Accepted in Int. Conference on Pattern Recognition (ICPR), 2014

Self-Updating with Facial Trajectories for Video-to-Video Face Recognition

Miguel De-la-Torre, Eric Granger, Paulo Radtke, Robert Sabourin
Laboratoire d'imagerie de vision et d'intelligence artificielle
École de technologie supérieure, Université du Québec, Montreal, Canada
miguel@livia.etsmtl.ca, eric.granger@etsmtl.ca
radtke@livia.etsmtl.ca, robert.rabourin@etsmtl.ca

Dmitry O. Gorodnichy
Science and Engineering Directorate
Canada Border Services Agency
Ottawa, Canada
dmitry.gorodnichy@cbsa-asfc.gc.ca

***Abstract*—For applications of face recognition (FR) in video surveillance, it is often costly or unfeasible to collect several high quality reference samples a priori to design representative facial models. Moreover, changes in capture conditions and human physiology create divergence between facial models and input captures. Multiple classifier systems (MCS) have been successfully applied to video-to-video FR, where the face of each individual of interest is modeled using an ensemble of 2-class classifiers (trained on target vs. non-target samples). However, the reliable self-update of these individual-specific ensembles with relevant target and non-target samples raises several challenges. In this paper, an adaptive MCS is proposed that allows for self-updating facial models given face trajectories captured during operations. Different faces appearing in a camera viewpoint are tracked, and ensemble predictions for facial captures are accumulated along each track for robust video-to-video FR. When the number of positive predictions over time surpasses an *update threshold*, the target face samples extracted from the trajectory are combined with non-target samples selected from the cohort and universal models for efficient self-update the corresponding face model. A learn-and-combine strategy is then employed to avoid knowledge corruption during self-update of an ensemble. At a transaction level, the adaptive MCS outperforms the reference systems that do not allow self-updating on Face in Action videos. Analysis at a trajectory level indicates that the proposed system allows for robust spatio-temporal recognition, which translates to enhanced security and situation analysis.**

I. Introduction

Face recognition (FR) systems are increasingly employed in video surveillance applications to rapidly determine if facial regions detected across a network of video cameras correspond to the facial model[1] of individuals of interest. For instance, FR is used in person re-identification applications for search and retrieval, which involves video-to-video FR. In video-to-video FR, facial regions of interest (ROIs) extracted from video streams are employed to design of facial models.

In video surveillance, tracking information may be used to record a complete trajectory from arrival until the individual leaves the scene. A *facial trajectory* is defined as a set of facial ROIs (produced by a face detection process) that correspond to a same high quality track of an individual across consecutive frames. In addition, individuals in a scene may be tracked, and the corresponding ROIs isolated through face detection may regrouped. A system for spatio-temporal FR will recognize individuals over time based on a group of ROIs.

The design of facial models is ideally performed during enrollment, using high-quality reference ROI samples captured for the target individual. This requirement is challenging in practical video-to-video FR. Given semi- and unconstrained capture conditions, video-to-video FR systems must recognize faces that exhibit changes in illumination, scale, blur, pose, occlusion and expression. Therefore, enrollment of an individual relies on a limited number of reference sample, resulting in facial models that are poor representatives of faces to be recognized during operations.

Several adaptive classifiers have been proposed in literature [1]–[4] for incremental learning of labeled samples. These can be used to update facial models from new reference data captured after enrollment, allowing to maintain or increase matching performance. Adaptive multiple classifier systems (MCS) have been successfully applied for FRiVS [5]. In these systems, the face model of each individual is encoded using an ensemble of 2-class classifiers or detectors (EoD), allowing a high level of discrimination between target and non-target individuals. In this paper, it is assumed that face matching is performed with an EoD per individual of interest [1].

An issue with the supervised update of face models is the analysis and labelling of new reference videos captured during operations. This costly process must be addressed manually by a domain expert that isolates target faces in video surveillance footage. Rather than relying on a human expert, the system may perform self-updating of facial models with operational videos. An approach to exploit both labeled and unlabeled data in adaptive biometrics is self-update [6].

In this paper, an adaptive MCS is proposed for self-update of the facial models under semi- and uncontrolled capture conditions seen in video-to-video FR using facial trajectories. During operations, information from a tracker and an individual-specific EoD is integrated at a decision-level for enhanced spatio-temporal FR. A detection threshold is applied to the accumulated positive EoD predictions over trajectories to produce a decision, and a second (higher) update threshold allows to select update trajectories. When a new trajectory is suitable for update (i.e., surpasses the higher update threshold), its facial ROIs are combined with those of non-target samples selected from the cohort and universal models. This block of data is comprised of diverse facial regions associated with target and non-target trajectories, and allows to generate a new pool of 2-class classifiers, and to update the fusion function of the user specific EoD. This learn-and-combine strategy has been shown to reduce the impact of knowledge corruption in adaptive ensembles [1]. Practical

[1] A facial model is defined as either a set of one or more reference samples (for template matching), or a statistical model estimated during training with reference samples (for classification).

DOI: 10.1109/ICPR.2014.294

memory limitations impose the need for a method to select and manage the most relevant validation samples. A long term memory (LTM) is maintained over time with a fixed number of validation samples per individual. These samples are ranked and selected according to their relative entropy with the Kullback-Leibler (KL) divergence [7].

One challenge with self-update is the reliable selection operational samples from the target individual to update facial models. A high level of confidence is required to avoid updating models with impostor or non-target data. The proposed adaptive MCS employs the tracker quality to regroup detected facial regions in facial trajectories, and applies a threshold to the accumulated EoD predictions over a trajectory to produce accurate decisions. A second (higher) threshold is applied to select high confidence trajectories that can be used for update. The system then performs self-update of the corresponding facial models using all facial regions of interest (ROIs) linked to a high confidence update trajectory. A single face trajectory may contain target ROIs that were incorrectly classified by the MCS, this allows facial models to be adapted with a diversified set of reference samples that are close to the boundaries between target and non-target distributions, and thereby improve the generalization performance.

The proposed MCS was validated with the Faces in Action video dataset, and each EoD in the MCS was designed with ARTMAP neural classifiers. After supervised learning of an EoD with enrollment videos, new videos from different operational sessions are processed by the system, allowing to self-update face models for high confidence trajectories. Performance is assessed at the transaction and trajectory levels.

II. Adaptive video-to-video FR

Assume that video-to-video FR is performed on video streams that are captured using one or more video surveillance cameras. First, the face detection (segmentation) process isolates the facial ROIs corresponding to faces captured in frames. Then, invariant and discriminant features are extracted for tracking and classification functions. Tracking follows the movement or expression of faces in consecutive video frames and regroup facial regions of a same person, whereas classification matches ROIs to the facial models of individuals enrolled to the system. A track ID is typically initialized with an ROI that is detected at different locations than other faces. Finally, the decision function combines the tracking IDs and classification scores in order to predict a list of likely individuals in the scene.

In the literature, FR in video surveillance (FRiVS) is addressed as an open set or open-world problem, where the number of individuals of interest is greatly outnumbered by other individuals. A multi-class classifier designed to address the open set problem in video FR is the TCM-kNN proposed by Li and Wechsler in [8]. This matcher takes advantage of transductive inference to generate a class prediction based on randomness deficiency. Tax and Duin also proposed propose a heuristic to combine any type of one-class classifiers for multi-class classification with outlier rejection. It allows to adjust the rejection threshold per class, and to combine class models that are not based on probability densities [9].

Modular architectures with one detector per individual have been proposed to address the problem with individual-specific 1- or 2-class classifier [5]. For instance, Kamgar-Parsi et al. propose an approach based on the identification of the decision region(s) in the feature space of individual-specific faces by training a dedicated feed-forward neural network for each individual of interest [10]. Another recent example is the SVM-based modular system proposed by Ekenel et al., for access control [11]. The architectures have been extended to train an ensemble of detectors per (EoD) per individual. An example of such systems is the EoD (2-class classifiers) proposed by Pagano et al. It allows for the generation of a diversified pool of ARTMAP neural networks using a DPSO based training strategy, and detectors are then selected and combined in the ROC space using Boolean combination (BC) [5]. Non-target samples are retrieved from the cohort and universal background models.

Spatio-temporal approaches for FR merge spatial information (e.g. face appearance) with the sequential variations presented over time (e.g behavior). Zhang and Martinez use probabilities accumulated by matching ROIs to the individual-specific Gaussian mean estimated from gallery reference samples, and normalize to produce posterior probabilities. This temporal analysis is independent of the matching or tracking algorithm [12]. Liu and Chen used HMMs to model the appearance and dynamics of a person, obtaining high confident results on sequences that were then used to adapt the models. A potential problem with the modeling of probability distributions of the motion is the assumption that the movement will be very similar, regardless of the new scenario [13]. Accumulating classification responses over time eliminates the assumption, and still takes into account the time information. For instance, the work of Ekenel et al. evaluates a video-to-video FR system for individuals entering into a room, which progressively combines confidence scores of the matchers using a sum rule over the full sequences to estimate the identity in video [11]. In their approach, they use a k-NN classifier on a DCT representation of face images, and use min-max normalization on the distance-based output scores, and then compare their proposed approaches: distance-to-model, distance-to-second-closest and a combination of both.

One of the main challenges encountered in video surveillance is that facial models lose their representativeness over time because they are designed *a priori* using limited numbers of reference samples captured from semi- and uncontrolled environments. Facial ROIs incorporate considerable variations due to limited control over capture conditions (e.g., illumination, and pose), and to changes is physiology over time (e.g., aging). These factors often result in facial models that are poor representatives of faces to be recognized during operations.

A. Self-Updating in Biometrics:

Strategies for the design of adaptive biometric systems involve (1) the *selection* of diversified, relevant reference samples to update a template gallery or an LTM of reference validation samples, and (2) the actual *update* of template galleries or classifier parameters using supervised or semi-supervised learning schemes. Techniques that are suitable for the selection of relevant samples in adaptive MCS have been reviewed in [7].

Several approaches in literature are suitable for semi-supervised learning of face models in adaptive biometrics [14]. Self-update methods have been proposed for template matching [6], where biometric models are first designed by storing samples from a labeled data set D_L in a template gallery $\mathcal{G}$. Then, during operations, similarity scores for unlabeled samples are produced through template matching. Positive prediction is output if the score surpasses the decision threshold γ^d. Predictions linked to a high degree of confidence (surpassing a higher updating threshold, $\gamma^u \geq \gamma^d$), are integrated to the gallery $\mathcal{G}$, thereby updating the corresponding biometric models. Similar self-updating strategies methods (e.g., [15]–[17]) have been proposed for neural or statistical classification systems that estimate of biometric models. Although self-update methods can improve accuracy [14], adapting a biometric system using operational data carries an inherent risk. There exists a trade-off between the false updates and false rejections that affect of performance, and the decision threshold is crucial in self-update system.

B. Adaptive Face Recognition:

In the literature, adaptive FR systems have traditionally incorporated newly-acquired reference samples to update the selection of a user's template from a gallery, via clustering and editing techniques. Processing thus allows an improved representation of intra-class variations to be obtained with a single template. Some adaptive biometric systems have been proposed to refine facial models according to intra-class variations in input samples [18].

Approaches for self-update of facial models are generally based on the matching scores. For instance, in [19], Euclidean distance-based measure of similarity is used, and at each iteration, the PCA-based feature space for matching is updated with the newly-acquired soft-labeled samples. An extension to the self-update algorithm named the Graph Mincut [20], has been proposed to update templates by analyzing the underlying graphical structure of input operational data. A pairwise similarity measure between operational data and existing templates is used to draw a graph that relates these samples. A representative example of adaptation in spatio-temporal FR, that exploits classification similarity and video information have been proposed by Franco et al. [21]. The authors propose an incremental template update strategy that is based on the similarity between captured ROIs and templates. It exploits the frequency of face detections over a complete video sequences of the different subjects in the scene, and their last position within the frame in the sequences.

More recently, adaptive MCS have been proposed for *supervised update* of facial models in video-to-video FR. An incremental learning strategy based on DPSO has been proposed for video-based access control [4]. It allows the evolution of an ensemble of heterogeneous multi-class classifiers from new videos. An ensemble of two-class classifiers or detectors (EoDs) per individual has been proposed by dela Torre et al. [1]. When a new data block becomes available, a diversified pool of PFAM classifiers is generated with a DPSO learning strategy and combines to others using Boolean combination (BC). Learn++ is a well-known ensemble-based technique for incremental learning that has been tested on FR problems [2]. This technique is inspired by the AdaBoost algorithm – it performs supervised incremental learning by incorporating a new set of weak classifiers to the ensemble each time new data becomes available.

III. Self-Updating with Facial Trajectories

In this paper, an adaptive MCS for video-to-video FR is proposed, where new trajectories captured during operations allows for self-updating facial models. As shown in Fig. 1, the proposed system is comprised of a segmentation module for face detection, a face tracker, a modular classification system with one EoD per individual of interest, a decision fusion module, and a design/update module.

During operations, information from a tracker and individual-specific EoDs are integrated at a decision level. The face tracker initializes a new trajectory with the first facial ROI captured by the segmentation system in a different area of the scene. As the tracker follows the facial region through the scene, the segmentation system captures facial ROIs for some of the frames, allowing to produce a trajectory T. The diverse set of ROIs belongs to the same individual is defined by the tracker. Feature vectors are extracted from ROIs segmented in each frame, and presented to each EoD$_k$. Each EoD$_k$ is comprised of a pool of 2-class classifiers $\mathcal{P}_k = \{c_{1,k}, ..., c_{M,k}\}$, and a fusion function $\mathcal{F}_k$ that is designed using a validation set D_k^c, for $k \in \{1, ..., K\}$. Ensemble member $c_{m,k}$ produces an output score $s_{m,k}^{+}(\mathbf{a})$ for a given feature vector $\mathbf{a}$ corresponding to an input ROI. The scores are then combined using $\mathcal{F}_k$. Each EoD_k produces an output prediction $p_k(\mathbf{a})$. Positive predictions are then accumulated over time in the decision fusion module. Depending on the strategy used for fusion, a subset of the classifiers in the pool $\mathcal{P}_k$ is selected to maximize performance.

Faces in a video sequence are tracked from frame to frame and regrouped, and the positive EoD$_k$ predictions p_k along a trajectory T are accumulated over time for robust spatio-temporal recognition. Finally, an individual-specific threshold is applied to the accumulation curves of each EoD$_k$ in order to generate an overall prediction d_k for each EoD$_k$. There are several evidence accumulation modules per track ID, to simultaneously recognize several people at a time in the scene. A *highly confident* trajectory T is associated with an individual of interest k when the number of accumulated positive predictions from the EoD_k (over a fixed-size window) surpasses the update threshold, $A_k \geq \gamma_k^u$, the *design/update* system assigns the label k to the trajectory T_k for update.

The adaptive MCS detects the presence of individuals of interest based on the number of positive EoD_k predictions over trajectories. Given a high quality trajectory T, each EoD_k generates a prediction $p_k(\mathbf{a}_n)$ for each sample $\mathbf{a}_n$ associated with a ROI in the trajectory. Output predictions from EoD_k over the ROI samples of a trajectory T, at the selected operations point, are defined by the set $\mathbf{P}_k = \{p_k(\mathbf{a}_1), ..., p_k(\mathbf{a}_N)\}$, associated with each input ROI sample $\mathbf{a}_n$. Negative predictions set $p_k(\mathbf{a}_n) = 0$, and positive ones set $p_k(\mathbf{a}_n) = 1$. The decision fusion system accumulates the number of positive predictions A_k of each EoD_k on fixed size window W according to:

$$A_k = \sum_{i=0}^{W-1} p_k(\mathbf{a}_{(W-i)}) \quad \in [0, W] \qquad (1)$$

For instance, a window of size $W = 30$ accumulates the last 30 positive predictions from the same trajectory. Each EoD_k accumulates a sequence of positive predictions that range from 0 (EoD_k made only negative predictions for W), to a maximum of W (EoD_k made only positive predictions for the last W ROIs).

Based on these accumulations A_k, for $k = 1, ..., K$, the system produces decisions. If A_k surpasses threshold γ_k^d, the system detects the presence of individual k and alerts the operator. Furthermore, if A_k surpasses the update threshold γ_k^u,

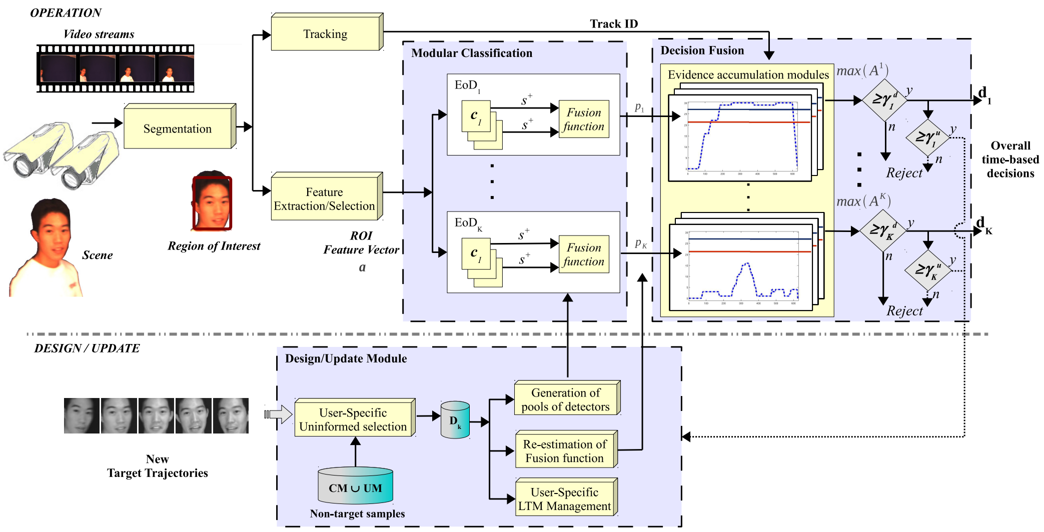


Fig. 1. Block diagram of the proposed MCS that for video-to-video FR that allows for self-updating of facial models.

the trajectory is suitable for self-updating of the corresponding EoD_k. Given the negative effects on performance caused by false updates, threshold γ_k^u is greater or equal to γ_k^d. For each EoD_k, the detection threshold γ_k^d is estimated using a validation set composed of one positive and several negative trajectories. In this way, a single target trajectory is required for design and update of the facial model.

In the design/update phase, when a new facial trajectory T_k becomes available for individual k, One-Sides Selection is used to form a individual-specific training set D_k with all its target samples and non-target samples selected from CM and UM. An ensemble EoD_k is updated with new ROIs from a trajectory T_k by generating new pool base detectors, and adding these to a pool $\mathcal{P}_k$ of previously trained detectors, and updating the fusion function according to the old and new validation samples. [1]. A fixed size LTM is maintained with validation samples that are representative of the decision bound between target and non-target distributions. When a new validation set D with target and non-target samples becomes available for individual k, all samples are ranked according to the Kullback-Leibler divergence. Then, the $\lambda_k/2$ highest ranked target samples, as well as the $\lambda_k/2$ highest ranked non-target samples are preserved, whereas the rest are discarded. The procedure followed by the management strategy to rank and select representative validation samples to be stored in the LTM$_k$ [7]. The decision-level fusion function is updated based on new data and pre-stored reference samples (from the LTM).

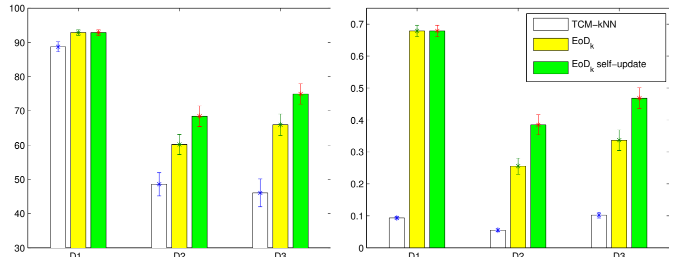
IV. Experimental Results

The adaptive MCS proposed in this paper for video-to-video FR is characterized for person re-identification application, using videos from the Carnegie Mellon University Face in Action (FIA) database [22]. This database consists of 20 second videos captures from 244 subjets under semi-constrained conditions mimicking a passport checking scenario. An array of 6 cameras horizontally positioned at face level, and positioned at 0^o (frontal) and $\pm72^o$ (left and right) angle with respect to the individual. Three cameras were set to an 8-mm focal-length (zoomed), resulting in face regions of about300×300 pixels, and the other three to a 4-mm focal-length (unzoomed) with regions of about 100×100 pixels. Faces are captured at 30 frames per second, a Sony ICX424 camera at a resolution of 640x480 pixels. Data has been captured on three sessions separated by a three months interval for each individual.

Ten individuals were randomly selected for re-identification (k = FIA ID = 2, 58, 72, 92, 147, 151, 176, 188, 190 and 209), and one EoD$_k$ is designed for each. Of the remaining individuals, 88 are selected as part of the universal model (UM), and the rest are considered as unknown test individuals. Face trajectories from individuals of interest contain between 80 and 239 facial ROIs, and non-target training and test samples differ in each dataset. Facial trajectories were formed with frontal facial ROIs segmented using the Viola-Jones algorithm, and the CAMSHIFT algorithm for face tracking. All ROIs are scaled to 70x70 pixels, the resolution of the smallest face captures after face detection. The Multi Scale LBP [23] feature extractor has been used with three different block sizes (3×3, 5×5 and 9×9), along with pixel intensity features. Resulting features are combined into feature vectors, and PCA is applied to select the 32 most discriminant projected features.

Prior to computer simulations, four data subsets have been prepared. Trajectories in the design dataset D are comprised of target ROIs from the the zoomed view of capture session 1. The test/adaptation datasets D_1 to D_3 have been constructed with ROIs from the unzoomed view of capture sessions 1 to 3 respectively. Non-target samples are independently selected for each of the training/validation sets picked from the cohort model (CM) and UM, using One-Side Selection. The CM comprises trajectories from non-target individuals enrolled to the system. The MCS used for simulations is comprised of an ensemble of 2-class Probabilistic Fuzzy ARTMAP (PFAM) classifiers per individual of interest, EoD$_k$(PFAM). The DPSO learning strategy was used for classifiers generation, and Boolean Combination was applied for decision-level fusion of classifiers in the ROC space [1]. The reference approaches for baseline comparison are the multi-class TCM-kNN [8] and the same MCS that does not allow for self-update (see [5]). After performance evaluation on D_1 the classifiers were updated with trajectories in D_1 and tested on D_2. The same process was repeated for update/test on D_2 and D_3 respectively.

Evaluation was performed following 2×5-fold cross-



(a) pAUC-5%

(b) F_1

Fig. 2. Average transaction-based performance of the proposed MCS and referencee systems for 10 independent experiments, and over the 10 individuals of interest, in terms of (a) pAUC(5%) and (b) F_1 (at $fpr = 1\%$). Systems were designed and updated with D, D_1 and D_2, and performance is shown after testing on $D_1 \rightarrow D_2 \rightarrow D_3$

TABLE I. AVERAGE PERFORMANCE OF PROPOSED MCS SYSTEM FOR INDIVIDUALS k = 58 AND 188 OVER 10 REPLICATIONS OF THE EXPERIMENT.

	EoD$_{58}$					EoD$_{188}$				
fpr ↓										
	0.233 ±0.094	→	0.863 ±0.085	→	1.619 ±0.385	2.544 ±0.567	→	1.175 ±0.201	→	0.310 ±0.098
tpr ↑										
	84.432 ±3.328	→	35.442 ±8.100	→	51.163 ±14.319	89.576 ±4.256	→	89.884 ±3.093	→	93.698 ±1.744
$\mathbf{F_1}$ ↑										
	0.849 ±0.023	→	0.353 ±0.066	→	0.384 ±0.093	0.472 ±0.054	→	0.667 ±0.031	→	0.920 ±0.013
pAUC(5%) ↑										
	98.455 ±0.225	→	74.578 ±3.539	→	80.443 ±6.337	91.120 ±2.408	→	96.387 ±0.480	→	99.723 ±0.050

validation for 10 independent trials. Target samples from the learning set were randomly split according to a uniform distribution, in 5 folds of the same size. The folds were first distributed in three different design sets, including two folds for training (D_t^t), $1\frac{1}{2}$ folds to stop training epochs (D_t^e), and $1\frac{1}{2}$ folds for fitness evaluation (D_t^f). Validating the number of training epochs for classifier convergence was performed on D_t^e, whereas particle fitness was evaluated on D_t^f. The DPSO algorithm was initialized with a swarm of 60 particles, and a maximum of 5 particles within each of the 6 subswarms. The algorithm was set to run a maximum of 30 iterations, allowing 5 extra iterations to ensure convergence. Once the global best particle is found, its classifier and the 6 local bests from each subswarm were added to the EoD$_k$(PFAM). Once the classifiers were trained, D_t^e and D_t^f are combined, randomized and divided in two equally distributed subsets to produce a validation data for threshold/fusion function estimation (D_t^c), and to select the operations point (D_t^s). Each fold was assigned to a different training/validation set at each replica of the experiment. At replication 5, the five folds were regenerated after a randomization of the sample order for each class, and the process was repeated to generate a standard error on ten different assignments. Evaluation was performed at the transaction level (in the ROC and Precision-Recall spaces), and at the trajectory level (time-based analysis of the entire system over video sequences).

Figure 2 presents the average transaction-level performance for the reference and proposed systems. The performance is evaluated using the partial AUC for a $0 \leq fpr \leq 0.05$: pAUC (5%) and the scalar F_1 measure for a desired operations point of $fpr = 1\%$. Performance for modular systems were measured for each individual (EoD$_k$), and average values are presented. In order to have comparable results for TCM-kNN, empirical ROC curves were estimated on validation for each individual. The selection of the operations point, as well as performance evaluation were computed after applying the specialized rejection threshold of the TCM-kNN. Note that this rejection threshold is estimated on the training data, taking advantage of the peak-side-ratio that characterizes the distributions of p-values for each class.

Overall results for all approaches show a degradation in the system performance after testing on D_2, with a slight recovery after testing on D_3. This decline in performance underscores the importance of adapting facial models as new reference videos become available. Except for the initial test (on D_1) results indicate that the self-update of ensembles during operations with high quality trajectories allows a better recovery in performance than both reference systems. An advantage of the proposed system is the incorporation of diversified information into facial models of detected individual. Self-updating provides EoDs with a greater diversity of samples captured under various conditions (pose, lighting, etc). These samples allow for a more accurate definition of the boundaries between target and non-target individuals in accordance with the most recent facial samples.

By observing the performance of the system for specific individuals (see Table I), it can be observed that the individual 58 initially exhibits a high level of performance. EoD$_{58}$ is negatively affected by updates – results for EoD$_{58}$ show a significant decline in performance after updating on D_1 (testing on D_2). However, EoD$_{188}$ presents a constant increase in performance. Despite the incorrect updates, the fpr decreases after each self-update. This suggests that lamb-like individuals stand to benefit somewhat from diverse of samples from incorrect updates. In fact, incorrect self-updates favors diversity between old and new classifiers, and may result in an increase in the performance of the updated EoD.

In the proposed MCS, the face tracker groups ROIs corresponding to trajectories initiated in each video sequence. EoD prediction for each ROI in each trajectory are accumulated over time. To assess the overall system performance over time, Figures 3 and 4 show the result of a trajectory-based analysis for individuals with ID 58 and 188 on the same experimental trial. Results are shown for 3 different evidence accumulation strategies according to time and in the ROC space. Accumulation of EoD predictions over the ROIs in the trajectory provides the better discrimination, specially if segmentaion does not capture many ROIs in a track due to poor capture quality. Results also suggest that decision-level fusion with a threshold-optimized techniques like Boolean

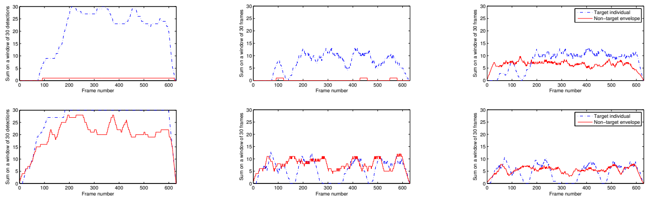


Fig. 3. Example of a trajectory-based analysis with EoD$_{58}$ (top row) and EoD$_{188}$ (bottom row) for the 3 evidence accumulation strategies: accumulation of EoD predictions over the ROIs in the trajectory (left), over each frame (center), and accumulation of scores over ROIs in a trajectory (right).

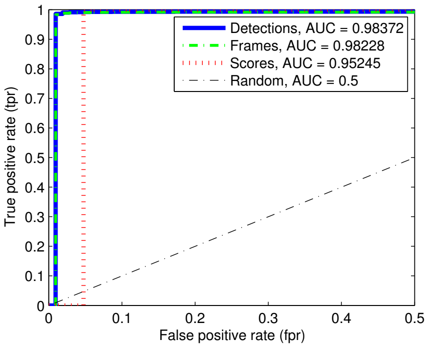


Fig. 4. ROC performance for 3 evidence accumulation strategies from trajectory-based analysis with EoD_{58}.

combination provide a higher level of performance because thresholds are specialized to base detectors.

V. Conclusion

In this paper, a modular and adaptive MCS, with user-specific ensembles of detectors was proposed for video-to-video FR. It allows for self-updating of facial models based on trajectories defined by the tracker. During operations, it integrates track IDs of a face tracker and predictions of a individual-specific ensemble at a decision-level for enhanced video-to-video FR. Accumulated predictions of a window of W frames over each trajectory define the corresponding accumulation curve for a given module k. When the accumulation curve surpasses a detection threshold, the individual is detected as a target. If the curve surpasses a more conservative update threshold, the ROI samples of the trajectory are used for self update of the system. In order to update facial models, target facial regions from the trajectory are combined with non-target samples selected from the cohort and universal models, using an extended condensed nearest neighbor selection.

Transaction-based results obtained with ensembles of 2-class ARTMAP classifiers generated using a DPSO strategy on videos from the CMU-FIA dataset indicate that the proposed adaptive MCS outperforms reference systems that do not perform self-update. Trajectory-based analysis shows the increased discrimination achieved when EoD predictions are accumulated according to a trajectory, leading to robust spatio-temporal recognition. The proposed system has been characterized using data that incorporates a gradual pattern of changes for facial models, over different capture sessions. However, future research should include system performance under both gradual and abrupt patterns of change, as seen in variations of illumination and pose.

Acknowledgment

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada, and the Mexican National Council for Science and Technology.

References

[1] M. De-la Torre, E. Granger, P. V. W. Radtke, R. Sabourin, and D. O. Gorodnichy, "Incremental update of biometric models in face-based video surveillance," in *Proc. IJCNN*, Brisbane, Australia, June 2012, pp. 1–8.

[2] R. Polikar, L. Udpa, S. S. Udpa, and V. Honavar, "Learn++: An Incremental Learning Algorithm for MLP Networks," *IEEE Trans. SMC*, vol. 31, no. 4, pp. 497–508, 2001.

[3] R. Singh, M. Vatsa, A. Ross, and A. Noore, "Biometric classifier update using online learning: A case study in near infrared face verification," *Image and Vision Computing*, vol. 28, pp. 1098–1105, 2010.

[4] J.-F. Connolly, E. Granger, and R. Sabourin, "Evolution of heterogeneous ensembles through dynamic particle swarm optimization for video-based face recognition," *Pattern Recognition*, vol. 45, no. 7, pp. 2460 – 2477, 2012.

[5] C. Pagano, E. Granger, R. Sabourin, and D. O. Gorodnichy, "Detector ensembles for face recognition in video surveillance," in *IJCNN*, Brisbane, Australia, June 2012, pp. 1–8.

[6] F. Roli, L. Didaci, and G. Marcialis, "Template co-update in multimodal biometric systems," in *International Conference on Biometrics*, vol. 4642, Seoul, Korea, August 2007, pp. 1194 – 202.

[7] M. De-la Torre, E. Granger, R. Sabourin, and D. O. Gorodnichy, "An individual-specific strategy for management of reference data in adaptive ensembles for person re-identification," in *Proc. ICDP*, London, UK, December 2013, pp. 1–7.

[8] F. Li and H. Wechsler, "Open set face recognition using transduction," *IEEE Trans. on PAMI*, vol. 27, no. 11, pp. 1686–97, 2005.

[9] D. Tax and R. Duin, "Growing a multi-class classifier with a reject option," *Pattern Recognition*, vol. 29, no. 10, pp. 1565 – 70, 2008.

[10] B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi, "Toward development of a face recognition system for watchlist surveillance," *IEEE Trans. PAMI*, vol. 33, no. 10, pp. 1925 – 37, 2011.

[11] H. K. Ekenel, J. Stallkamp, and R. Stiefelhagen, "A video-based door monitoring system using local appearance-based face models," *Computer Vision Image Understanding*, vol. 114, no. 5, pp. 596–608, May 2010.

[12] Y. Zhang and A. Martinez, "From stills to video: Face recognition using a probabilistic approach," in *Computer Vision and Pattern Recognition Workshop Conference on*, 2004, p. 78.

[13] X. Liu and T. Cheng, "Video-based face recognition using adaptive Hidden Markov Models," in *Proceedings 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, Los Alamitos, CA, USA, 2003, pp. 340 – 5.

[14] A. Rattani, B. Freni, G. L. Marcialis, and F. Roli, "Template update methods in adaptive biometric systems: A critical review," in *Lecture Notes in Computer Science (included Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5558, Alghero, Italy, 2009, pp. 847 – 856.

[15] K. Okada, L. Kite, and C. von der Malsburg, "An adaptive person recognition system," in *Proceedings 10th IEEE International Workshop on Robot and Human Interactive Communication*, Piscataway, NJ, USA, 2001, pp. 436–41.

[16] K. Lu, Z. Ding, J. Zhao, and Y. Wu, "A novel semi-supervised face recognition for video," in *Proc. of the International Conference on Intelligent Control and Information Processing*, 2010, pp. 313–316.

[17] G. Yu, G. Zhang, C. Domeniconi, Z. Yu, and J. YouZ, "Semi-supervised classification based on random subspace dimensionality reduction," *Pattern Recognition*, vol. 45, no. 3, pp. 1119 – 1135, 2012.

[18] F. Roli, L. Didaci, and G. L. Marcialis, "Adaptive biometric systems that can improve with use," in *Advances in Biometrics: Sensors, Systems and Algorithms*, N. R. V. Govindaraju, Ed. Springer, 2008, pp. 447–471.

[19] F. Roli and G. L. Marcialis, "Semi-supervised pca-based face recognition using self-training," in *JIAPR - Int. Workshop on Structural and Syntactical Pat. Rec. and Statistical Techniques in Pat. Rec.*, vol. 4109. Hong Kong, China: Springer, August 2006, pp. 560–568.

[20] A. Rattani, G. Marcialis, and F. Roli, "Capturing large intra-class variations of biometric data by template co-updating," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Piscataway, NJ, USA, 2008, pp. 1–6.

[21] A. Franco, D. Maio, and D. Maltoni, "Incremental template updating for face recognition in home environments," *Pattern Recognition*, vol. 43, no. 8, pp. 2891 – 903, 2010.

[22] R. Goh, L. Liu, X. Liu, and T. Chen, "The CMU Face In Action Database," in *Analysis and Modelling of Faces and Gestures*. Carnegie Mellon University, 2005, pp. 255–263.

[23] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Tr. PAMI*, vol. 24, no. 7, pp. 971–87, 2002.