Information Fusion 24 (2015) 31–53

Contents lists available at ScienceDirect

Information Fusion

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

Partially-supervised learning from facial trajectories for face recognition in video surveillance

Miguel De-la-Torre [a,b,*], Eric Granger [a], Paulo V.W. Radtke [a], Robert Sabourin [a], Dmitry O. Gorodnichy [c]

[a] Laboratoire d'imagerie de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montréal, Canada
[b] Centro Universitario de Los Valles, Universidad de Guadalajara, Ameca, Mexico
[c] Science and Engineering Directorate, Canada Border Services Agency, Ottawa, Canada

ARTICLE INFO

Article history:
Available online 21 June 2014

Keywords:
Multiple classifier systems
Adaptive biometrics
Incremental learning
Video-to-video face recognition
Video surveillance

ABSTRACT

Face recognition (FR) is employed in several video surveillance applications to determine if facial regions captured over a network of cameras correspond to a target individuals. To enroll target individuals, it is often costly or unfeasible to capture enough high quality reference facial samples a priori to design representative facial models. Furthermore, changes in capture conditions and physiology contribute to a growing divergence between these models and faces captured during operations. Adaptive biometrics seek to maintain a high level of performance by updating facial models over time using operational data. Adaptive multiple classifier systems (MCSs) have been successfully applied to video-to-video FR, where the face of each target individual is modeled using an ensemble of 2-class classifiers (trained using target vs. non-target samples). In this paper, a new adaptive MCS is proposed for partially-supervised learning of facial models over time based on facial trajectories. During operations, information from a face tracker and individual-specific ensembles is integrated for robust spatio-temporal recognition and for self-update of facial models. The tracker defines a facial trajectory for each individual that appears in a video, which leads to the recognition of a target individual if the positive predictions accumulated along a trajectory surpass a detection threshold for an ensemble. When the number of positive ensemble predictions surpasses a higher update threshold, then all target face samples from the trajectory are combined with non-target samples (selected from the cohort and universal models) to update the corresponding facial model. A learn-and-combine strategy is employed to avoid knowledge corruption during self-update of ensembles. 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 the ensembles evolves. For proof-of-concept, a particular realization of the proposed system was validated with videos from Face in Action dataset. Initially, trajectories captured from enrollment videos are used for supervised learning of ensembles, and then videos from various operational sessions are presented to the system for FR and self-update with high-confidence trajectories. At a transaction level, the proposed approach outperforms baseline systems that do not adapt to new trajectories, and provides comparable performance to ideal systems that adapt to all relevant target trajectories, through supervised learning. Subject-level analysis reveals the existence of individuals for which self-updating ensembles with unlabeled facial trajectories provides a considerable benefit. Trajectory-level analysis indicates that the proposed system allows for robust spatio-temporal video-to-video FR, and may therefore enhance security and situation analysis in video surveillance.

1. Introduction

In video surveillance applications, automated face recognition (FR) systems are increasingly employed to match facial regions of interest (ROIs) captured across a network of video cameras to individuals of interest enrolled to the system. These applications range from watchlist screening, which involves still-to-video FR, to person re-identification (for search and retrieval), which involves

* Corresponding author at: Laboratoire d'imagerie de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, 1100, rue Notre-Dame Ouest, Montréal (Qc) H3C 1K3, Pavillon principal (A), office A-3600, Montréal, Canada. Tel.: +1 514 396 8800x7687.

E-mail addresses: miguel@livia.etsmtl.ca (M. De-la-Torre), eric.granger@etsmtl.ca (E. Granger), radtke@livia.etsmtl.ca (P.V.W. Radtke), robert.sabourin@etsmtl.ca (R. Sabourin), dmitry.gorodnichy@cbsa-asfc.gc.ca (D.O. Gorodnichy).

http://dx.doi.org/10.1016/j.inffus.2014.05.006
1566-2535/© 2014 Elsevier B.V. All rights reserved.

video-to-video FR. Regardless, systems for FR in video surveillance (FRiVS) must operate under semi- and unconstrained capture conditions, where scale, pose, occlusion, blur/resolution, expression and illumination vary over time.

A facial model used for matching may be defined as a set of one or more reference samples (for a template matching system), or a statistical model estimated through training with reference samples (for a neural or statistical classification system). In video-to-video FR, reference samples extracted from ROIs captured in video streams are employed to design of facial models, integrating time and space information in facial models [1,2]. In still-to-video FR, reference samples are extracted from one or more still images.

In video surveillance, individuals in a scene may be tracked, and the facial ROIs captured in videos that correspond to different individuals may be regrouped over multiple frames for robust spatio-temporal recognitions [3]. Tracking information can, for instance, be used to record a complete trajectory,[1] from the arrival of individual in the scene until he leaves. Predefined thresholds have been applied to matching scores and image quality measurements to produce overall decisions based on the consecutive ROIs [4]. In addition, the sum rule has been applied over the matching scores produced by ROIs in a trajectory [5]. Tracking information as also been used to model the joint posterior distribution of the motion and identity for the individual in the scene [6].

This paper concerns system for video-to-video FR, where facial models for matching are defined as a statistical model. Facial models are usually designed during enrollment, ideally using several high quality reference ROIs captured for the target individual under controlled conditions. In video-to-video FR, these reference ROIs are extracted along one or more reference trajectories. This requirement is rarely fulfilled in practical applications, and enrollment of individuals often relies on a limited number of lower quality ROIs. FR performance tends to decline since facial models are not representative of the faces to be recognized during operations. Both abrupt and gradual changes in capture conditions (due to, e.g., aging and variations in pose and lighting) also lead to a decline in FR performance due to a growing divergence between these facial models and faces captured during operations. Several adaptive classifiers have been proposed in literature for supervised incremental learning of labeled samples [2,7–9]. These can be used to update facial models after enrollment, as new reference data becomes available, allowing to maintain or increase matching performance. Adaptive multiple classifier systems (MCS) have been successfully applied for FRiVS [2,10]. In these systems, the facial model of each individual is encoded using an ensemble of 2-class classifiers or detectors (EoD), trained to discriminate between samples of a target individual and non-target individuals.

An issue with the supervised update of classifiers is the analysis and extraction of labeled reference samples from operational videos. A domain expert must isolate target faces manually or semi-automatically in video surveillance footage, which involves undesirable costs and delays. Instead of relying on a human expert, the system may self-update face models with operational videos. Several semi-supervised learning approaches have been proposed to update biometric models using a combination of labeled and unlabeled samples [11–13]. In the area of adaptive biometrics, two representative approaches for semi-supervised learning are the self-update and co-update techniques [14]. The first applies an update threshold (higher than the detection threshold) to each matching scores to select input biometric samples as new templates, and the second seeks corroboration of scores from two or more matchers for cross-updating.

To the authors' knowledge, a FR system that allows for self-updating facial models in video surveillance applications has not been proposed in literature. An issue encountered with self-updating is the reliable selection of operational samples from the target individual to adapt facial models. A high level of confidence is required to avoid updating models with non-target data. In contrast, a facial model should also be adapted with a diversified set of reference samples to improve the generalization performance. Given an adaptive MCS proposed in [2,10], information from a face tracker and individual-specific ensembles may be integrated to provide a variety of high confidence reference samples.

In video surveillance, an abundance of reference samples may be extracted from non-target facial trajectories acquired in the scene during routine system operation. Two databases may be formed with samples extracted (1) from trajectories of other individuals of interest besides the target individual (known as the cohort model, CM), and (2) from unknown people appearing in scene (known as the universal model, UM) [1,2,10,15]. This imposes the need to sub-sample non-target data in order to design accurate facial models, using an ensemble of 2-class classifiers. Moreover, adaptive MCSs require reference data to be stored in memory for validation [2,9]. Practical memory limitations impose the need for a method to rank and select the most relevant validation samples for each individual (EoD).

In this paper, an adaptive MCS is proposed for video-to-video FR in semi- and unconstrained video surveillance environments. Within the adaptive MCS, an EoD encodes and updates the facial model of each individual of interest. This novel system allows for spatio-temporal recognition and self-update of facial models based on high-confidence trajectories. During operations, a face tracker defines facial trajectories for different individuals that appear in a video. Track ID numbers are integrated with predictions of individual-specific ensembles at a decision-level for enhanced video-to-video FR. The proposed system relies on tracker quality to regroup ROIs into facial trajectories, and applies a double thresholding scheme to curves produced by accumulating positive EoD predictions for a trajectory. An individual of interest is recognized if the number of positive predictions accumulated over some time window of a trajectory surpass a *detection* threshold for an EoD.

A second (higher) *update* threshold is applied to select high-confidence trajectories that are suitable for self-updating a facial model. If the number of positive predictions surpasses this threshold for an EoD, then all samples extracted from the target ROIs of the trajectory are combined with non-target samples (selected from the CM and UM) to update the corresponding face model. Since a trajectory may contain target ROIs that were incorrectly classified by the EoD, facial models are adapted with a diversified set of reference samples that may refine the decision boundary between target and non-target distributions, and thereby improve the generalization performance. A sub-sampling technique based on condensed nearest neighbor (CNN) [16] is employed to select non-target samples along this boundary. The data for EoD update is comprised of diverse facial regions associated with target and non-target trajectories, and is employed to generate a new pool of 2-class classifiers, and to update the fusion function of the user specific EoD. To avoid issues related to knowledge corruption in incremental learning classification systems, the self-update of EoD employs a learn-and-combine strategy [2]. Finally, a long term memory (LTM) is maintained over time with a fixed number of reference validation samples per individual. A memory management strategy based on the Kullback–Leibler (KL) divergence criteria [17] is proposed to rank and select the most relevant target and non-target reference samples. This criteria seeks to preserve the highest relative entropy of ensemble over time. In other words, the KL divergence becomes higher for samples that contain a higher level

[1] A *facial trajectory* is defined as a set of ROIs (isolated through face detection) that correspond to a same high quality track of an individual across consecutive frames.