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].