Adaptive Classification for Person Re-Identification Driven by Change Detection

C. Pagano[1], E. Granger[1], R. Sabourin[1], G. L. Marcialis[2] and F. Roli[2]
[1]*Lab. d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Université du Québec, Montreal, Canada*
[2]*Pattern Recognition and Applications Group, Dept. of Electrical and Electronic Engineering, University of Cagliari, Cagliari, Italy*
cpagano@livia.etsmtl.ca, {eric.granger, robert.sabourin}@etsmtl.ca, {marcialis,roli}@diee.unica.it

Keywords: Multi-Classifier Systems, Incremental Learning, Adaptive Biometrics, Change Detection, Face Recognition, Video Surveillance.

Abstract: Person re-identification from facial captures remains a challenging problem in video surveillance, in large part due to variations in capture conditions over time. The facial model of a target individual is typically designed during an enrolment phase, using a limited number of reference samples, and may be adapted as new reference videos become available. However incremental learning of classifiers in changing capture conditions may lead to knowledge corruption. This paper presents an active framework for an adaptive multi-classifier system for video-to-video face recognition in changing surveillance environments. To estimate a facial model during the enrolment of an individual, facial captures extracted from a reference video are employed to train an individual-specific incremental classifier. To sustain a high level of performance over time, a facial model is adapted in response to new reference videos according the type of concept change. If the system detects that the facial captures of an individual incorporate a gradual pattern of change, the corresponding classifier(s) are adapted through incremental learning. In contrast, to avoid knowledge corruption, if an abrupt pattern of change is detected, a new classifier is trained on the new video data, and combined with the individual's previously-trained classifiers. For validation, a specific implementation is proposed, with ARTMAP classifiers updated using an incremental learning strategy based on Particle Swarm Optimization, and the Hellinger Drift Detection Method is used for change detection. Simulation results produced with Faces in Action video data indicate that the proposed system allows for scalable architectures that maintains a significantly higher level of accuracy over time than a reference passive system and an adaptive Transduction Confidence Machine-kNN classifier, while controlling computational complexity.

1 INTRODUCTION

Face recognition (FR) has become an important function in several types of video surveillance (VS) applications. For instance, in *watch-list screening*, FR systems seek to determine if a target face captured in video streams corresponds to an individual of interest in a watchlist. In *person re-identification*, a FR system seek to alert a human operator as to the presence of individuals of interest appearing in either live (real-time monitoring) or archived (post-event analysis) video streams. These applications rely on the design of a representative facial model[1] to perform template matching or classification. *Watch-list screening* uses one or more regions of interest (ROIs) extracted from reference still images or mugshots, while in *person re-identification* ROIs are extracted from reference videos and tagged by a human operator.

This paper focuses on the design of robust face classification systems for video-to-video FR in changing surveillance environments, as required in *person re-identification* or *search and retrieval*. For example, in such applications, the operator can isolate a facial trajectory[2] for an individual over a network of cameras, and enrol a face model to the system. Then, during operations, facial regions captured in

[1]A facial model is defined as either a set of one or more reference face captures (used for template matching), or a statistical model (used for classification).

[2]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.

live or archived video streams are matched against facial models of target individuals of interest to be followed. It is assumed that holistic facial models are estimated by training a neural network or statistical classifier on reference ROI patterns extracted from operational videos using a face detector. In this context, the performance of state-of-the-art commercial and academic systems is limited by the difficulty in capturing high quality facial regions from video streams under semi-controlled (e.g., at inspection lanes, portals and checkpoint entries) and uncontrolled (e.g., in cluttered free-flow scenes at airports or casinos) capture conditions. Performance is severely affected by the variations in pose, scale, orientation, expression, illumination, blur, occlusion and ageing.

More precisely, given a face classifier, the various conditions under which a face can be captured by video cameras are representative of different *concepts*, i.e. different data distributions in the input feature space. These concepts contribute to the diversity of an individual's face model, and underlying class distributions are composed by information from all possible capture conditions (e.g. pose orientations and facial expressions that could be encountered during operations).

However, in practice, ROIs extracted from videos are matched against facial models designed a priori, using a limited number of reference captures collected during enrolment. Incomplete design data and changing distributions contribute to a growing divergence between the facial model and the underlying class distribution of an individual. In person re-identification applications, reference video containing an individual of interest may become available during operations or through some re-enrolment process. Under semi or uncontrolled capture conditions, the corresponding ROIs may be sampled from various concepts (e.g., with different facial orientation), but the presence of all the possible concepts inside a single reference sequence cannot be guaranteed. For this reason, a system for video-to-video FR should be able to assimilate new reference trajectories over the time (as they become available) in order to add newly available concepts to the individuals' facial models, as they may be relevant to perform FR under future observation conditions. Therefore, adapting facial models to assimilate new concepts without corrupting previously-learned knowledge is an important feature for FR in changing real-world VS environments.

In this paper, an active framework for an adaptive multi-classifier system is proposed for video-to-video FR as seen in person re-identification applications. It maintains a high level of performance in changing VS environments by adapting its face models to concepts emerging in new reference videos, without corrupting the previously acquired knowledge. A specific implementation is proposed using, for each target individual enrolled to the system, a pool of two-class incremental ARTMAP neural network classifiers (Carpenter et al., 1992) optimized using an incremental learning strategy based on Dynamic Niching PSO (DNPSO) (Nickabadi et al., 2008; Connolly et al., 2012). Pools are combined using the weighted-average score-level fusion. When a new reference trajectory becomes available for enrolment or adaptation of an individual's face model, a change detection mechanism based on Hellinger histogram distances (Ditzler and Polikar, 2011) evaluates whether the corresponding ROI patterns exhibit gradual or abrupt changes w.r.t. the previously-learned knowledge. If the new reference samples exhibit gradual changes w.r.t. a previously-stored reference distribution, the corresponding classifier is updated using the DNPSO-based learning strategy. If the new reference samples present significant (or abrupt) changes compared to all the previously-stored distributions, a new reference distribution is stored. A new classifier is then trained on the new ROI patterns and combined with the individual's previously learned classifiers at the score level.

The accuracy and resource requirements of the proposed approach are compared to a passive version (incremental only) of the framework, as well as an adaptive version of a Transduction Confidence Machine-kNN (TCM-kNN) system (Li and Wechsler, 2005), using ROIs extracted from real-world video surveillance streams of the publicly-available Faces in Action database (Goh et al., 2005). It is composed of over 200 individuals captured over 3 sessions (several months), and exhibits both gradual (e.g. expression, ageing) and abrupt (e.g. orientation, illumination) changes. A person re-identification scenario is considered, where an analyst can label ROIs captured in operational videos, and provide new sets of reference ROI patterns for adaptation. Each new set can incorporate a different concept, for example a different facial pose or illumination, and the system may encounter ROIs from every possible concept during its operation.

2 VIDEO-TO-VIDEO FACE RECOGNITION

Many video FR techniques have been proposed in the literature, relying on both spatial and temporal information to perform recognition (Zhou et al., 2006; Barry and Granger, 2007; Matta and Dugelay, 2009). However, only a small subset is suitable for video-to-video FR in video surveillance applications (C. Pagano, E. Granger, R. Sabourin, 2012). For example, research by Connolly et al. (Connolly et al., 2012) are focused on N-class classifiers for video FR in access control applications. In addition, some specialized classification architecture have been proposed for an *open-set* recognition environment, such as FR in VS. Among them, the open-set TCM-kNN is a global multi-class classifier employed with a specialized rejection option for unknown individuals (Li and Wechsler, 2005).

This research focuses on modular systems designed with individual-specific detectors (one or two-class classifiers). In fact, individual-specific detectors have been shown to outperform global classifiers in applications where the design data is limited w.r.t. the complexity of underlying class distributions and to the number of features and classes (Oh and Suen, 2002). For example, Tax and Duin (Tax and Duin, 2008) proposed a heuristic to combine one-class classifiers for solving multi-class problems, where rejection thresholds are class-dependent. Given the limited amount of reference patterns and the complexity of environments, class-modular approaches have been extended to improve classification performance, by assigning a classifier ensemble to each individual. Pagano et al. (C. Pagano, E. Granger, R. Sabourin, 2012) proposed a system for FR in VS comprised of an ensemble of 2-class ARTMAP classifiers per individual, each one designed using target and non-target patterns. In addition to the performance improvement, the advantages of class-modular architectures in FR in VS (and biometrics in general) include the ease with which biometric models of individuals (classes) may be added, updated and removed from the systems, and the possibility of specializing feature subsets and decision thresholds to each specific individual.

To integrate new reference data, several adaptive methods have been proposed in the literature, which can be differentiated by the level of the adaptation. While incremental classifiers (like ARTMAP (Carpenter et al., 1992) and self-organizing (Fritzke, 1996) neural networks), adapt their internal parameters in response to new data, ensembles of classifiers (EoC) allow for two levels of adaptation, updating the internal parameters of a swarm of classifiers, and/or the selection and fusion function (Kuncheva, 2004). Updating a single classifier can translate to low system complexity, but incremental learning of ROI patterns extracted from videos that represent significantly different concepts can corrupt the previously acquired knowledge (Connolly et al., 2012; Polikar and Upda, 2001). On the other hand, classifier ensembles are well suited to prevent knowledge corruption, as previously acquired knowledge can be preserved by training a new classifier on the new data. However, the benefits of EoC (accuracy and robustness) are achieved at the expense of system complexity (the number of classifiers grows). The time required for face classification grows with the number of classifiers, and the structure of ROI pattern distributions. The trade off between accuracy and complexity is critical in VS applications, as the recognition may be performed in real time.

More recently, active approaches for adaptive classification have been proposed in the literature, exploiting a change detection mechanism to drive on-line learning, such as the diversity for dealing with drifts algorithm (Minku and Yao, 2012) and the Just-in-Time architecture that regroups reference templates per concept (Alippi et al., 2013). However these approaches have been developed for on-line learning, where the goal is to adapt to the concept currently observed by the system. Their adaptation focuses on the more recent concepts, through weighting or by discarding of previously-learned concepts, which may degrade system performance w.r.t. other concepts.

Although relevant to video-to-video face recognition due to their open-set nature and ability to adapt to new data, these methods are not designed for a re-identification scenario. They either increase the system's complexity with each newly available reference sequence, or consider a single operational concept at the expense of the previously-acquired knowledge. In this paper, a new framework is proposed to perform active adaptation, allowing to refine facial models of individuals over time using new reference trajectories without corrupting the previously acquired knowledge, and controlling the system's growth. Depending on the detected pattern of change, it relies on a hybrid updating strategy that dynamically adapts an ensemble of classifiers on the three possible levels: the ensemble (adding new classifiers), the classifiers (adapting their internal parameters), and the decision.

3 CONCEPT CHANGE AND FACE RECOGNITION

In this paper, a mechanism is considered to detect changes in the underlying data distribution, as can be observed in new sets of reference ROI patterns provided by an operator in face re-identification applications. This mechanism triggers different updating strategies depending on the nature of concepts ob-

Type of change	Examples in video-to-video FR
1) random noise	– inherent noise of system (camera, matcher, etc.)
2) gradual changes	– ageing of user over time
3) abrupt changes	– new unseen capture conditions (e.g. new pose angle, scale, etc.)
4) recurring contexts	– unpredictable but recurring changes in capture conditions (e.g. daily variations in artificial or natural illumination.)

Table 1: Types of changes occurring in video surveillance environments.

served by the system in these sequences. This section illustrates the relation between the abstract notion of concepts and the real-world recognition problem - the actual facial captures.

A *concept* can be defined as the underlying data distribution of the problem at some point in time (Narasimhamurthy and Kuncheva, 2007), and a *concept change* encompasses various types of noise, trends and substitutions in the underlying data distribution associated with a class or concept. A categorization of changes has been proposed by Minku et al. (Minku et al., 2010), based on severity, speed, predictability and number of re-occurrences, but the following four categories are mainly considered in the literature: noise, abrupt changes, gradual changes and recurring changes (Kuncheva, 2008).

In the context of video-to-video FR, a concept is related to a specific capture condition of physiological characteristic, and concept changes originate from variations in those capture conditions and/or individuals' physiology, which have yet to be integrated into the system's facial models. As shown in Table 1, they may range from minor random fluctuations or noise, to sudden abrupt changes of the underlying data distribution, and are not mutually exclusive in real-word surveillance environments. In this paper, video-to-video FR is performed under semi- and uncontrolled capture conditions, and concept changes are observed in new reference ROI patterns. The refinement of previously-observed concepts (e.g., new reference ROIs are captured for previously seen pose angles), corresponds to gradual changes, and data corresponding to newly-observed concepts (e.g., new ROIs are captured under previously unseen illumination conditions, or pose angles), corresponds to abrupt changes. A new concept can also correspond to a recurring change as specific observation conditions may be re-encountered in the future (e.g., faces captured under natural vs. artificial lighting).

In proof of concept simulations, the system proposed in Section 4 processed ROI patterns from the Faces in Action (FIA) database (Goh et al., 2005). It contains reference videos captured over 3 sessions, and using camera for 0° and ±72.6° pose angles.

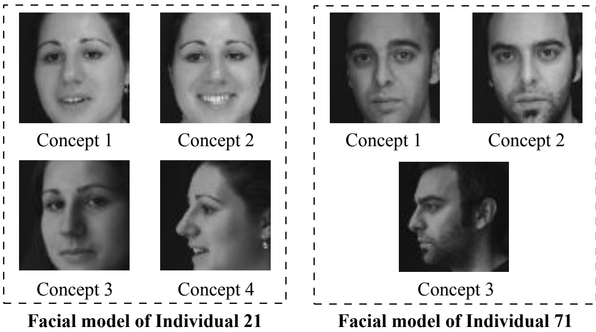


Figure 1: The most representative reference ROIs of different concepts detected by the proposed system for individuals 21 and 71 of the Faces in Action database.

Changes in the reference ROI patterns have been detected for each individual of interest, and the corresponding concepts have been integrated into the system. Fig. 1 shows the most representative ROIs of the different concepts detected for individuals 21 and 71 (the smallest Hellinger distance between an ROI pattern and the histogram representation of the concept by the system). Note that the system detected 4 different concepts for individual 21, corresponding respectively to: 2 frontal orientations with different facial expressions, and 2 different profile views. In the same way, 3 concepts have been detected for individual 71: 2 frontal orientations with different facial hair, and a profile view. This illustrates the relation between concepts detected by the system in the feature space, and the capture conditions of the ROIs - these concepts correspond to different observation conditions encountered in ROIs from reference videos.

4 AN ADAPTIVE MULTI-CLASSIFIER SYSTEM FOR VIDEO-TO-VIDEO FR

Figure 2 presents an active framework for an adaptive multi-classifier system (AMCS) with change detection and weighting that is specialized for video-to-video FR in changing environments, as seen in person

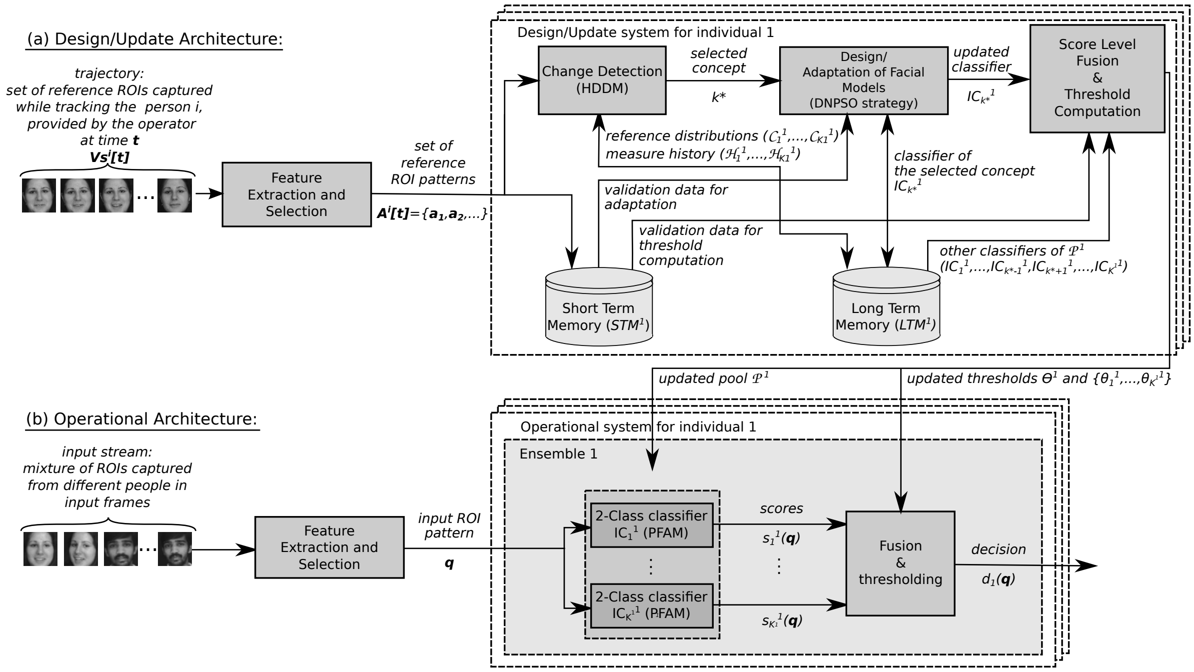


Figure 2: Architecture of the proposed $AMCS_w$ for video-to-video FR in changing environments. The design and update architecture for each individual of interest i is presented in (a), and the operational architecture (for all I individuals) in (b).

re-identification applications. In this figure, the reference trajectories are presented as sets of ROIs for simplification purposes, but the system can incorporate a segmentation module prior to the feature extraction and selection one to automatically extract ROIs from a reference sequence.

Depending on the nature of ROI patterns extracted from new reference videos, the proposed system relies on three different levels of adaptation to maintain the level of accuracy: (1) internal parameters of the classifiers are updated through incremental learning of data from already known concepts, (2) new classifiers are added to assimilate new concepts, and (3), the fusion of classifiers is updated. This hybrid approach allows to preserve past knowledge of concepts, as classifiers are only updated incrementally with ROI patterns from similar concepts, otherwise new classifiers are trained. This mechanism controls the growth of the system, as new classifiers are only added when necessary, i.e. when a set of significantly different ROI pattern is presented to the system.

In this paper, a specific implementation of the proposed weighted AMCS framework (called $AMCS_w$) is presented using probabilistic fuzzy-ARTMAP (PFAM) (Lim and Harrison, 1995) classifiers. PFAM classifiers are incremental learning neural-networks known to provide a high level of accuracy with moderate time and memory complexity (Lim and Harrison, 1995). They rely on an unsupervised categorization of the feature space into hyper-rectangles associated to output classes through a MAP field, which is then modelled as mixtures of Gaussian distributions to provide probabilistic prediction scores instead of binary decisions. These classifiers are optimized with a DNPSO algorithm (Nickabadi et al., 2008), as this updating strategy has already been successfully applied to FR in video in (Connolly et al., 2012). More precisely, DNPSO is a dynamic population based stochastic optimization technique inspired by the behaviour of a flock of birds (Eberhart and Kennedy, 1995), which is used to determine optimal sets of hyper-parameters $\mathbf{h} = (\alpha, \beta, \varepsilon, \bar{\rho}, r)$ of PFAM classifiers w.r.t. validation data.

In addition, following the recommendations in (Kittler and Alkoot, 2003) on the fusion of correlated classifiers, an average score-level fusion rule is considered for the ensembles of PFAM classifiers. More precisely, to filter out ambiguities, the average is weighted to favour scores that are highest w.r.t. their threshold: for an individual i with a concept-specific threshold θ_k^i (determined with validation data for concept k), each score $s_k^i(\mathbf{q})$ is weighted by ω_k^i, defined by the confidence measure:

$$\omega_k^i = \max\{0, (s_k^i(\mathbf{q}) - \theta_k^i)\} \quad (1)$$

This weight reflects the quality of the input pattern $\mathbf{q}$ in reference to concept k. Finally, for change detection, the Hellinger Drift Detection Method (HDDM) presented in (Ditzler and Polikar, 2011) has been chosen for its low computational and memory costs.

For each enrolled individual $i = 1,...,I$, this modular system is composed by a pool of K^i two-class PFAM classifiers $\mathcal{P}^i = \{IC_1^i,...,IC_{K^i}^i\}$, $K^i \geq 1$ being the number of concepts detected in the individual's reference ROI pattern sets. Decisions are produced using classifier-specific (concept) thresholds $\{\theta_1^i,...,\theta_{K^i}^i\}$, and a global user-specific threshold Θ^i. The supervised learning of new reference ROI pattern sets by the 2-class PFAM classifiers is handled using the DNPSO-training strategy presented in (Connolly et al., 2012). $AMCS_w$ is an active system, where the adaptation strategy is guided by change detection, using HDDM (Ditzler and Polikar, 2011). In order to compare a new set of reference ROI patterns to all the K_i previously-encountered concepts, histogram representations $\{\mathcal{C}_1^i,...,\mathcal{C}_{K_i}^i\}$ are stored into a long-term memory LTM^i. In addition, a short term memory STM^i is used to store reference data for design or adaptation and for validation.

The class-modular architecture of $AMCS_w$ allows to design and update facial models independently for each individual of interest i, according to Alg. 1 and Fig. 2a. When a new set of reference ROIs $Vs^i[t]$ is provided by the operator at time t, relevant features are first extracted and selected from each ROI in order to produce the corresponding set of ROI patterns $\mathbf{A}^i[t]$. STM^i temporarily stores validation data used for classifier design and threshold selection. The change detection process assess whether the underlying data distribution exhibits significant changes compared to previously-learned data. For this purpose, the system compares previously-observed concepts $\{\mathcal{C}_1^i,...,\mathcal{C}_{K^i}^i\}$ stored in LTM^i and $\mathbf{A}^i[t]$ using the Hellinger distance, following:

$$\delta_k^i[t] = \frac{1}{D}\sum_{d=1}^{D}\sqrt{\sum_{b=1}^{B}\left(\sqrt{\frac{\mathbf{A}(b,d)}{\sum_{b'=1}^{B}\mathbf{A}(b',d)}}-\sqrt{\frac{\mathcal{C}_k^i(b,d)}{\sum_{b'=1}^{b}\mathcal{C}_k^i(b',d)}}\right)^2} \quad (2)$$

where D is the dimensionality of the feature space, B the number of bins in $\mathbf{A}$ and $\mathcal{C}_k^i$, and $\mathbf{A}(b,d)$ and $\mathcal{C}_k^i(b,d)$ the frequency count in bin b of feature d.

If a significant (abrupt) change is detected between $\mathbf{A}^i[t]$ and all the stored concept models, or if $Vs^i[t]$ is the first reference sequence for the individual (no previous concept has been stored), a new concept is assumed. More precisely, an abrupt change between $\mathcal{C}_k^i$ and $\mathbf{A}^i[t]$ is detected if $\delta_k^i[t] > \beta_k^i[t]$, with $\beta_k^i[t]$ an adaptive threshold computed from the previous distance measures following:

$$\beta_k^i[t] = \hat{\delta}_k^i + t_{\alpha/2}.\frac{\hat{\sigma}}{\sqrt{\Delta_t}} \quad (3)$$

Algorithm 1 Strategy to design and adapt the facial model of individual i with the proposed $AMCS_w$

Input: Set of new reference ROIs for individual i provided by the operator at time t, $Vs^i[t]$
Output: Updated classifier pool $\mathcal{P}^i$ ($K^i = 1$ or $K^i > 1$)

Perform feature extraction and selection on $Vs^i[t]$ to obtain a set of ROI patterns $\mathbf{A}^i[t]$
$STM^i \leftarrow \mathbf{A}^i[t]$
for each concept $k = 1$ to K^i **do**
  Measure $\delta_k^i[t]$ the distance between $\mathbf{A}^i[t]$ and the concept representation $\mathcal{C}_k^i$ using Hellinger distance
  Compare $\delta_k^i[t]$ to the change detection threshold $\beta_k^i[t]$ of the concept k
end for
if $\delta_k^i[t] > \beta_k^i[t]$ for each concept $k \in [1,K_i]$, or $K_i = 0$ **then** {Abrupt change or 1^{st} concept}
  $K^i \leftarrow K^i + 1$
  Set index of the chosen concept $k^* \leftarrow K^i$
  Generate the concept representation $\mathcal{C}_{K^i}^i$ from $\mathbf{A}^i[t]$ and store it into LTM^i
  Initiate a DNPSO-learning strategy using data from STM^i, to obtain the best classifier $IC_{K^i}^i$
  Update $\mathcal{P}^i \leftarrow \{\mathcal{P}^i, IC_{K^i}^i\}$
else {Gradual change}
  Determine the index of the closest concept $k^* = min\{\delta_k^i[t] : k = 1,...,K^i\}$
  Re-initiate a DNPSO-learning strategy using data from STM^i, to obtain the updated best classifier IC_{k^*}
end if
for each concept $k = 1$ to K^i **do**
  Compute the classifier specific threshold θ_k^i using data from STM^i {see Section 5.3}
end for
Compute the user specific threshold Θ^i using data from STM^i {see Section 5.3}

where α is the confidence interval of the t-statistic test, Δ_t the total amount of past distance measures, and $\hat{\delta}_k^i$ and $\hat{\sigma}$ their average and variance. In this case, K^i is incremented, and a new incremental classifier $IC_{K^i}^i$ is designed for the concept (IC_1^i if the first concept) using the training and adaptation module with the data from STM^i. When a moderate (gradual) change is detected, the classifier $IC_{k^*}^i$ corresponding to the closest concept representation $\mathcal{C}_{k^*}^i$ is updated and evolved through incremental learning.

Finally, if several concepts are stored in the system, $\mathcal{P}^i$ is updated to combine the most accurate classifiers of the known concepts: if a new concept has been detected, a new classifier $IC_{K^i}^i$ is added to $\mathcal{P}^i$, and if a known concept k^* is updated, the corresponding classifier $IC_{k^*}^i$ is updated. If only one concept has been detected, a single classifier is assigned to the individual, $\mathcal{P}^i = IC_1^i$. The fusion of classifiers is performed at score level, using a weighted average to favour scores that are highest w.r.t. their threshold. For this purpose, classifier specific thresholds θ_k^i are determined with validation data for concept k, and a user specific threshold Θ^i is also computed.

During operations, when the $AMCS_w$ is not designing or updating facial models, it functions according to the architecture shown in Fig. 2b. The system extracts a pattern $\mathbf{q}$ in response to input ROI from face detection. Then, an overall score is computed for each individual pool $\mathcal{P}^i$ through fusion of PFAM classifiers' scores $s_k^i(\mathbf{q})$ ($k = 1,...,K^i$), using weighted average fusion. Each score $s_k^i(\mathbf{q})$ is multiplied by the weight ω_k^i computed following Eq. 1. The weighted average $\sum_{k=1}^{K^i} \omega_k^i.s_k^i$ is then compared to the class specific threshold Θ^i to produce the overall decision $d^i(\mathbf{q})$.

5 EXPERIMENTAL METHODOLOGY

5.1 Video Database

The Carnegie Mellon University Faces In Action (FIA) face database (Goh et al., 2005) has been used to evaluate the performance of the proposed system. It is composed of 20-second videos capturing the faces of 221 participants in both indoor and outdoor scenario, each video mimicking a passport checking scenario. Videos have been captured at three different horizontal pose angles (0° and ±72.6°), each one with two different focal length (4 and 8mm). For the experiments of this paper, all ROIs have been segmented from each frame, using the OpenCV v2.0 implementation of the Viola-Jones algorithm (Viola and Jones, 2004), and the faces have been rotated to align the eyes (to minimize intra-class variations (Gorodnichy, 2005)). ROIs have been scaled to a common size of 70x70 pixels, which was the smallest detected ROI. Features have finally been extracted with the Multi-Bloc Local Binary Pattern (LBP) (Ahonen, 2006) algorithm features for block sizes of 3x3, 5x5 and 9x9 pixels, concatenated with the grayscale pixel intensity values, and reduced to $D = 32$ features using Principal Component Analysis. The dimensionality of the final feature space has been determined through preliminary experiments, $D = 32$ being the smallest dimensionality that could be performed without reducing classification performance.

The FIA videos have been separated into 6 subsets, according to the different cameras (left, right and frontal face angle, with 2 different focal length, 4 and 8 mm) for each one of the 3 sessions, and for each individual. Only indoors videos for the the frontal angle (0°) and left angle (±72.6°) are considered for experiments in this paper.

5.2 Simulation Scenario

Ten (10) individuals of interests have been selected as target individuals, subject to two experimental constraints: 1) they appear in all 3 sessions, and 2), at least 30 ROIs for every frontal and left videos have been detected by the OpenCV segmentation. The ROIs of the remaining 200 individuals are mixed into a Universal Model (UM), to provide classifiers with non-target samples. Only 100 of those individuals have been randomly selected for the training UM, to ensure that the scenario contains unknown individuals in testing (i.e.the remaining 100 whose samples have never been presented to the system during training).

To avoid bias due to the more numerous ROI samples detected from the frontal sessions, the original FIA frontal sets have been separated into two subsets, forming a total of 9 sets of reference ROI patterns for design and update (see Table 2). Simulations emulate the actions of a security analyst in a decision support system that provides the systems with new reference ROI pattern sets. The reference sets $Vs^i[t]$ are presented to update the face models of individuals $i = 1,..,10$ at a discrete time $t = 1,2,...,9$.

Reference sets used for design are populated using the ROI patterns from the same individual, from the cameras with 8-mm focal length in order to provide ROI patterns with better quality. ROIs captured during 3 different sessions and 2 different pose angles may be sampled from different concepts, and the transition from sequence 6 to 7 (change of camera angle) represents most abrupt concept change in the reference ROI patterns. Changes observed from one session to another, such as from sequences 2 to 3, 4 to 5, 7 to 8 and 8 to 9 depends on the individual. As faces are captured over intervals of several months, both gradual and abrupt changes may be detected.

For each time step $t = 1,2,...,9$, the systems are evaluated after adaptation on the same test dataset, emulating a practical security checkpoint station where different individuals arrive one after the other. The test dataset is composed by ROI patterns from every session and pose angle to simulate face re-identification applications where different concepts may be observed during operations, but where the analyst gradually tags and submits new ROI patterns to the system to adapt face models. Every different concept (face capture condition) for which the system can adapt is present in the test data, and thus should be preserved over time. In order to present different

Time step t	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Reference ROI pattern sets	$Vs[1]$	$Vs[2]$	$Vs[3]$	$Vs[4]$	$Vs[5]$	$Vs[6]$	$Vs[7]$	$Vs[8]$	$Vs[9]$
Corresponding FIA sequence	Frontal camera, session 1		Frontal camera, session 2		Frontal camera, session 3		Left camera, session 1	Left camera, session 2	Left camera, session 3

Table 2: Correspondence between the 9 reference ROI pattern sets of the experimental scenario and the original *FIA* video sequences.

facial captures than the ones used for training, only the cameras with 4-mm focal length are considered for testing. While every facial capture is scaled to a same size, the shorter focal length adds additional noise (lower quality ROIs), thus accounting for reference ROIs that do not necessarily originate from the same observation environment in a real-life surveillance scenario.

5.3 Protocol for Validation

For each time step $t = 1, ..., 9$, and each individual $i = 1, ..., 10$, a temporary dataset $dbLearn^i$ is generated, and used to perform training and optimization of 2-class PFAM networks. It is composed of ROI patterns (after feature extraction and selection) from the reference set of the individual of interest (target) at time t, as well as twice the same amount of non target patterns equally selected from the UM dataset and the Cohort Model (CM) of the individual (samples from the other individuals of interest). Selection of non target pattern is performed using the *Condensed Nearest Neighbor* (CNN) algorithm (Hart, 1968). About the same amount of target and non-target patterns is generated using CNN, as well as the same amount of patterns not selected by the CNN algorithm, in order to have patterns close to the decision boundaries between target and non-target, as well as some patterns corresponding to the center of mass of the non target population.

The experimental protocol follows the (2x5 fold) cross-validation process to produce 10 independent replications, with pattern order randomization at the 5th replication. For each independent replication, $dbLearn^i$ is divided into the following subsets based on the 2x5 cross-validation methodology (with the same target and non-target proportions): (1) $dbTrain^i$ (2 folds): the training dataset used to design and update the parameters of PFAM networks, (2) ${dbVal_{ep}}^i$ (1 fold): the first validation dataset used to select the number of PFAM training epochs (the amount of presentations of patterns from $dbTrain^i$ to the networks) during the DNPSO optimization, and (3), STM^i (2 folds): the second validation dataset, used, to perform the DNPSO optimization. Using recommended parameters in (Connolly et al., 2012), an incremental learning strategy based on DNPSO is then employed to conjointly optimize all parameters of these classifiers (weights, architecture and hyper-parameters) such that the area under the P-ROC curve is minimized.

When a gradual change is detected, and a previously-learned concept is updated, an existing swarm of classifiers is re-optimized using the DNPSO training strategy. The optimization resumes from the last state – the parameters of each classifier of the swarm. On the other hand, when an abrupt change is detected, a completely new swarm is generated and optimized for the new concept $C^i_{K^i}$. The classifier specific threshold $\theta^i_{k^*}$ is computed from a ROC curve produced by the classifier $IC^i_{k^*}$ over validation data from the concept k^*, satisfying the constraint $fpr \leq 5\%$ for the highest tpr value. The classifiers from each concept are then combined into $\mathcal{P}^i = \{IC^i_1, ..., IC^i_{K^i}\}$, and another validation ROC curve is produced for the combined pool response, from which the class specific threshold Θ^i is selected with the same constraint.

The proposed system is compared to a modular version of the original system proposed in (Connolly et al., 2012), which is a passive approach. In essence, it behaves like an $AMCS_w$ that would never detect a change, and thus always incrementally learn new data for the same concept with the same incremental classifier. In addition, an adaptive version of the open-set TCM-kNN (Li and Wechsler, 2005) is also evaluated, as such system has already been applied to video-to-video FR. The same reference sequences are provided to the TCM-kNN system, and, since it is based on the kNN classifier, the update of the prototypes is straightforward. In addition, to adapt its whole architecture, its parameters are also updated at every time step, as well as the value of k (for the kNN) which is validated through (2x5 folds) cross validation. Finally, a final decision threshold Θ^i is validated for each individual of interest using the same methodology than $AMCS_w$.

To measure system performance, the classifiers are characterized by their precision-recall operating characteristics curve (P-ROC), and the area under this P-ROC ($AUPROC$). Precision is defined as the ratio $TP/(TP+FP)$, with TP and FP the number of true and false positive, and recall is another denomination of the true positive rate (tpr). The precision and recall measures can be summarized by the scalar F_1 measure for a specific operational point. Precision-recall measures enable to consider to focus on the performance over target samples, which is of a definite interest in a face re-identification application where the system is presented with a majority of non-target samples. Finally, as the number of prototypes is directly proportional to the time and memory complexity required to classify and input ROI pattern during operations, system complexity is measured as the sum of the number of prototypes ($F2$ layer neurons for all the PFAM classifiers in a pool) for $AMCS_w$ and the passive reference system, and the number of stored reference ROI pattern in TCM-kNN.

6 RESULTS AND DISCUSSIONS

ID	**Time step** t									**Tot.**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
2	X				X			X		3
21	X				X		X		X	4
69	X		X			X	X			4
72	X		X				X			3
110	X		X		X		X			4
147	X		X		X		X			4
179	X		X		X			X		4
190	X				X		X			3
198	X				X		X			3
201	X		X		X		X		X	5
Tot.	10	0	6	0	8	1	8	2	2	

Table 3: Changes detected per individual of interest (marked as a X) for each time step. The ID correspond to the IDs of the 10 individuals selected as target.

For each target individual, Table 3 presents the time steps when changes have been detected, as well as the total number of detections. $t = 1$ corresponds to the initialization of the first concepts of each individual, which is when the maximum number of changes (10) have been detected. Then, it can be observed that the 3 highest detection counts (6, 8 and 8 individuals) occur at $t = 3$, 5 and 7. These changes correspond to the introduction of training samples from the second and third frontal session, and the first profile session (left face angle). This result confirms the relation between change detection in the feature space and the observation environment. In fact, those 3 time steps are the most likely to exhibit significant abrupt changes: $t = 3$ and $t = 5$ respectively present data captured at least 2 and 3 months after the data presented at $t = 1$, and $t = 7$ is the first introduction of faces captured from a different angle.

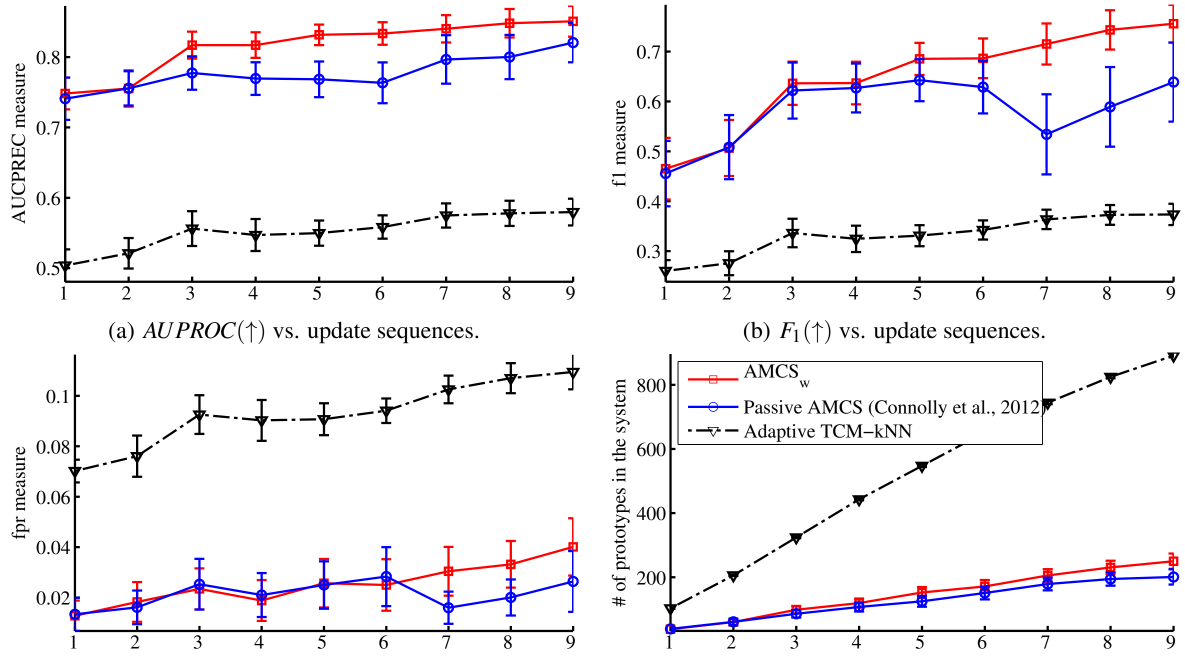
Fig. 3 shows the average overall transaction-level performance of the compared systems, for the 10 individuals of interest according to the global $AUPROC$ measure over all fpr values (Fig. 3a), and F_1 measures (Fig. 3b) at an operating point selected (during validation) to respect the constraint $fpr \leq 5\%$. Performance is assessed on predictions for each ROI pattern captured in test sequences (transactional level), after the systems are updated on each adaptation ROI pattern set.

It can be observed that the $AUPROC$ performance (Fig. 3a) for the proposed $AMCS_w$ is significantly higher than the adaptive TCM-kNN throughout the entire simulation. In addition, although higher than the adaptive TCN-kNN, the performance of the passive AMCS is also significantly lower than $AMCS_w$ from $t = 3$ until the end. $AMCS_w$ starts at 0.75 ± 0.03, and continues to increase as new ROI pattern sets are used to adapt face models, to end at 0.89 ± 0.02. Although starting at the same performance level, the passive AMCS exhibits a less significant improvement over the time, ending at 0.82 ± 0.03. Finally, TCM-kNN starts at 0.51 ± 0.02, and gradually increases to 0.58 ± 0.02 after the last reference set,

The same observations can be made for the F_1 performance (Fig. 3b) of $AMCS_w$ and TCM-kNN. $AMCS_w$ starts at 0.47 ± 0.06 and increases to end a 0.76 ± 0.04, while TCM-kNN starts at 0.26 ± 0.02 to end at 0.37 ± 0.02. In addition, the F_1 performance of the passive AMCS illustrates the knowledge-corruption that may occur when training an incremental classifier with data originating from different concepts. Although close to $AMCS_w$ up to $t = 6$, its performance significantly drops from 0.63 ± 0.05 to 0.53 ± 0.08 at $t = 7$, as a consequence of the presentation of reference data from the first profile session, and remains below $AMCS_w$ for the rest of the simulation, to end at 0.64 ± 0.08.

It can also be noted that the fpr measure (Fig. 3c) of $AMCS_w$ and the passive AMCS remain under the operation constraint of 5% fixed in validation, starting at $1.3\% \pm 0.6$ and ending at respectively $4.0\% \pm 1.1$ and $3\% \pm 1.2$. However, the fpr measure of TCM-kNN is always above the operational constraint, starting at $7.0\% \pm 0.5$ and ending at $10.1\% \pm 0.7$.

Finally, in addition to exhibiting significantly better classification performance, the memory complexity of $AMCS_w$ is significantly lower than TCM-kNN



(a) $AUPROC(\uparrow)$ vs. update sequences.

(b) $F_1(\uparrow)$ vs. update sequences.

(c) $fpr(\downarrow)$ vs. update sequences.

(d) Memory complexity($\downarrow$) vs update sequences.

Figure 3: Average overall transaction-level $AUPROC$(a), F_1(b) and fpr(c) performance of $AMCS_w$ and TCM-kNN, after the integration of the 9 pattern sets. $t = [1,2]$ corresponds to the 1st frontal angle set, $t = [3,4]$ the 2nd frontal angle set, $t = [5,6]$ the 3rd frontal angle set, and $t = \{7,8,9\}$ to the 1st, 2nd and 3rd left angle sets respectively. Memory complexity (d) is measured as the number of prototypes for the $AMCS_w$ pools and TCM-kNN systems after adaptation for each ROI pattern set. Average values of all measures and confidence interval over 10 replications are averaged for the 10 individuals of interest.

(Fig. 3d). The memory complexity of TCM-kNN grows to about 900 prototypes after the 9 adaptation sequences, while $AMCS_w$ ends with 250 ± 13.7 prototypes. As only a single incremental classifier is used for the passive AMCS, its memory complexity is the lowest, with 201 ± 28 prototypes. Considering that a prototype or reference sample is stored using 128 bytes (a vector of 32-bit floats), the reference sample stored by the TCM-kNN system after the 9 adaptation ROI pattern sets use up to 115 kBytes, while the prototypes of $AMCS_w$ use around 32 kBytes, and the incremental passive system around 26 kBytes.

7 CONCLUSION

In this paper, an adaptive framework for an AMCS is proposed for face re-identification in video surveillance, using an hybrid strategy that allows to compromise between incremental learning and ensemble generation to preserve the knowledge of historic capture conditions. A specific implementation $AMCS_w$ is used for experimentations, using an ensemble of 2-class PFAM classifiers for each enrolled individual, where all parameters are optimized using a DNPSO-training strategy, and using a Hellinger based Drift Detection Method to detect possible changes in reference videos.

Simulation results indicate that the proposed $AMCS_w$ is able to maintain a high level of performance when significantly different reference videos are learned for an individual. The proposed $AMCS_w$ exhibits higher classification performance than a reference open-set TCM-kNN system. In addition, when compared to a passive AMCS where the change detection process is bypassed, it can be observed that the proposed active methodology enables to increase the overall performance and mitigate the effects of knowledge corruption when presented with reference data exhibiting abrupt changes, yet controlling the system's complexity as the addition of new classifiers (and thus the increase of complexity) is only triggered when a significantly abrupt change is detected. The proposed $AMCS_w$ thus provides a scalable architecture that avoids issues related to knowledge corruption, and thereby maintains a high level of accuracy and robustness while bounding its computational complexity.

In the proposed scenario, the change detection has been performed with the assumption of a single concept per reference video, while different observation conditions could be observed inside a single sequence. In future research, the proposed AMCS framework could be further improved with a detection of changes inside those sequences for a better modeling of the facial models. Finally, this paper focuses on face classification of ROI patterns. In video surveillance, classification responses should be combined over several cameras and frames for robust spatio-temporal recognition.

REFERENCES

Ahonen, T. (2006). Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):2037–2041.

Alippi, C., Boracchi, G., and Roveri, M. (2013). Just-in-time classifiers for recurrent concepts. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4):620–634.

Barry, M. and Granger, E. (2007). Face recognition in video using a what-and-where fusion neural network. In *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, pages 2256–2261.

C. Pagano, E. Granger, R. Sabourin, D. G. (2012). Detector ensembles for face recognition in video surveillance. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., Rosen, D. B., and Member, S. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, 3(5):698–713.

Connolly, J.-F., Granger, E., and Sabourin, R. (2012). An adaptive classification system for video-based face recognition. *Information Sciences*, 192:50–70.

Ditzler, G. and Polikar, R. (2011). Hellinger distance based drift detection for nonstationary environments. In *Computational Intelligence in Dynamic and Uncertain Environments (CIDUE), 2011 IEEE Symposium on*, pages 41–48.

Eberhart, R. C. and Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Proceedings of the sixth international symposium on micro machine and human science*, volume 1, pages 39–43. New York, NY.

Fritzke, B. (1996). Growing self-organizing networks - why? In *In ESANN96: European Symposium on Artificial Neural Networks*, pages 61–72. Publishers.

Goh, R., Liu, L., Liu, X., and Chen, T. (2005). The CMU face in action (FIA) database. In *Analysis and Modelling of Faces and Gestures*, pages 255–263.

Gorodnichy, D. (2005). Video-based framework for face recognition in video. In *Proceedings Canadian Conference on Computer and Robot Vision*, pages 330–338.

Hart, P. (1968). The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14(3):515–516.

Kittler, J. and Alkoot, F. M. (2003). Sum versus vote fusion in multiple classifier systems. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 25, pages 110–115.

Kuncheva, L. I. (2004). *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience.

Kuncheva, L. I. (2008). Classifier ensembles for detecting concept change in streaming data: Overview and perspectives. In *2nd Workshop SUEMA 2008 (ECAI 2008)*, pages 5–10.

Li, F. and Wechsler, H. (2005). Open set face recognition using transduction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(11):1686–1697.

Lim, C. and Harrison, R. (1995). Probabilistic fuzzy ARTMAP: an autonomous neural network architecture for bayesian probability estimation. In *Proceedings of 4th International Conference on Artificial Neural Networks*, pages 148–153.

Matta, F. and Dugelay, J.-L. (2009). Person recognition using facial video information: A state of the art. *Journal of Visual Languages & Computing*, 20(3):180 – 187.

Minku, L., White, A., and Yao, X. (2010). The impact of diversity on online ensemble learning in the presence of concept drift. *EEE Transactions on Knowledge and Data Engineering*, 22(5):730–742.

Minku, L. L. and Yao, X. (2012). DDD: A New Ensemble Approach for Dealing with Concept Drift. *IEEE Transactions on Knowledge and Data Engineering*, 24(4):619–633.

Narasimhamurthy, A. and Kuncheva, L. (2007). A framework for generating data to simulate changing environments. In *25th IASTED International Multi-Conference: artificial intelligence and application*, pages 384–389.

Nickabadi, A., Ebadzadeh, M. M., and Safabakhsh, R. (2008). DNPSO: A dynamic niching particle swarm optimizer for multi-modal optimization. In *2008 IEEE Congress on Evolutionary Computation, CEC 2008*, pages 26–32.

Oh, I.-S. and Suen, C. Y. (2002). A class-modular feed-forward neural network for handwriting recognition. *Pattern Recognition*, 35(1):229 – 244. Shape representation and similarity for image databases.

Polikar, R. and Upda, L. (2001). Learn++ : An Incremental Learning Algorithm for supervised neural networks. In *IEEE Transactions on Systems, Man and Cybernetics*, volume 31, pages 497–508.

Tax, D. and Duin, R. (2008). Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29:1565–1570.

Viola, P. and Jones, M. J. (2004). Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57:137–154.

Zhou, S. K., Chellappa, R., and Zhao, W. (2006). *Unconstrained face recognition*, volume 5. Springer.