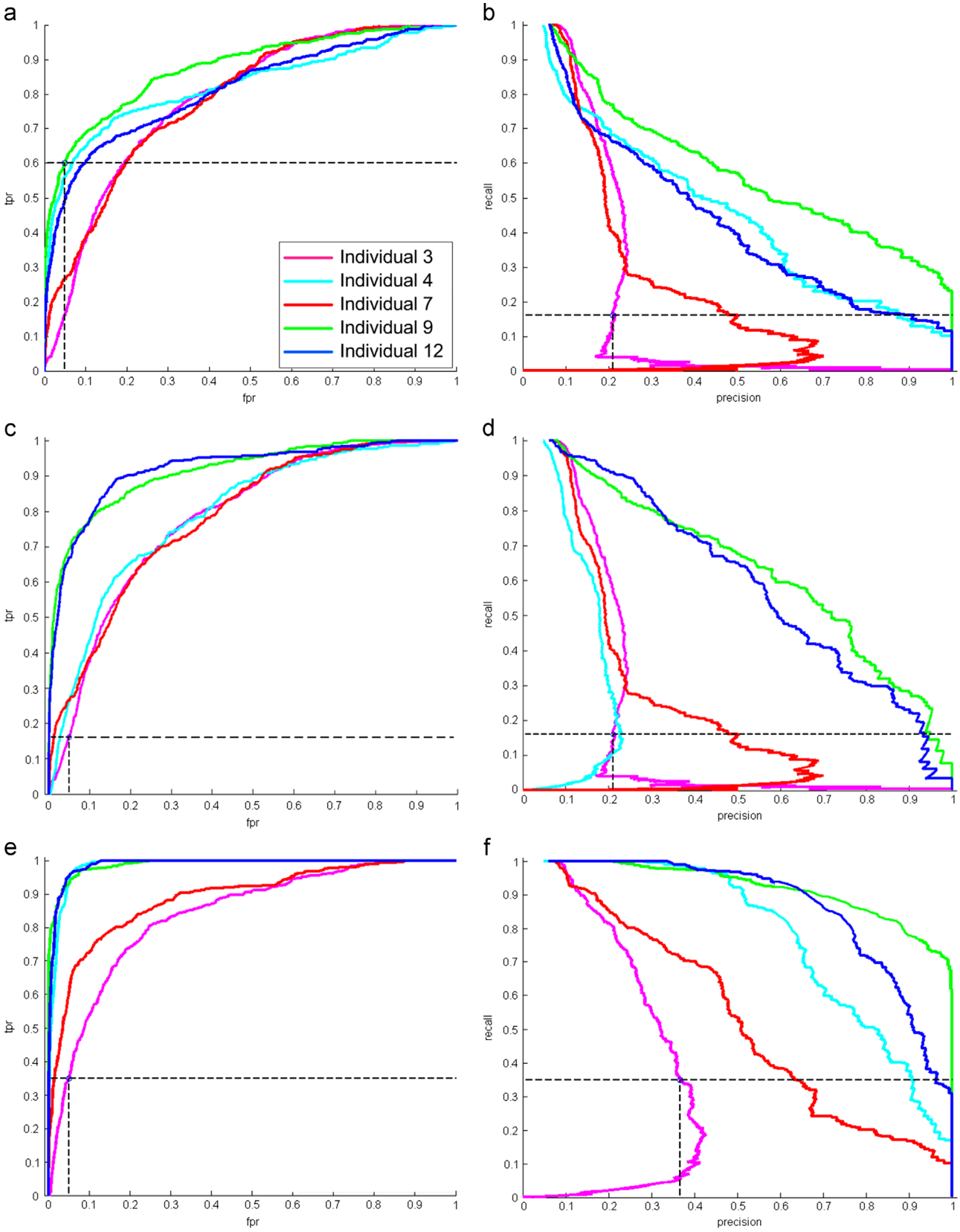

Fig. 8. Example of ROC curves (right side) and inverted-PR curves (left side) for TM-FR, TM-FR with self-update, and AAMT-FR systems using all the leaving video sequences of Chokepoint for Camera 2. (a), (b) Leaving, Camera 2 sequences (TM-FR). (c), (d) Leaving, Camera 2 sequences (TM-FR with self-update). (e), (f) Leaving, Camera 2 sequences (AAMT-FR).

cumulative probability density function. In Fig. 4, the score distribution is shown in blue, whereas its cumulative distribution is shown in red. At $fpr = 5\%$, the decision threshold for this specific individual (ID 03) in the illustration is $\delta^3 \cong 0.8251$.

4. Experimental methodology

4.1. Video data

To compare the performance of the proposed method, video sequences from the Chokepoint dataset [15] are used. They are recorded for a VS scenario, where an array of 3 cameras is placed above different portals (natural choke points in terms of pedestrian traffic) to capture individuals walking through in a natural way. The sequences are named according to the recording conditions (e.g. P2E_S1_C3) where P, S, and C stand for portal, sequence and camera, respectively. E and L indicate individuals either entering or leaving the portal.

The dataset contains 54 videos, among which 6 contain a mixture of people (crowded scene) and 48 are with one person at a time. Each one of the videos with one person at a time captures 29 individuals, where 19 subjects are male and 6 are female. All videos are captured in two portals and 5 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. In total, the dataset consists of 54 video sequences (27 leaving and 27 entering) and 64,204 labeled face images each of which are cropped with size 96×96. This dataset is challenging for FR as the videos are captured under uncontrolled conditions with variations in pose accumulation, lighting, scale, and blur.

4.2. Protocol

The AAMT-FR technique proposed in this paper is compared to the systems for still-to-video FR based on template matching (TM-FR [8]), adaptive biometrics (TM-FR with self-update [9]), Sparse Variation Dictionary Learning (SVDL [41]), and Multiple Face Representation (MFR [27]). In TM-FR, input ROI patterns are extracted from the ROIs detected in a frame and compared with all the gallery-face-models 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 gallery-face-model for each target individual. In TM-FR with self-update, a FR system similar to TM-FR is employed, where the gallery-face-models is changed adaptively over time. To update the gallery-face-models, only those input ROI patterns are selected for which the similarity scores surpass a second update-threshold for the target individuals. In SVDL, the gallery-face-models 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 gallery-face-model. Multiple feature extraction techniques (LBP, LPQ, HOG, and Haar feature) are applied to patches isolated from the gallery-face-models to generate diverse face-part representations. Finally, an ensemble of template matchers is applied on multiple face representations for FR.

To measure average system performance, the following experiments are repeated 10 times, and each time randomly selecting 5 different individuals as targets and 10 different individuals as non-targets for WL screening. The initial enrollment process involves the random selection of 5 target WL individuals of interests among the 25 individuals of Chokepoint dataset. As an example, the ROIs captured in 5 mugshots of 5 individuals are shown in Fig. 5 (left column). During operation, 15 individuals are considered for recognition in video sequences, which includes the 5 target individuals and 10 non-target individuals selected at random to appear in the scene. Selected ROI pattern captured from videos for the target WL individuals are also shown in Fig. 5 (right column). Since the experiments are conducted on 25 individuals, and we consider 48 video sequences in the Chokepoint dataset, the performance of the FR systems are measured on $48 \times 25 = 1200$ facial trajectories, with a total number of about 58,000 ROIs. The remaining 10 individuals $(25 - 15)$ are used as universal background model (UBM) for individual-specific threshold selection and to compute PCA projection matrix.

In experiments, faces are detected using Viola and Jones [16] algorithm. The number of ROIs captured for the experiments from the entering and leaving sequences with different camera views are enlisted in Table 1. A particle filter based tracker [21] is used to follow the motion of faces in the scene, and form trajectories. Trajectories are initialized with the bounding box surrounding a new face detected in the scene. Different parameters for the tracker are set as follows. For predicting the new location of the target in a frame, $P = 600$ particles are used with the particle filter. For affine transformation, the parameters σ_x, σ_y, σ_θ, σ_s, σ_α, and σ_φ are set to 9, 9, 0.05, 0.05, 0.005, and 0.001, respectively. These parameters represent the standard deviations for x-translation, y-translation, rotation, scale, aspect-ratio, and skew direction changes, respectively, that are allowed for transformation functions with the tracking system. The forgetting factor f and batch size g for the SKL method to update track-face-model online are set to 0.99 and 5.00, respectively. In SVDL, three regularization parameters, λ_1, λ_2, λ_3, are set 0:001, 0:01 and 0:0001, respectively, and the number of dictionary atoms as 400 in the initialization. In MFR, the patch based method is used for feature extraction, where 16 patches with size 12×12 have been extracted from each of the facial captures in the experiments.

The facial ROIs are scaled into a common size of 48×48 pixels. For recognition, 81-dimensional HOG (Histogram of Oriented Gradient) features are extracted from each ROI. Then HOG features are extracted by using 9 rectangular cells and 9 bin histogram per cell. The nine histograms with nine bins are then concatenated to make a 81-dimensional feature vectors. The histogram bins are calculated over rectangular cells by the computation of unsigned gradient. The cells overlap half of their area, meaning that each cell contributes more than once to the final feature vectors. Finally, the 81 features are reduced into 32 using PCA projection.

At a given time, one videos in the Chokepoint dataset are recorded using 3 cameras with different viewing angles, one of the cameras is likely to capture a face near-frontal. Since faces captured in reference stills have a frontal pose and higher resolution, ROI captured from video with a near frontal pose and higher quality are expected to generate higher similarity scores. As seen in Table 1, Camera 2 is more likely to capture a face with near-frontal view of the 48 Chokepoint sequences used for experiments, only one individual is presented at a time. Therefore, by selecting the maximum score among the responses obtained from 3 camera views, it can be expected that the most similar frontal pose will be selected to populate a trajectory. All the transaction and trajectory level results are always considered with or without score fusion across cameras.

An architecture for the AAMT-FR technique with camera-fusion is shown in Fig. 6. Given $T = 15$ individuals walking through a portal one at a time, surveillance cameras will provide different view points. For a single camera view C_n (here, $n = 1$, 2, and 3), 15 track-face-models are

Table 3
Average AUC, pAUC(5 %) and AUPR performance for TM-FR, TM-FR with self-update, and AAMT-FR systems at transaction level analysis on all the videos.

IDs	Entering/Leaving	TM-FR			TM-FR with self-update			AAMT-FR		
		AUC	pAUC	AUPR	AUC	pAUC	AUPR	AUC	pAUC	AUPR
03	Entering	0.739 ±0.06	0.160 ±0.05	0.237 ±0.07	0.739 ±0.06	0.160 ±0.05	0.237 ±0.07	0.823 ±0.04	0.123 ±0.06	0.244 ±0.07
	Leaving	0.818 ±0.02	0.118 ±0.02	0.233 ±0.03	0.818 ±0.02	0.118 ±0.02	0.233 ±0.03	0.805 ±0.03	0.168 ±0.01	0.256 ±0.02
04	Entering	0.771 ±0.02	0.339 ±0.01	0.373 ±0.02	0.771 ±0.02	0.339 ±0.01	0.373 ±0.03	0.936 ±0.01	0.484 ±0.07	0.521 ±0.08
	Leaving	0.824 ±0.02	0.324 ±0.07	0.347 ±0.07	0.815 ±0.02	0.226 ±0.06	0.247 ±0.06	0.972 ±0.01	0.642 ±0.04	0.685 ±0.05
07	Entering	0.753 ±0.02	0.131 ±0.01	0.215 ±0.02	0.753 ±0.02	0.131 ±0.01	0.215 ±0.02	0.812 ±0.03	0.234 ±0.06	0.347 ±0.07
	Leaving	0.799 ±0.02	0.174 ±0.05	0.238 ±0.03	0.799 ±0.02	0.174 ±0.05	0.238 ±0.05	0.912 ±0.01	0.403 ±0.04	0.539 ±0.03
09	Entering	0.673 ±0.02	0.119 ±0.02	0.134 ±0.02	0.673 ±0.04	0.119 ±0.02	0.134 ±0.02	0.933 ±0.02	0.374 ±0.10	0.445 ±0.08
	Leaving	0.833 ±0.15	0.381 ±0.09	0.473 ±0.12	0.884 ±0.16	0.463 ±0.12	0.544 ±0.14	0.976 ±0.19	0.728 ±0.15	0.804 ±0.14
12	Entering	0.726 ±0.03	0.236 ±0.06	0.286 ±0.07	0.813 ±0.07	0.304 ±0.08	0.376 ±0.09	0.954 ±0.01	0.532 ±0.07	0.619 ±0.06
	Leaving	0.805 ±0.01	0.276 ±0.05	0.360 ±0.05	0.878 ±0.04	0.387 ±0.09	0.473 ±0.10	0.968 ±0.01	0.721 ±0.05	0.780 ±0.04
Average		0.774 ±0.04	0.226 ±0.04	0.290 ±0.05	0.794 ±0.05	0.242 ±0.05	0.307 ±0.05	0.908 ±0.04	0.441 ±0.07	0.524 ±0.07

Table 4
Average *P*, *R*, and *F*1 performance for TM-FR, TM-FR with self-update, and AAMT-FR systems on all videos.

IDs	Entering/ Leaving	TM-FR			TM-FR with self-update			AAMT-FR		
		P	R	F1	P	R	F1	P	R	F1
03	Entering	0.277 ±0.06	0.256 ±0.08	0.265 ±0.07	0.277 ±0.05	0.256 ±0.07	0.265 ±0.06	0.248 ±0.09	0.235 ±0.11	0.240 ±0.10
	Leaving	0.249 ±0.02	0.225 ±0.03	0.236 ±0.03	0.249 ±0.02	0.224 ±0.03	0.236 ±0.03	0.292 ±0.04	0.277 ±0.04	0.285 ±0.04
04	Entering	0.331 ±0.01	0.441 ±0.01	0.368 ±0.01	0.331 ±0.01	0.441 ±0.01	0.378 ±0.01	0.430 ±0.15	0.675 ±0.04	0.525 ±0.02
	Leaving	0.310 ±0.05	0.458 ±0.08	0.370 ±0.06	0.259 ±0.04	0.359 ±0.08	0.301 ±0.06	0.459 ±0.04	0.854 ±0.04	0.597 ±0.02
07	Entering	0.257 ±0.02	0.200 ±0.01	0.225 ±0.01	0.257 ±0.02	0.200 ±0.01	0.225 ±0.01	0.363 ±0.03	0.335 ±0.05	0.347 ±0.04
	Leaving	0.275 ±0.05	0.276 ±0.08	0.273 ±0.07	0.275 ±0.05	0.276 ±0.08	0.273 ±0.07	0.463 ±0.02	0.591 ±0.03	0.519 ±0.03
09	Entering	0.201 ±0.03	0.205 ±0.04	0.203 ±0.03	0.201 ±0.03	0.205 ±0.04	0.203 ±0.03	0.409 ±0.04	0.570 ±0.12	0.474 ±0.07
	Leaving	0.400 ±0.08	0.491 ±0.11	0.440 ±0.09	0.435 ±0.09	0.576 ±0.14	0.494 ±0.10	0.456 ±0.10	0.882 ±0.17	0.674 ±0.12
12	Entering	0.249 ±0.04	0.300 ±0.07	0.271 ±0.06	0.301 ±0.06	0.404 ±0.11	0.342 ±0.08	0.440 ±0.02	0.694 ±0.07	0.537 ±0.03
	Leaving	0.339 ±0.03	0.395 ±0.06	0.364 ±0.05	0.398 ±0.06	0.523 ±0.11	0.451 ±0.08	0.533 ±0.02	0.872 ±0.05	0.662 ±0.03
Average		0.289 ±0.03	0.325 ±0.06	0.302 ±0.05	0.398 ±0.05	0.346 ±0.07	0.317 ±0.05	0.418 ±0.04	0.599 ±0.07	0.486 ±0.05

initiated and generate T trajectories by tracking their faces in the scene. Classification compares each track-face-model against the $L=5$ templates $\mathbf{b}_l$ in the gallery. For each frame, every track-face-model generates a matching score for the L templates in the gallery. The scores for different camera views are combined using the maximum score-level fusion rule [51]. Finally, the maximum scores are accumulated over the trajectory for robust spatio-temporal recognition. If an accumulated score in a fixed window of $W=30$ frames surpasses the individual-specific decision threshold, the presence of the target individual is detected in the scene.

4.3. Performance measures

Performance of a still-to-video FR system is measured at transaction- and trajectory-levels, each one evaluated with only the center camera or with score-level fusion over the 3 cameras. Transaction-level analysis shows the performance of FR systems for face matching on each ROI. Table 2 gives an overview of the experiments presented in Section 5.

In Receiver Operational Characteristic (ROC) space, the Area Under Curve (AUC) provides a global measure of system performance. ROC is a parametric curve in which the true positive rate *tpr* is plotted against *fpr*, where the *tpr* is the proportion of targets correctly classified over the total number of target ROIs and the *fpr* is the proportion of non-target ROIs incorrectly classified (as positives) over the total number of non-target ROIs. In practice, an empirical ROC curve is obtained by connecting the observed (*tpr*, *fpr*) pairs for a classifier at each threshold. The AUC assesses ranking in terms of class separation – the fraction of positive–negative pairs that are ranked correctly. For instance, with an AUC=1, all positives are ranked higher than negatives indicating a perfect discrimination between classes. A random classifier has an AUC=0.5, and both classes are ranked at random. The partial AUC, pAUC(5%), is measured by taking the AUC at $0 < fpr \leq 5\%$ from the ROC curve.

Class priors for target and non-target individuals may vary over time in real scenario. Traditional ROC analysis cannot distinguish between two classifiers for specific class miss-classification costs. ROC curves and the AUC allow for a performance evaluation that is independent of costs and priors by integrating performance over a range of decision thresholds. However, it is important to observe performance as the proportion of the correctly predicted positive ROIs out of the total number of ROIs predicted to belong to a given individual. Otherwise, when processing highly

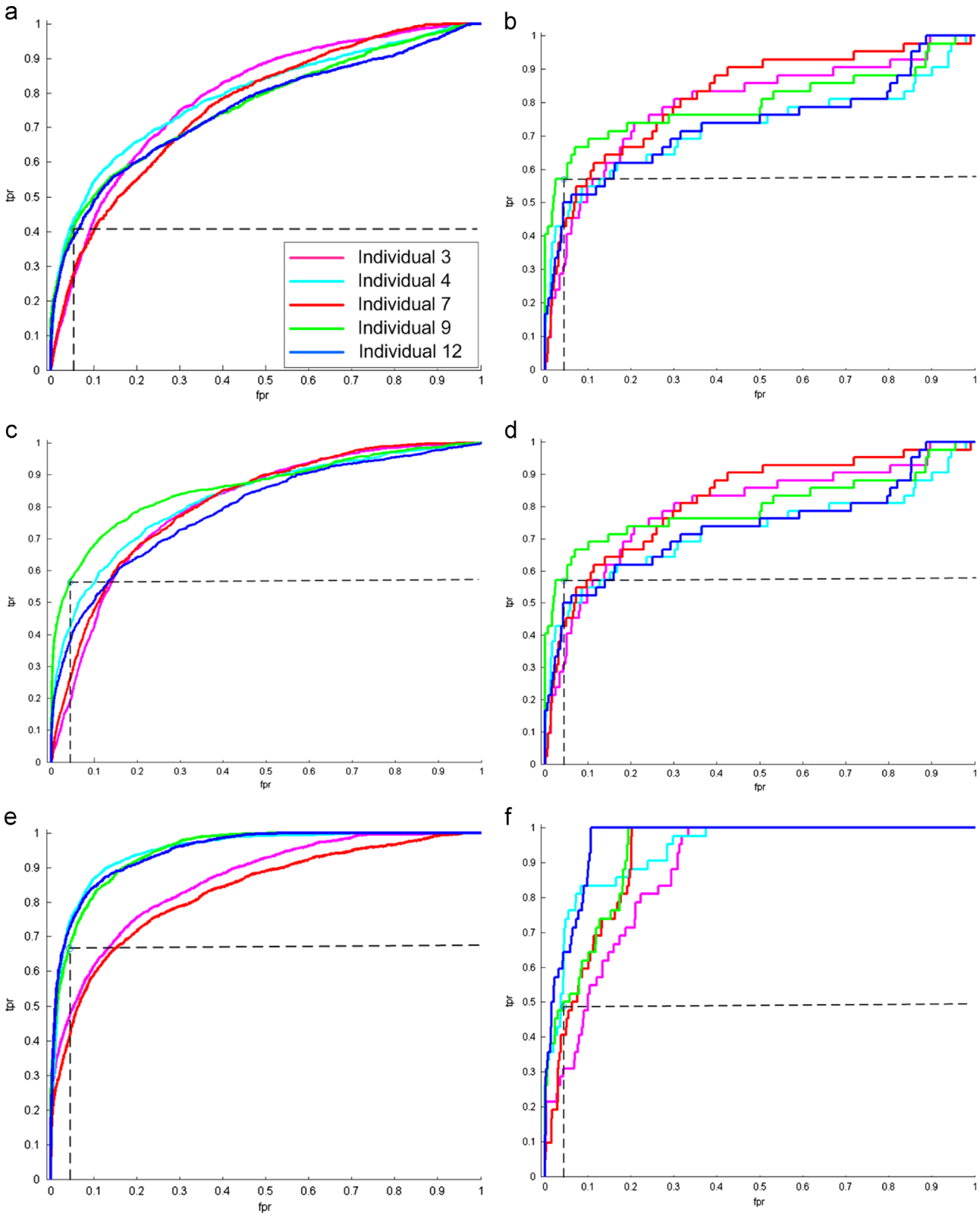

Fig. 9. ROC curves for TM-FR, TM-FR with self-update, and AAMT-FR systems at transaction and trajectory levels for all videos Captured by Camera 2. (a) ROCs at transaction level (TM-FR). (b) ROCs at trajectory level (TM-FR). (c) ROCs at transaction level (TM-FR with self-update). (d) ROCs at trajectory level (TM-FR with self-update). (e) ROCs at transaction level (AAMT-FR). (f) ROCs at trajectory level (AAMT-FR).

imbalanced data, and the minority positive samples are of interest, a FR system may outperform others by predicting a very large number of samples as minority, resulting in an increased *tpr* at the expense of an increased *fpr*. Given the imbalance between target and non-target captures, performance is assessed in the Precision–Recall (PR) space [52], where the area under the PR curve (AUPR) provides another global scalar measure. The PR space focuses on inter-class measures. Precision $(P) = t_p/(t_p + f_p)$ is the proportion of correct positive predictions (TP) against the total positive

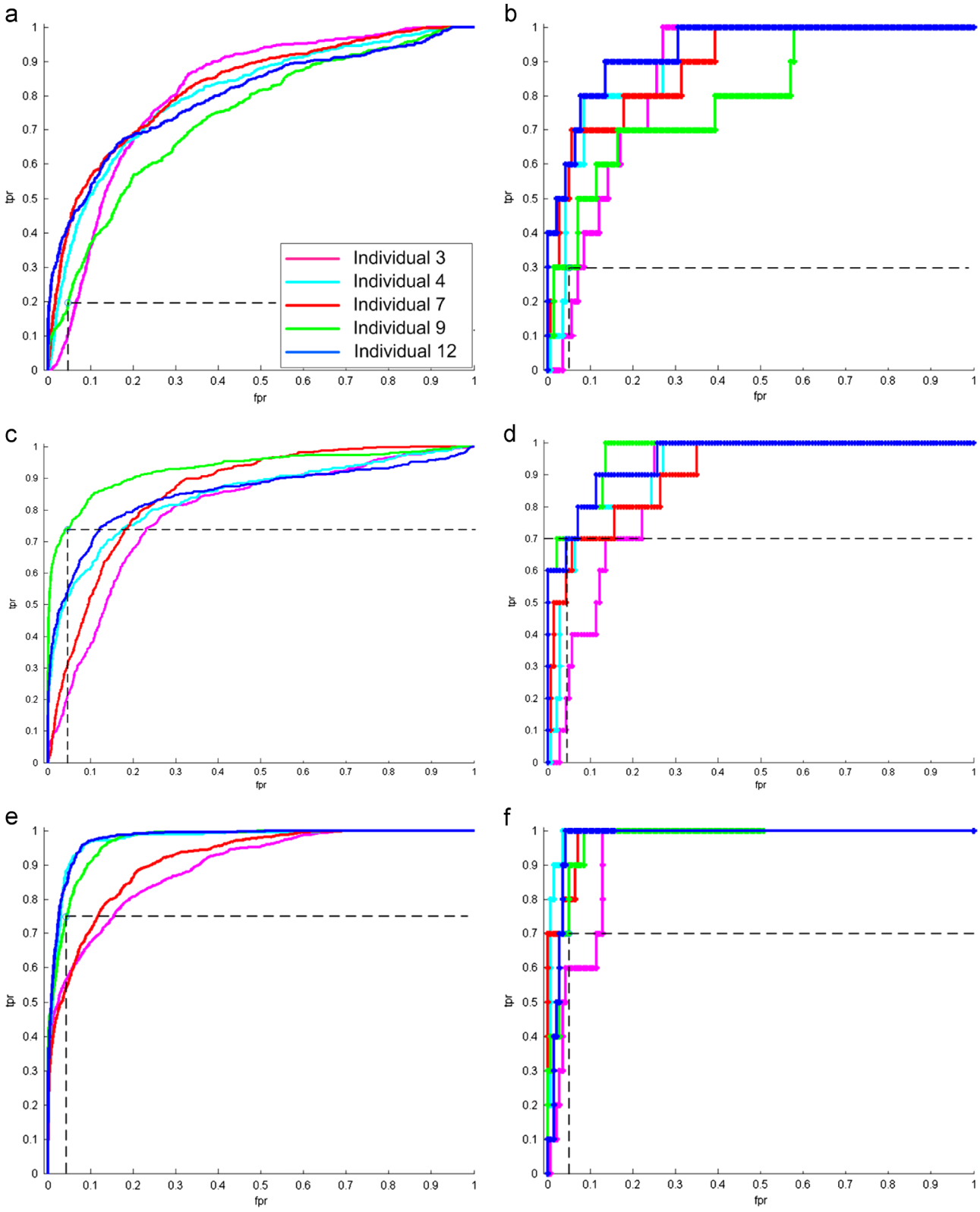

Fig. 10. ROC curves for TM-FR, TM-FR with self-update, and AAMT-FR systems at transaction and trajectory levels with camera fusion of 3 Chokepoint Cameras. (a) ROCs at transaction level (TM-FR). (b) ROCs at trajectory level (TM-FR). (c) ROCs at transaction level (TM-FR with self-update). (d) ROCs at trajectory level (TM-FR with self-update). (e) ROCs at transaction level (AAMT-FR). (f) ROCs at trajectory level (AAMT-FR).

predictions (t_p+f_p) and Recall $(R)=t_p/(t_p+f_n)$ which is same as the *tpr*. At a particular operating point, the scalar F_1 produces a single performance indicator, $F1=2P\cdot R/(P+R)$.

With trajectory-level analysis, the performance of an entire still-to-video FR system is analyzed after accumulating its output scores (obtained from score-level fusion for 3 camera views) over a trajectory. Information from FT is used to accumulate the scores for facial regions corresponding to a same person in a scene over some window (e.g., $W=30$ frames). Fig. 7(a) illustrates the accumulation of matching score according to the

Table 5
Average pAUC(5%) and AUPR for TM-FR, TM-FR with self-update, MFR, SVDL, and AAMT-FR systems on all Chokepoint videos over 10 replications.

Systems	With Camera 2				With Camera Fusion			
	Transaction		Trajectory		Transaction		Trajectory	
	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR
TM-FR	0.243 ±0.03	0.353 ±0.02	0.312 ±0.05	0.369 ±0.03	0.249 ±0.02	0.457 ±0.04	0.318 ±0.02	0.423 ±0.03
TM-FR with self-update	0.291 ±0.02	0.389 ±0.01	0.373 ±0.04	0.478 ±0.03	0.347 ±0.03	0.452 ±0.05	0.391 ±0.02	0.478 ±0.02
MFR	0.397 ±0.05	0.414 ±0.01	0.413 ±0.03	0.485 ±0.04	0.417 ±0.03	0.512 ±0.05	0.472 ±0.02	0.532 ±0.02
SVDL	0.423 ±0.02	0.432 ±0.01	0.492 ±0.05	0.516 ±0.03	0.439 ±0.03	0.533 ±0.05	0.583 ±0.02	0.631 ±0.02
AAMT-FR	0.452 ±0.04	0.553 ±0.03	0.512 ±0.03	0.596 ±0.02	0.494 ±0.02	0.588 ±0.04	0.649 ±0.07	0.793 ±0.03

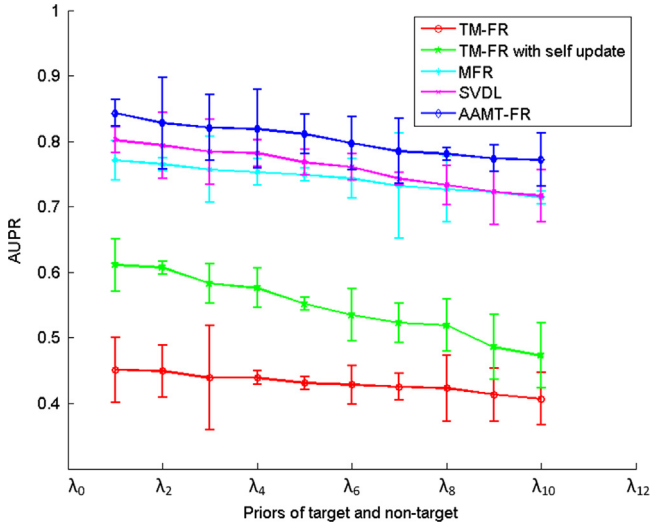

Fig. 11. Performance of the FR systems at different priors in the operational data.

frame count over time for a single trajectory of a video. Once individuals appear before a camera in video stream, and the related trajectories are processed, the performance of a FR system based on accumulated scores may be assessed over the range of decision threshold, and represented in the ROC space (see Fig. 7(b)).

The quality of the facial ROI captures in videos are also assessed with respect to the reference ROIs of the still gallery for describing best and worst case scenarios. For quality assessment, the luminance, contrast and structure measures are considered within a global quality measure called Structured Similarity Index (SSIM) [53], where

- *Luminance:* $l(\mathbf{x},\mathbf{y}) = (2\mu_x\mu_y + C_1)/(\mu_x + \mu_y + C_1)$, where $\mathbf{x}$ and $\mathbf{y}$ are the features extracted from a facial capture and a reference ROI pattern, respectively, and μ_x and μ_y are their means. C_1 is a constant included to avoid instability when $\mu_x + \mu_y + C_1$ is very close to zero.
- *Contrast:* $c(\mathbf{x},\mathbf{y}) = (2\sigma_x\sigma_y + C_2)/(\sigma_x + \sigma_y + C_2)$, where σ_x and σ_y are the standard deviations of $\mathbf{x}$ and $\mathbf{y}$, respectively. C_2 is a constant.
- *Structure:* $s(\mathbf{x},\mathbf{y}) = (\sigma_{xy} + C_3)/(\sigma_x\sigma_y + C_3)$, where $\sigma_{xy} = (1/(N-1)) \sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)$ and C_3 is a constant.

Luminance, contrast, and structure are combined into a global quality measure as follows:

$$SSIM(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})^{\alpha} c(\mathbf{x},\mathbf{y})^{\beta} s(\mathbf{x},\mathbf{y})^{\gamma}] = \frac{(2\mu_x\mu_y + C_1)(2\sigma_x y + C_2)}{(\mu_x + \mu_y + C_1)(\sigma_x + \sigma_y + C_2)} \tag{4}$$

where $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$.

5. Results and discussion

5.1. Transaction-level analysis

Fig. 8 shows an example of ROC and PR curves obtained when matching ROI patterns extracted from input videos (leaving sequences captured in Camera 2) with the templates of target IDs 03, 04, 07, 09, and 12 using TM-FR, TM-FR with self-update, and AAMT-FR systems. Fig. 8 (a) and (b) shows the ROC and inverted-PR curves obtained using the TM-FR systems on all leaving sequences captured in Camera 2, respectively. The dotted line in the figures indicates the operating point at $fpr = 5\%$ related to target ID 09. Results show that the TM-FR system performs poorly in most of the sequences, due to the limited representativeness of templates, and due to the significant discrepancy between probe and gallery data. The gallery represents templates extracted from high quality frontal shots, whereas the faces captured are of lower quality and vary with several nuisance factors. Along with camera interoperability issues, this results in poor overall performance.

Table 6
Average pAUC(5%), AUPR, and computation time (s/frame) for TM-FR, TM-FR with self-update, AAMT-FR, SVDL, and MFR systems.

Systems	Time complexity (s/frame)	Accuracy	
		pAUC	AUPR
TM-FR [8]	0.191 ± 0.05	0.312 ± 0.05	0.369 ± 0.03
TM-FR with self-update [9]	0.292 ± 0.07	0.373 ± 0.04	0.478 ± 0.03
MFR [27]	0.250 ± 0.08	0.413 ± 0.03	0.485 ± 0.04
SVDL [41]	0.237 ± 0.05	0.492 ± 0.05	0.516 ± 0.03
AAMT-FR	0.217 ± 0.06	0.512 ± 0.03	0.596 ± 0.02

Fig. 8 (c) and (d) shows the ROC and inverted PR curves for the TM-FR system with self-update for all the leaving sequences captured in Camera 2, respectively. Self-update allows for improved performance for IDs 09 and 12, yet degraded performance for ID 04. For IDs 03 and 07, the system shows comparable performance to the TM-FR system. Self-update improves performance if gallery-face-models are correctly updated with ROIs from the same target individual during operation. However, if ROI patterns are incorrectly selected for update, performance is degraded.

Fig. 8 (e) and (f) show the ROC and inverted PR curves obtained using AAMT-FR system for all the leaving sequences captured in Camera 2, respectively. This system learns faces during operation as in track-face-models. These models represent intra-class variations and improve face matching scores even if ROIs of a target individual differ somewhat. However, if the appearance of target faces vary abruptly within a trajectory, track-face-model may accept non-targets as targets and produce higher similarity scores even for the non-target individuals. Moreover, if the variation of a face appearance in the facial captures is very different from the templates in the gallery, track-face-model may produce poor classification scores for the target individual in the scene. Fig. 8(e) and (f) shows that the IDs 03 and 07 perform comparatively worse compare to the IDs 04, 09, and 12. However, AAMT-FR system performs better compare to the other methods.

Table 3 shows average AUC, pAUC, and AUPR, and Table 4 shows the average Precision (P), Recall (R), and F1 measures for a defined $fpr = 5\%$ for the TM-FR, TM-FR with self-update, and AAMT-FR systems considering all the entering and leaving sequences captured with Camera 2. ROC and inverted-PR curves for all the entering and leaving sequences (captured in all 3 Cameras) using the 3 FR systems are shown in Appendix A.1. Detailed results include all 3 camera views are shown in Appendix A.2, Tables A.7 and A.8. It is observed from Tables 3 and 4 that the AAMT-FR system significantly outperforms the others.

5.2. Trajectory-level analysis

Fig. 9 shows the transaction and trajectory level performance using only for Camera 2. For reference, Fig. 9(a), (c), and (e) shows the ROC curves for the TM-FR, TM-FR with self-update, and AAMT-FR systems, respectively, at the transaction level. Without score level fusion based on three camera views, only the frontal or near frontal views from a single camera provide ROIs for face matching. Trajectory level performance is shown in Fig. 9(b), (d), and (f) only for Camera 2. In all cases, accumulating scores over trajectories provide a significant higher level of performance through still-to-video recognition. The ROC curves obtained using all the Chokepoint videos processed with the TM-FR systems at transaction-level after score-level fusion of 3 cameras are shown in Fig. 10(a), (c), and (e). From these curves, it is apparent that overall the AAMT-FR also outperforms the others at transaction level after camera score fusion, where the fused scores are accumulated over time. Fig. 10(b), (d), and (e) shows the corresponding trajectory-level ROC curves (1 s, or 30 frames), along each trajectory. At trajectory level, output predictions are result of scores accumulated over time, and provide a higher level of accuracy and robustness through spatiotemporal recognition. The TM-FR, TM-FR with self-update, and AAMT-FR systems achieved improved performance for all the WL individuals compared to transaction-level ROC curves. The overall performance of the AAMT-FR system is the highest.

Table 5 presents a comparison of the overall transaction- and trajectory-level performance of the TM-FR, TM-FR with self-update, MFR, SVDL, and AAMT-FR systems. Both levels of analysis are shown with- and without camera fusion and considering all the sequences of the Chokepoint. To compare the global performance of the proposed system, the experiments are repeated 10 times, each time randomly selecting five targets (to form the WL) and 10 non-targets. It is observed in Table 5 that the proposed AAMT-FR outperforms other methods in all the cases. In MFR, the multiple representation of faces may fail to incorporate information of face variations that occur due to change in capture conditions (e.g., pose, scale, resolution, illumination, blur and expression) and camera inter-operability. Moreover, fusion of multiple representation is also challenging. In SVDL, the variation of face appearance in a generic set learned within gallery-face-model may be quite different from that of the single ROI set enlisted in the gallery, which may not be suitable to distinguish the target individuals of the gallery. AAMT-FR incorporates greater diversity of facial representation in the track-face-model that are captured during operation, and this leads to better discrimination of the target individuals for spatiotemporal recognition.

5.3. Performance analysis at different priors

In Fig. 11, performance of FR systems are observed in terms of AUPR by varying the proportion of matched ROIs of the target and non-target individuals in the operational data. This analysis uses the following levels of class imbalance of targets and non-targets, $\Lambda = \{\lambda_0, \lambda_1, \ldots, \lambda_{10}\} = \{1:1, 1:5000 \ldots, 1:50,000\}$. To measure system performance, the experiment is repeated 10 times, each time a different individual is randomly selected as target, and along with a growing number of different individuals as non-targets in the operational data. AUPR performance for 10 target individuals are calculated at different priors. Finally, the system performance at different priors is estimated by taking average of the AUPRs over the 10 different target individuals.

The performance of the AAMT-FR system is compared with the TM-FR, TM-FR with self update, MFR, and SVDL 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 are plotted in Fig. 11. The figure shows that the performance of all the systems are slightly declined as the priors of non-target

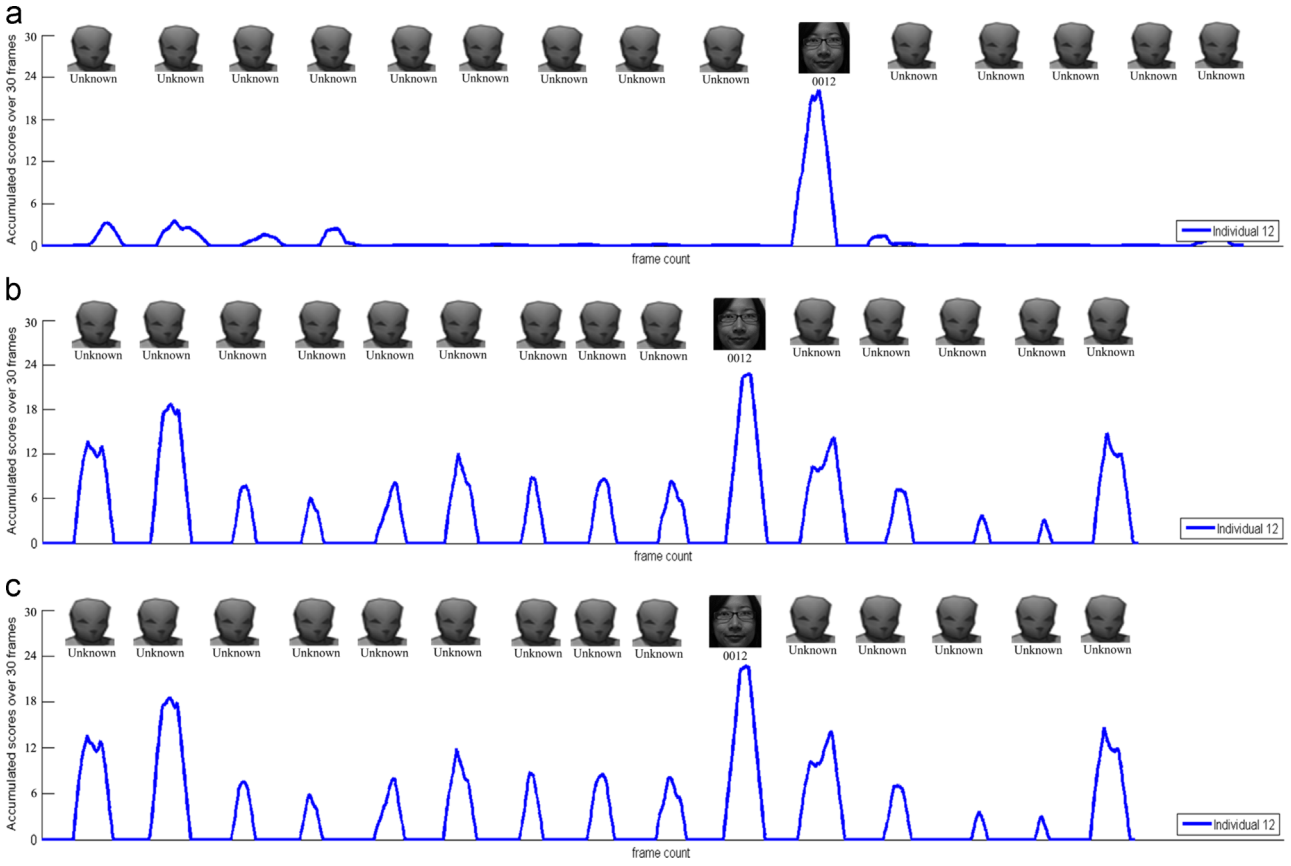

Fig. 12. Accumulated scores when inputs are matched against reference for ID 0004 in the gallery. (a) AAMT-FR. (b) TM-FR. (c) TM-FR with self-update.

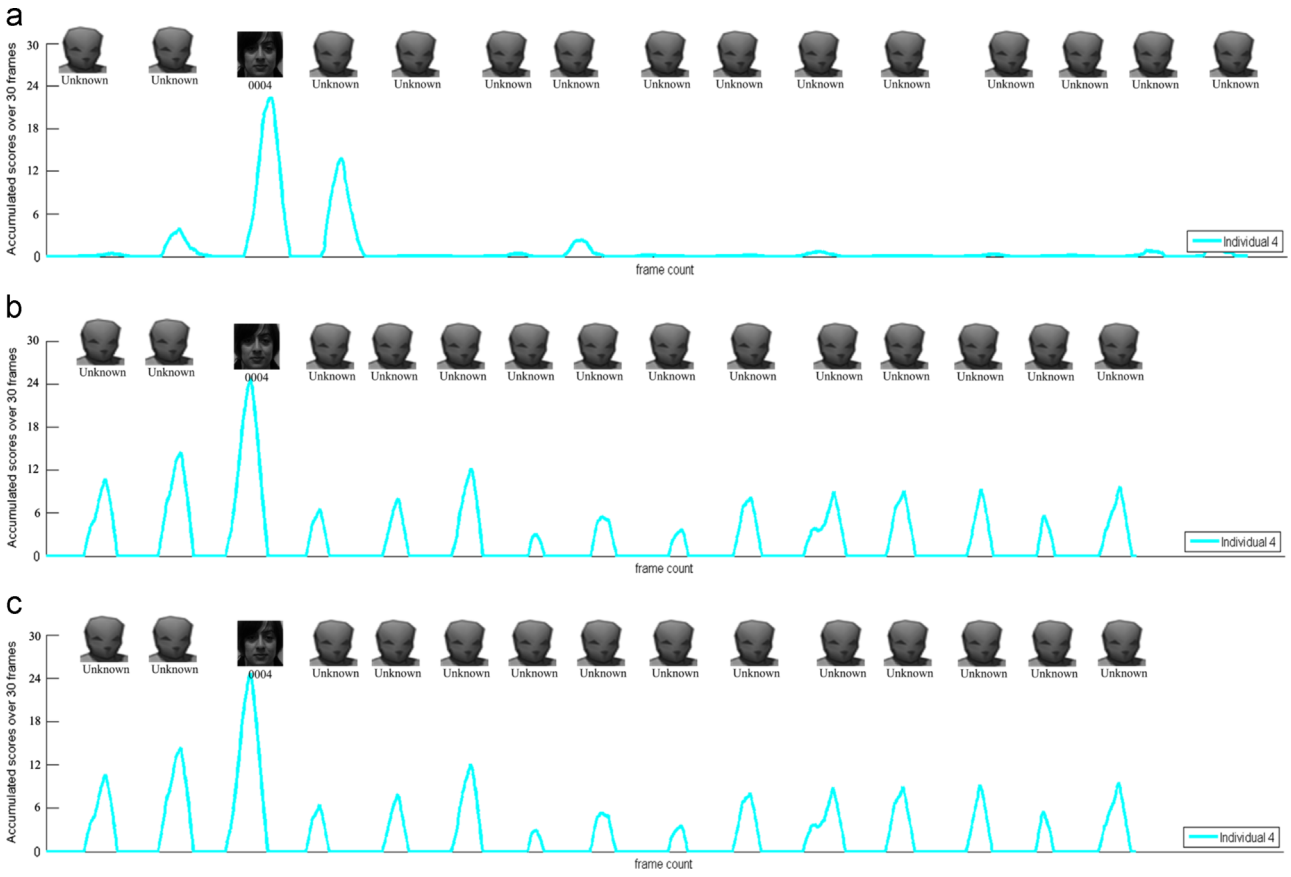

Fig. 13. Accumulated scores when inputs are matched against reference for ID 03 in the galley. (a) AAMT-FR. (b) TM-FR. (c) TM-FR with self-update.

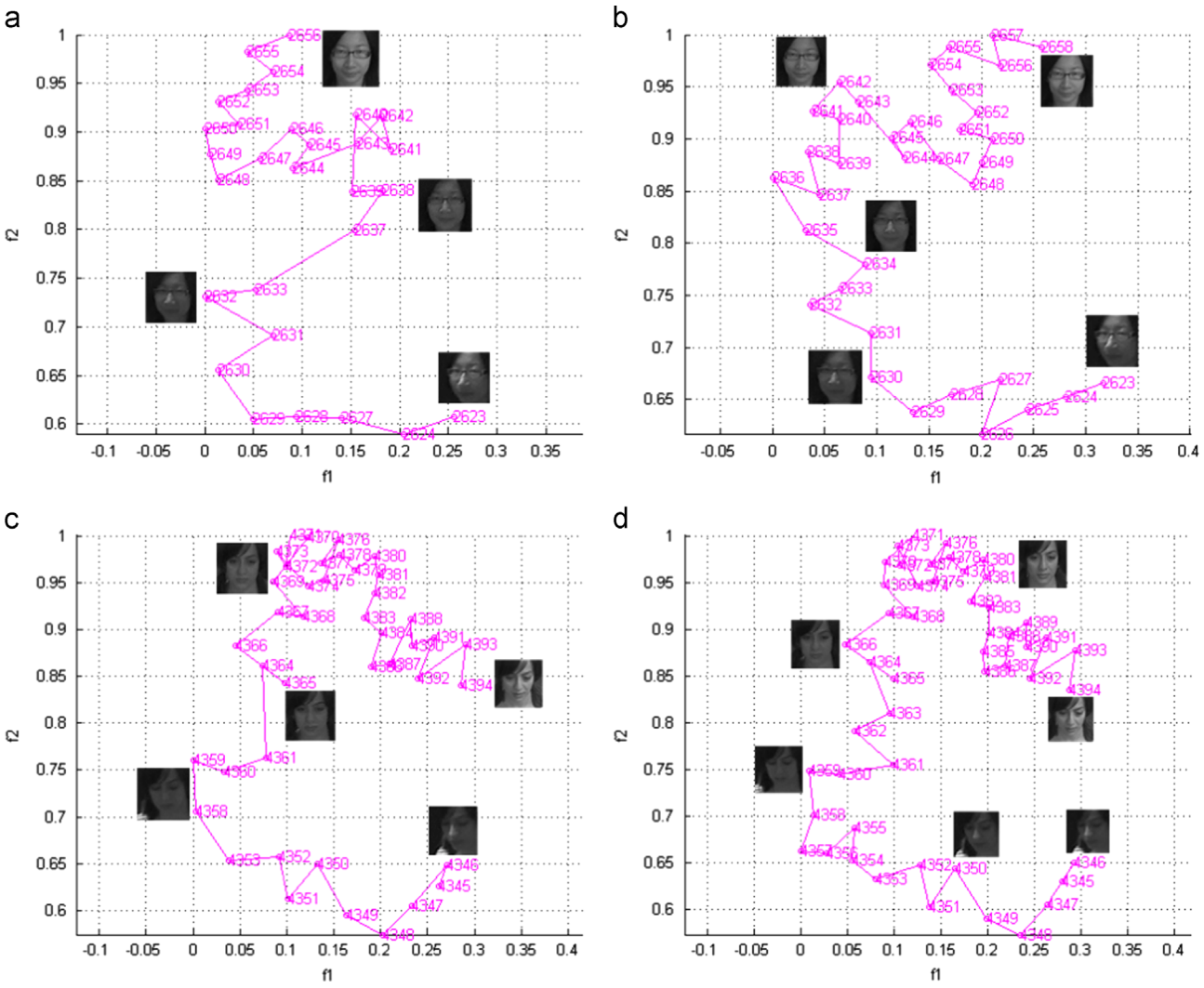

Fig. 14. Example of facial captures used in TM-FR and AAMT-FR systems in a good and a worse case scenario. (a) Captures of ID 12 used with TM-FR (good case). (b) Captures of ID 12 used in AAMT-FR (good case). (c) Captures of ID 04 used in TM-FR (worse case). (d) Captures of ID 04 used in AAMT-FR (worse case).

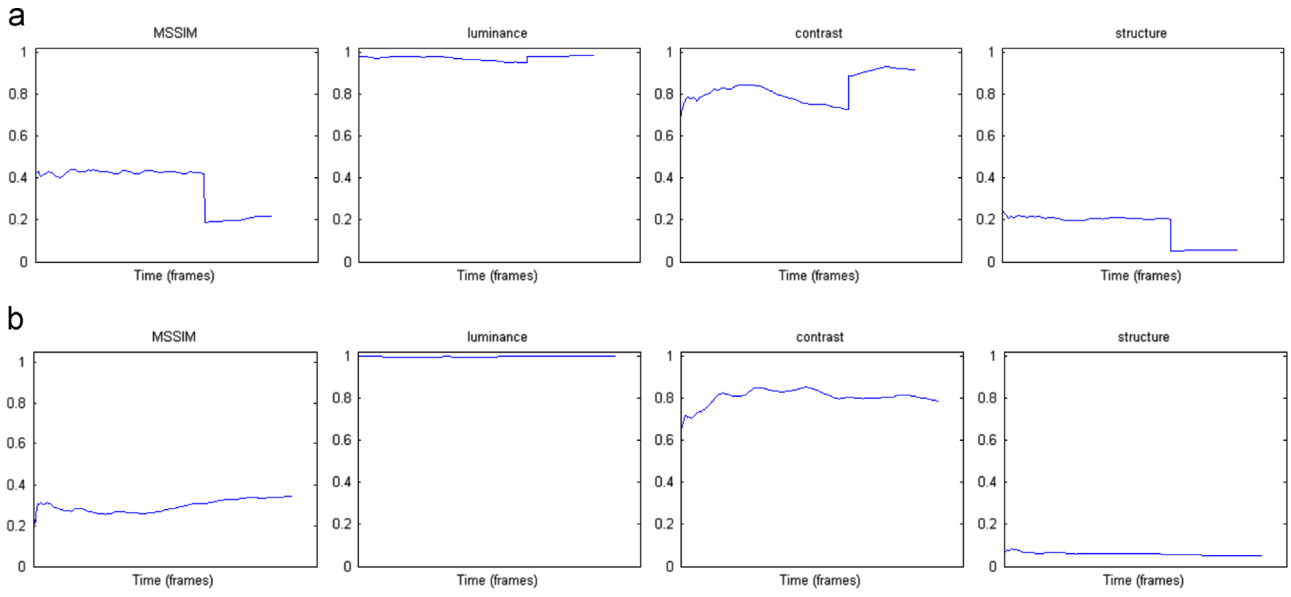

Fig. 15. Quality of facial captures used in AAMT-FR and TM-FR systems. (a) Captures of ID 04 with respect to template ID 04 (good case). (b) Captures of ID 03 with respect to template ID 03 (worse case).

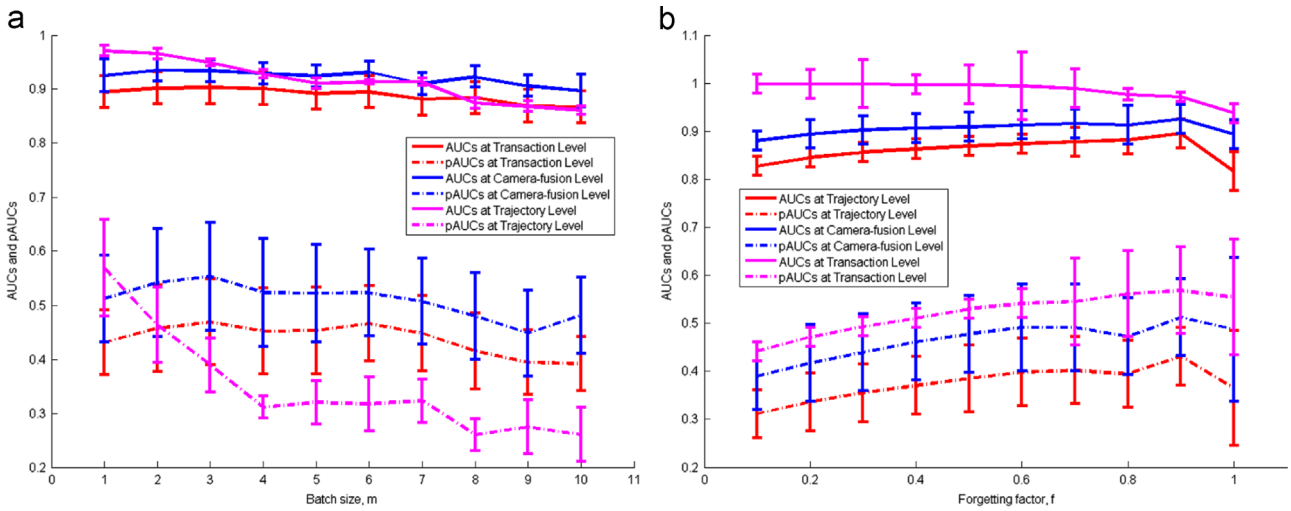

Fig. 16. The AUCs and pAUCs(5%) for different batch sizes and forgetting factors.

individuals increase compared to the target individuals in the operational data. Results indicate that the AUPR performance of all systems tends to decrease gradually as imbalance grows. Performance for TM-FR with self-update degrades sharply because of incorrect updates of the gallery-face-model. However, the proposed AAMT-FR system still outperforms others because the face-tracking-model incorporates diverse information of the facial captures through tracking.

5.4. Time complexity analysis

The computation effort required by the AAMT-FR is mainly in processing steps for face model update during tracking. For face model update, the AAMT-FR uses the SKL algorithm [13] whose computational complexity is $O(dm^2)$, where d and m refer to the dimensionality of the input feature vectors (HOG of the facial captures) 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 [54].

In SVDL, the complexity of commonly used l_1-*regularized* sparse coding is $O(m+d)^{\epsilon}$, where m is the number of dictionary atoms, d is the dimensionality of the features, and ϵ is an error constant. In TM-FR with self-update, the main computation is required for gallery-face-model update, where the eigenspace is updated by re-computing the principal components matrix with the increased training set. This operation requires $O(d^3+d^2m)$ computations, where d and m refer to the dimensionality and the number of feature vectors used, respectively [55]. The computational complexity for TM-FR is $O(d)$, where d is the dimensionality of feature vectors. It does not update face model and the computational complexity that it requires is mainly for the template matching. In MFR, patch based method is used for feature extraction, where the selection of patch size is important, because the amount of information that can be provided by a patch affects the performance of FR system. However, bigger patches provide much information about the region but increase the complexity of processing. In the experiment, for each facial capture, 16 patches with size 12×12 have been used.

The computation time, and pAUC(5%) and AUPR performance for different FR systems are shown in Table 6. At the trajectory level, the performance for each system has been improved over the transaction level because of score accumulation. 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 for the FR systems include the times required for the tracking and the recognition. Result indicate that the AAMT-FR provides its highest level of performance for a relatively low time complexity.

5.5. Detailed analysis

5.5.1. Best and worst cases performance

Fig. 12(a)–(c) shows the matching scores over time obtained using TM-FR, TM-FR with self-update, and AAMT-FR systems, respectively, where the template for ID 12 is compared with faces captured in the P1E_S1_C1 video. This is a good case for the AAMT-FR system, since it provides discrimination scores between ROI of targets and non-target individuals in this scene. This leads to a significantly higher level of performance than the other systems because it allows incorporating diverse information on facial-appearance changes into the track-face-models during operation. Although the TM-FR with self-update system allow upgrading gallery-face-models based on operational data, the use of conservative threshold to avoid incorporating non-target information prevents updating properly, even when target facial captures become available in the scene. In contrast, the AAMT-FR system exploits kinematic models within the tracker to support online learning of tracker face models. The scores over trajectory for the models for TM-FR, TM-FR with self-update, and AAM-FR systems for all 5 target WL individuals in the scene are presented in Appendix B, Figs. B.23, B.24, and B.25, respectively.

A worse case scenario for AAMT-FR is shown in Fig. 13, where the AAMT-FR, TM-FR, and TM-FR with self-update systems compare template of ID 04 with the faces captured in the video P1E_S1_C1 video. Fig. 13(a) shows that the AAMT-FR method produces high classification scores for facials captures linked to ID 04 (target individual) and ID 05 (another) in the scene. In such situation, lower range is allowed for the selection of decision threshold which may lead to false detection. However, the TM-FR and TM-FR with self-update provide better discriminant scores in this case.

In Fig. 13, the ROIs for individuals with ID 04 and ID 12 are plotted in the 81 feature HOG space of ROI patterns and then projected in a 2D space using Sammon mapping. Since both TM-FR and TM-FR with self-update systems use ROIs captured through face detection, fewer high quality captures are available w.r.t. to the AAMT-FR. Thus, TM-FR and even TM-FR with self-update process facial information with less diversity of appearance than with AAMT-FR.

Since the AAMT-FR uses facial captures obtained from a tracker, it can incorporate more diversity information on facial appearances during operations. With AAMT-FR, the faces of a person in the scene are continually tracked and used for on-line learning of track-face-model without any interruption until the track is dropped (due to lower track confidence). However, if the appearance of a face changes abruptly, the track-face-model may become corrupted with tracker ROI patterns that are significantly different from the corresponding template, and may also result in degraded FR performance.

Fig. 14(a) and (b) shows the scatter plots of the facial captures for corresponding ID 12 used in TM-FR and AAMT-FR methods, respectively. It is observed from the figures that the TM-FR used fewer facial captures from segmentation than from tracker. In videos, the diversity of facial appearance (from segmentation) in video is moderate and complies with the template stored for WL individual ID 12. Both the systems perform well for ID 12 in this video. Fig. 14(c) and (d) shows the plot of the facial captures for ID 04 used in TM-FR and AAMT-FR systems, respectively. The diversity of facial appearance for ID 04 in video is high with respect to the corresponding template in the galley. The abrupt changes in appearance do not correspond to its template and result in a low level of performance for both systems.

Plots of SSIM image quality along with its components (luminance, contrast, and structure) are shown in Fig. 15(a)–(c), respectively, for the facial captures in video of individual ID 03 (with respect to template ID 03) and ID 04 (with respect to template 04).

The AAMT-FR appears to perform better when most of the facial captures of an individual have relatively high SSIM quality with respect to the template from the same individual. Fig. 15(a) shows the relatively high SSIM quality of the facial captures of ID 04 with respect to its template which corresponds to higher performance in Fig. 12(a), Tables 3 (row 5) and 4 (row 5).

If the quality of few facial captures with respect to the template from the same individual is high in a video, TM-FR may perform better than the AAMT-FR. In such situation, the TM-FR system with self-update may degrade the performance if the face models are incorrectly updated. Fig. 15(b) shows that the quality for the facial captures from ID 03 are lower with respect to the template of ID 03, mostly due to the SSIM structure component and poor discrimination for ID 03 leads to lower AUC and AUPR performance for AAMT-FR (see Tables 4 and 5). However, in case of TM-FR and TM-FR with self-update method, they achieved comparable performance.

Fig. 14 suggests that limitation of AAMT-FR is that track-face-models not generated with high quality facial captures may adapt to operational information which is not related to target template. In such situation, AAMT-FR may not produce better discrimination than TM-FR. In case of TM-FR, it does not adapt the face models; thus, it is not affected by knowledge corruption from operational captures as with the AAMT-FR system.

5.5.2. *Sensitivity analysis*

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. 16 (a) shows the AUC and pAUC(5%) produced by AAMT-FR system while varying m. In this case m is changed from 1 to 10 while keeping the $f = 0.9$. This figure shows that if the batch size is increased, the performance declines as track-face-models 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. 16 (b) shows the performance of the AAMT-FR while varying f between 0 and 1 and fixing $m = 1$. Here, $f = 0$ indicates forget everything whereas with the higher value of f, it allows us to remember more past observations. AAMT-FR performance tends to increase with the value of f.

6. Conclusion

This paper presents a system, Adaptive Appearance Model Tracker-based Face Recognition (AAMT-FR), for still-to-video face recognition that is specialized for watch-list screening applications, to address problems with SSPP (single sample per person). Inside the system, an adaptive appearance model tracker (AAMT) is used that gradually learns a track-face-model per individual with the facial captures appearing in the scene. The track-face-models are gradually matched over time against the reference still images from the individuals of interests, and the matching scores are accumulated over several frames and multiple cameras for spatiotemporal recognition. Performance of the proposed system has been validated with a generic still-to-video FR systems, where each gallery-face-model corresponds to a template (ROI pattern). However, each gallery-face-model can also be defined by several templates or a statistical model. In all cases, AAMT-FR system will provide the same over state-of-the-art systems for still-to-video FR.

There are several advantages of the face-track-model for still to video FR. It can learn the intra-class variability of the facial appearance online while tracking the face in the scene. Since the face-track-model relies on facial captures obtained from the tracker for face modeling, and updates the model incrementally, it performs reasonably well in FR even if a limited number of ROI patterns are stored in the WL gallery. Moreover, face modeling with the tracking faces helps us to incorporate information on face appearance changes under different real-world capture conditions, which is not possible in synthetic ROI pattern generation or enlarging the training set using auxiliary data. Because of incremental update, the robustness of the face model in matching for different viewing condition is gradually improved over time.

Any separate preprocessing steps to select key frames has not been applied in the proposed system. Key-frame selection techniques have their limitations [56], such as neglecting of essential contents and limited maintenance of dynamic contents in the video. The intuition for learning track-face-model in the proposed system is to incorporate more representative intra-class variations of the facial appearance in order to improve face recognition accuracy. Selecting key frames may significantly reduce the number facial captures as well as diversity information in face model. This may negatively affect FR performance. Since the track-face-model is updated within a tracker in the proposed system, it is more likely to update the model with captures from the same person in a scene. However, some incorrect update may occur due for example to occlusion, which typically leads to low tracking confidence, and eventually it drops the track considering that the face has disappeared from the scene. When a track is dropped for an individual, the corresponding track-face-model is also eliminated from the system and is no longer used

for FR. Thus, a track-face-model with incorrect update will have limited impact on FR performance. If the face appears in the scene again, the segmentation module detects the face and initiate a new track-face-model for that individual, and trajectory.

The limitation of track-face-model is that it may be incorrectly updated during online learning using tracking, which may increase false positives in FR. To solve this issue, research could be conducted to enforce constrained on the overall appearance of the face model to determine the acceptable level of plasticity of the track-face-model. AAMT based methods are still rather complex. These methods may be problematic for FR as their computational complexity grows with the number of facial tracks, size of faces, camera resolution and frame rate, and the number of cameras used by the surveillance system. Further research measures to reduce the complexity of these methods are also important.

Conflict of interest

There is no conflict of interest identified in this research.

Acknowledgments

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

Appendix A. Supplementary data

Supplementary data associated with this paper can be found in the online version at http://dx.doi.org/10.1016/j.patcog.2015.08.002.

References

[1] F. Matta, J.-L. Dugelay, Person recognition using facial video information: a state of the art, J. Vis. Lang. Comput. 20 (2009) 180–187.
[2] R. Chellappa, M. Du, P. Turaga, S.K. Zhou, Face Tracking and Recognition in Video, Springer-Verlag London Limited, Princeton, 2011, pp. 323–351.
[3] Z. Huang, X. Zhaoa, S. Shan, R. Wang, X. Chen, Coupling alignments with recognition for still-to-video face recognition, In: International Conference in Computer Vision, Sydney, Australia, 2013.
[4] S. Zhou, V. Krueger, R. Chellappa, Probabilistic recognition of human faces from video, Comput. Vis. Image Underst. 91 (2003) 214–245.
[5] C. Pagano, E. Granger, R. Sabourin, D. Gorodnichy, Detector ensembles for face recognition in video surveillance, In: International Joint Conference on Neural Networks, Brisbane, Australia, vol. 91, 2012, pp. 1–8.
[6] X. Tan, S. Chen, Z.-H. Zhou, F. Zhang, Face recognition from a single image per person: a survey, Pattern Recognit. 39 (9) (2006) 1725–1745.
[7] A. Rattani, B. Freni, G.L. Marcialis, Template update methods in adaptive biometric systems: a critical review, In: Advances in Biometrics, 2009, pp. 847–856.
[8] R. Brunelli, T. Poggio, Face recognition: features versus templates, IEEE Trans. Pattern Anal. Mach. Intell. 15 (10) (1993) 1042–1052.
[9] F. Roli, G. Marcialis, Semi-supervised pca-based face recognition using self-training, In: Lecture Notes in Computer Science, vol. 4109, 2006, pp. 560–568.
[10] X. Yan, X. Wu, I.A. Kakadiaris, Shishir K. Shah, To track or to detect? an ensemble framework for optimal selection, In: European Conference on Computer Vision, Firenze, Italy, 2012, pp. 594–607.
[11] M.A.A. Dewan, E. Granger, F. Roli, R. Sabourin, G.L. Marcialis, Comparison of adaptive appearance methods for tracking faces in video surveillance, In: International Conference on Imaging for Crime Detection and Prevention, Kingstone, UK, 2013, pp. 1–7.
[12] S. Salti, A. Cavallaro, Adaptive appearance modeling for video tracking: survey and evaluation, IEEE Trans. Image Process. 21 (10) (2012) 4334–4348.
[13] A. Levy, M. Lindenbaum, Sequential Karhunen–Loeve basis extraction and its application to images, IEEE Trans. Image Process. 9 (8) (2000) 1371–1374.
[14] M.D. la Torrea, E. Granger, P.V. Radtkea, R. Sabourina, D.O. Gorodnichyc, Partially-supervised learning from facial trajectories for face recognition in video surveillance, Inf. Fus. (2014) 1–23.
[15] Y. Wong, S. Chen, S.M.C. Sanderson, B.C. Lovell, Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition, In: IEEE Computer Vision and Pattern Recognition Workshops, CO, USA, 2011, pp. 74–81.
[16] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. 57 (2) (2004) 137–154.
[17] T. Poggio, Kah-Kay Sung, Example-based learning for view-based human face detection, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1) (1998) 39–51.
[18] H. Schneiderman, T. Kanade, Statistical method for 3d object detection applied to faces and cars, In: IEEE Conference on Computer Vision and Pattern Recognition, SC, USA, 2000.
[19] M.J. Black, A.D. Jepson, Eigentracking: robust matching and tracking of articulated objects using a view-based representation, Int. J. Comput. Vis. 26 (1) (1998) 63–84.
[20] D.J. Fleet, A.D. Jepson, T.F. El-Marahgi, Robust online appearance models for visual tracking, IEEE Trans. Pattern Anal. Mach. Intell. 25 (10) (2003) 1296–1311.
[21] D.A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (1) (2008) 125–141.
[22] B. Babenko, M.-H. Yang, S. Belongie, Visual tracking with online multiple instance learning, In: IEEE Conference on Computer Vision and Pattern Recognition, CA, USA, 2009, pp. 983–990.
[23] W.X.Q. Wang, F. Chen, M.-H. Yang, Online discriminative object tracking with local sparse representation, In: IEEE Workshop on Application of Computer Vision, Beijing, China, 2012, pp. 425–432.
[24] Z. Kalal, K. Mikolajczyk, J. Matas, Tracking-learning-detection, IEEE Trans. Pattern Anal. Mach. Intell. 34 (7) (2012) 1409–1422.
[25] F. Roli, L. Didaci, G. Marcialis, Template co-update in multimodal biometric systems, In: International Conference on Advances in Biometrics, Seoul, South Korea, 2007, pp. 1194–1202.
[26] A. Rattani, G. Marcialis, F. Roli, Capturing large intra-class variations of biometric data by template co-update, In: IEEE Workshop on Biometrics, Alaska, USA, 2008, pp. 1–6.
[27] S. Bashbaghi, R.S.E. Granger, G. Bilodeau, Watch-list screening using ensembles based on multiple face representations, In: International Conference on Pattern Recognition, Stockholm, Sweden, 2014, pp. 1–6.
[28] B. Klare, A.K. Jain, On a taxonomy of facial features, In: IEEE International Conference on Biometrics: Theory Applications and Systems, Washington, DC, USA, 2010, pp. 1–8.
[29] W. Gao, S. Shan, X. Chai, X. Fu, Virtual face image generation for illumination and pose insensitive face recognition, In: IEEE International Conference on Acoustics, Speech, and Signal, Beijing, China, 2003, pp. 149–152.
[30] F. Mokhayeriy, E. Granger, G.-A. Bilodeau, Synthetic face generation under various operational conditions in video surveillance, In: International Conference on Image Processing, Quebec, Canada, 2015.
[31] B. Kamgar-Parsi, W. Lawson, B. Kamgar-Parsi, Toward development of a face recognition system for watchlist surveillance, IEEE Trans. Pattern Anal. Mach. Intell. 33 (10) (2011) 1925–1937.
[32] Y. Xu, X. Zhu, Z. Li, G. Liu, Y. Lu, H. Liu, Using the original and symmetrical face training samples to perform representation based two-step face recognition, Pattern Recognit. 46 (2013) 1151–1158.
[33] S. Chen, D. Zhang, Z.-H. Zhou, Enhanced (pc)2a for face recognition with one training image per person, Pattern Recognit. Lett. 25 (10) (2004) 1173–1181.
[34] H.-C. Jung, B.-W. Hwang, S.-W. Lee, Authenticating corrupted face image based on noise model, In: IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, South Korea, 2006, pp. 1–6.
[35] J. Liu, S. Chen, Z.-H. Zhou, X. Tan, Single image subspace for face recognition, In: Analysis and Modeling of Faces and Gestures, Rio de Janeiro, Brazil, 2007, pp. 205–219.

[36] A. Majumdar, R. Ward, Single image per person face recognition with images synthesized by non-linear approximation, In: International Conference on Image Processing, San Diego, CA, USA, 2008, pp. 2740–2743.
[37] T. Shan, B.C. Lovell, S. Chen, Face recognition robust to head pose from one sample image, In: International Conference on Pattern Recognition, Hong Kong, China, 2006, pp. 515–518.
[38] F. Hafiz, A.A. Shafie, Y.M. Mustafah, Face recognition from single sample per person by learning of generic discriminant vectors, Proc. Eng. 41 (2012) 465–472.
[39] Y. Su, S. Shan, X. Chen, W. Gao, Adaptive generic learning for face recognition from a single sample per person, In: IEEE Conference on Computer Vision and Pattern Recognition, CA, USA, 2010, pp. 2699–2706.
[40] T. Poggio, T. Vetter, Recognition and Structure from One 2d Model View: Observations on Prototypes, Object Classes and Symmetries, Artificial Intelligence Laboratory, MIT, Cambridge, MA, A.I. Memo No. 1347.
[41] M. Yang, L. Van, L. Zhang, Sparse variation dictionary learning for face recognition with a single training sample per person, In: IEEE International Conference on Computer Vision, DC, USA, 2013.
[42] C. Shan, Face recognition and retrieval in video, In: Video Search and Mining Studies in Computational Intelligence, vol. 287, 2010, pp. 235–260.
[43] X. Liu, T. Cheng, Video-based face recognition using adaptive hidden Markov models, In: IEEE Computer Society Conference Computer Vision and Pattern Recognition, Toronto, Canada, 2003, pp. 340–345.
[44] Y. Zhang, A.M. Martinez, A weighted probabilistic approach to face recognition from multiple images and video sequences, Image Vis. Comput. 24 (2006) 626–638.
[45] S. Zhou, R. Chellappa, B. Moghaddam, Visual tracking and recognition using appearance-adaptive models in particle filters, IEEE Trans. Image Process. 13 (11) (2004) 1491–1506.
[46] E. Granger, M. Barry, Face recognition in video using a what-and-where fusion neural network, In: International Joint Conference on Neural Networks, Orlando, USA, 2007, pp. 2256–2261.
[47] E. Granger, M.A. Rubin, S. Grossberg, P. Lavoie, A what-and-where fusion neural network for recognition and tracking of multiple radar emitters, Neural Netw. 14 (3) (2001) 325–344.
[48] H. Ekenel, J. Stallkamp, R. Stiefelhagen, Avideo-based door monitoring system using local appearance-based face models, Comput. Vis. Image Underst. 114 (5) (2010) 596–608.
[49] V. Despiegel, S. Gentric, J. Fondeur, Border control: from technical to operational evaluation, In: International Biometric Performance Testing Conference, MD, USA, 2012, pp. 1–3.
[50] M. Isard, A. Blake, Contour tracking by stochastic propagation of conditional density, In: European Conference on Computer Vision Cambridge, Cambridge, UK, vol. 1, 1996, pp. 343–354.
[51] B. Han, S.-W. Joo, L.S. Davis, Multi-camera tracking with adaptive resource allocation, Int. J. Comput. Vis. 91 (1) (2011) 45–58.
[52] T. Landgrebe, P. Paclik, R. Duin, A. Bradley, Precision–recall operating characteristic (p-roc) curves in imprecise environments, In: International Conference on Pattern Recognition, Hong Kong, China, 2006.
[53] Z. Wang, A. Conrad, B. Hamid, R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 6000–6012.
[54] J.D. Hol, T.B. Schon, F. Gustafsson, On resampling algorithms for particle filters, Cambridge, UK, In: IEEE Nonlinear Statistical Signal Processing Workshop, 2006.
[55] A. Sharma, K.K. Paliwal, Fast principal component analysis using fixed-point algorithm, Pattern Recognit. Lett. 28 (2007) 1151–1155.
[56] X. Li, T. Xu, Face video key-frame extraction algorithm based on color histogram, In: International Conference on Computer Science and Information Technology, 2011.

M. Ali Akber Dewan received his B.Sc. degree in Computer Science and Engineering from Khulna University, Bangladesh, in 2003. He received his Ph.D. degree in Computer Engineering from Kyung Hee University, South Korea, in 2009. From 2003 to 2009, he was a Faculty Member in the Department of Computer Science and Engineering, Chittagong University of Engineering and Technology, Bangladesh. From 2009 to 2012, he was a Postdoctoral Fellow at Concordia University, Canada. From 2012 to 2014, he was a Postdoctoral Fellow at École de Technologie Supérieure, Université du Québec, Canada. In October 2014, he joined the Faculty of Athabasca University, Canada, where he is currently an Assistant Professor of Computing and Information Systems. His research interests include image processing, computer vision, machine learning, biometric recognition, motion detection, tracking, medical image analysis, and multimedia technology.

Eric Granger obtained a Ph.D. in Electrical Engineering from the École Polytechnique de Montréal in 2001. From 1999 to 2001, he was a Defence Scientist at Defence R&D Canada in Ottawa. Until then, his work focused primarily on neural networks for fast classification of radar signals in Electronic Surveillance (ES) systems. From 2001 to 2003, he worked in R&D with Mitel Networks Inc. on algorithms and electronic circuits to implement cryptographic functions in Internet Protocol (IP) based communication platforms. In 2004, he joined the ETS, Université du Québec, where he has developed applied research activities in the areas of patterns recognition, computer vision and microelectronics. He presently holds the rank of Full Professor in System Engineering. Since joining ÉTS, he has been a member of the Laboratoire d'imagerie, de vision et d'intelligence artificielle (LIVIA), and his main research interests are adaptive classification systems, incremental learning, change detection, and multiclassifier systems, with applications in biometrics, video surveillance, and intrusion detection in computer and network security.

Gian Luca Marcialis received his M.S. degree and Ph.D. degree in Electronic and Computer Science Engineering from the University of Cagliari, Italy, in 2000 and 2004, respectively. He is currently an Assistant Professor and member of the Pattern Recognition and Applications group at the Electrical and Electronic Engineering Department of the University of Cagliari. His research interests are in the fields of fusion of person recognition by multiple biometrics, vulnerability analysis and template updating of biometric systems. Gian Luca Marcialis is the member of the Institute of Electrical and Electronic Engineers (IEEE) and the International Association for Pattern Recognition (IAPR).

Robert Sabourin joined the Physics Department of the Montreal University in 1977 where he was responsible for the design, experimentation and development of scientific instrumentation for the Mont Mégantic Astronomical Observatory. His main contribution was the design and the implementation of a microprocessor based fine tracking system combined with a low light level CCD detector. In 1983, he joined the staff of the École de Technologie Supérieure, Université du Québec, in Montréal where he cofounded the Department of Automated Manufacturing Engineering where he is currently a Full Professor, and teaches Pattern Recognition, Evolutionary Algorithms, Neural Networks and Fuzzy Systems. In 1992, he joined also the Computer Science Department of the Pontifícia Universidade Católica do Paraná (Curitiba, Brazil) where he was co-responsible for the implementation in 1995 of a master program and in 1998 a Ph.D. program in Applied Computer Science. Since 1996, he is a Senior Member of the Centre for Pattern Recognition and Machine Intelligence (CENPARMI, Concordia University). Since 2012, he is the Research Chair holder specializing in Adaptive Surveillance Systems in Dynamic Environments. Dr. Sabourin is the author (and coauthor) of more than 300 scientific publications including journals and conference proceeding. He was co-chair of the program committee of CIFED'98 (Conférence Internationale Francophone sur l'Écrit et le Document, Québec, Canada) and IWFHR'04 (9th International Workshop on Frontiers in Handwriting Recognition, Tokyo, Japan). He was nominated as Conference co-chair of ICDAR'07 (9th International Conference on Document Analysis and Recognition) that has been held in Curitiba, Brazil in 2007. His research interests are in the areas of adaptive biometric systems, adaptive surveillance systems in dynamic environments, intelligent watermarking systems, evolutionary computation and bio-cryptography.

Fabio Roli received the M.S. (Hons.) and Ph.D. degrees in Electronic Engineering from the University of Genoa, Genoa, Italy. From 1988 to 1994, he was at the University of Genoa as a Research Group Member in Image Processing and Understanding. He was an Adjunct Professor at the University of Trento, Trento, Italy, from 1993 to 1994. In 1995, he joined the Department of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy, where he is currently a Professor of Computer Engineering and a Director of the Research Laboratory on Pattern Recognition and Applications. His current research interests include the design of pattern recognition systems and their applications to biometric personal identification, multimedia text categorization, and computer security. Prof. Roli is a Governing Board Member of the International Association for Pattern Recognition and the IEEE Systems, Man and Cybernetics Society. He is a fellow of the International Association for Pattern Recognition. He is a fellow of the IEEE and a fellow of the International Association for Pattern Recognition.