Information Sciences 286 (2014) 75–101

Contents lists available at ScienceDirect

Information Sciences

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

# Adaptive ensembles for face recognition in changing video surveillance environments

C. Pagano [a,*], E. Granger [a], R. Sabourin [a], G.L. Marcialis [b], F. Roli [b]

[a] Lab. d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montreal, Canada
[b] Pattern Recognition and Applications Group, Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy

ARTICLE INFO

*Article history:*
Received 11 October 2013
Received in revised form 26 May 2014
Accepted 6 July 2014
Available online 18 July 2014

*Keywords:*
Multi-classifier system
Incremental learning
Adaptive biometric
Change detection
Face recognition
Video surveillance

ABSTRACT

Recognizing faces corresponding to target individuals remains a challenging problem in video surveillance. Face recognition (FR) systems are exposed to videos captured under various operating conditions, and, since data distributions change over time, face captures diverge w.r.t. stored facial models. Although these models may be adapted when new reference videos become available, incremental learning with faces captured under different conditions may lead to knowledge corruption. This paper presents an adaptive multi-classifier system (AMCS) for video-to-video FR in changing surveillance environments. During enrolment, faces captured in reference videos are employed to design an individual-specific classifier. During operations, a tracker allows to regroup facial captures for individuals in the scene, and accumulate the predictions per track for robust spatiotemporal FR. Given a new reference video, the corresponding facial model is adapted according to the type of concept change. If a gradual pattern of change is detected, the individual-specific classifier(s) are adapted through incremental learning. To preserve knowledge, another classifier is learned and combined with the individuals previously-trained classifier(s) if an abrupt change is detected. For proof-of-concept, the performance of a particular implementation of this AMCS is assessed using videos from the Faces in Action dataset. By adapting facial models according to changes detected in new reference videos, this AMCS allows to sustain a high level of accuracy comparable to the same system that is always updated using a learn-and-combine approach, while reducing time and memory complexity. It also provides higher accuracy than incremental learning classifiers that suffer the effects of knowledge corruption.



## 1. Introduction

The global market for video surveillance (VS) technologies has reached revenues in the billions of $US as traditional analogue technologies are replaced by IP-based digital ones. VS networks are comprised of a growing number of cameras, and transmit or archive massive quantities of data for reliable decision support. The ability to automatically recognize and track individuals of interest across these networks, and under a wide variety of operating conditions, may provide enhanced screening and situation analysis.

* Corresponding author.
*E-mail addresses:* cpagano@livia.etsmtl.ca (C. Pagano), eric.granger@etsmtl.ca (E. Granger), robert.sabourin@livia.etsmtl.ca (R. Sabourin), marcialis@diee.unica.it (G.L. Marcialis), roli@diee.unica.it (F. Roli).

http://dx.doi.org/10.1016/j.ins.2014.07.005
0020-0255/

In decision support systems for VS, face recognition (FR) has become an important function in two types of applications. In *watch-list screening applications*, facial models[1] used for classification are designed using regions of interest (ROIs) extracted from the reference still images or mugshots of a watch-list. Then, still-to-video FR seeks to determine if faces captured in video feeds correspond to an individual of interest. In *person re-identification* for search and retrieval applications, facial models are designed using ROIs extracted from reference videos and tagged by a human operator. Then, video-to-video FR seeks to alert the operator when these individuals appear in either live (real-time monitoring) or archived (post-event analysis) videos.

This paper focuses on the design of robust classification systems for video-to-video FR in changing surveillance environments, as required in person re-identification. In this context, public security organizations have deployed many CCTV and IP surveillance cameras in recent years, but FR performance is limited by human recognition abilities. Indeed, accurate and timely recognition of ROIs is challenging under semi-controlled (e.g., in an inspection lane, portal or checkpoint entry) and uncontrolled (e.g., in cluttered free-flow scene at an airport or casino) capture conditions. Given the limited control during capture, the performance of state-of-the-art systems are affected by the variations of pose, scale, orientation, expression, illumination, blur, occlusion and ageing. Moreover, FRiVS is an open set problem, where only a small proportion of the faces captured during operations correspond to individuals of interest. Finally, ROIs captured in videos are matched against facial models designed a priori, using a limited number of high quality reference samples captured during enrolment. Accuracy of face classification is highly dependent on the representativeness of models, and thus number, relevance and diversity of these samples.

Some specialized classification architectures have been proposed for FRiVS. For instance, the open-set Transduction Confidence Machine-*k*NN (TCM-*k*NN) is comprised of a global multi-class classifier with a rejection option tailored for unknown individuals [37]. Classification systems for FRiVS should however be modeled as independent individual-specific detection problems, each one implemented using one- or two-class classifiers (i.e., detectors), with specialized thresholds applied to their output scores [48]. The advantages of class-modular architectures in FRiVS (and biometrics in general) include the ease with which face models (or classes) may be added, updated and removed from the systems, and the possibility of specializing feature subsets and decision thresholds to each specific individual. Individual-specific detectors have been shown to outperform global classifiers in applications where the reference design data is limited w.r.t. the complexity of underlying class distributions and to the number of features and classes [45,54]. Moreover, some authors have argued that biometric recognition is in essence a multi-classifier problem, and that biometric systems should co-jointly solve several classification tasks in order to achieve state-of-the-art performance [5].

Modular architectures for FRiVS have been proposed by Ekenel et al. [19], where 2-class individual-specific Support Vector Machines are trained on a mixture of target and non-target samples. Given the limited amount of reference samples and the complexity of environments, modular approaches have been extended by assigning a classifier ensemble to each individual. For example, Pagano et al. [48] proposed a system comprised of an ensemble of 2-class ARTMAP classifiers per individual, each one designed using target and non-target samples. A pool of diversified classifiers is generated using an incremental learning strategy based on dynamic PSO, and combined in the ROC space using a Boolean fusion function.

In person re-identification, new reference video become available during operations or through some re-enrolment process, and an operator can extract a set of facial ROIs belonging to a target individual. In order to adapt an individual's facial model in response to these new ROI samples, the parameters of a individual-specific classifier can be re-estimated through supervised incremental learning. For example, ARTMAP neural networks [9] and extended Support Vector Machines [52] have been designed or modified to perform incremental learning. However, these classifiers are typically designed under the assumption that data is sampled from a static environment, where class distributions remain unchanged over time [25].

Under semi- and uncontrolled capture conditions, ROI samples that are extracted from new reference videos may incorporate various patterns of change that reflect varying concepts.[2] While gradual patterns of change in operational conditions are often observed (due to, e.g., ageing over sessions), abrupt and recurring patterns (caused by, e.g., new pose angle versus camera) also occur in FRiVS. A key issue in changing VS environments is adapting facial models to assimilate samples from new concepts without corrupting previously-learned knowledge, which raises the *plasticity–stability* dilemma [26]. Although updating a single classifier may translate to low system complexity, incremental learning of ROI samples extracted from videos that reflect significantly different concepts can corrupt the previously acquired knowledge [13,50]. Incomplete design data and changing distributions contribute to a growing divergence between the facial model and the underlying class distribution of an individual.

Adaptive ensemble methods allow to exploit multiple and diverse views of an environment, and have been successfully applied in cases where concepts change in time. By assigning an adaptive ensemble to each individual, it is possible to adapt a facial model by updating the pool of classifiers and/or the fusion function [33]. For example, with iques like Learn++ [50] and other Boosting variants [47], a classifier is trained independently using new samples, and weighted such that accuracy is maximized. Other approaches discard classifiers when they become inaccurate or concept change is detected, while maintaining a pool with these classifiers allows to handle recurrent change. Classifier ensembles are well suited for adaptation in changing environments since they can manage the *plasticity–stability* dilemma at the classifier level – when samples are

[1] A *facial model* is defined as either a set of one or more reference captures (used in template matching systems), or a statistical model estimated through training with reference captures (used in neural or statistical classification systems) corresponding to a target individual.

[2] A *concept* can be defined as the underlying class distribution of data captured under specific condition, in our context due to different pose angle, illumination, scale, etc. [43].