Video-Based Face Recognition Using Ensemble of Haar-Like Deep Convolutional Neural Networks

Mostafa Parchami
Computer Science and Engineering Dept.,
University of Texas at Arlington, TX, USA
mostafa.parchami@mavs.uta.edu

Saman Bashbaghi and Eric Granger
École de technologie supérieure,
Université du Québec, Montréal, Canada
bashbaghi@livia.etsmtl.ca, eric.granger@etsmtl.ca

***Abstract*—Growing number of surveillance and biometric applications seek to recognize the face of individuals appearing in the viewpoint of video cameras. Systems for video-based FR can be subjected to challenging operational environments, where the appearance of faces captured with video cameras varies significantly due to changes in pose, illumination, scale, blur, expression, occlusion, etc. In particular, with still-to-video FR, a limited number of high-quality facial images are typically captured for enrollment of an individual to the system, whereas an abundance facial trajectories can be captured using video cameras during operations, under different viewpoints and uncontrolled conditions. This paper presents a deep learning architecture that can learn a robust facial representation for each target individual during enrollment, and then accurately compare the facial regions of interest (ROIs) extracted from a still reference image (of the target individual) with ROIs extracted from live or archived videos. An ensemble of deep convolutional neural networks (DCNNs) named HaarNet is proposed, where a trunk network first extracts features from the global appearance of the facial ROIs (holistic representation). Then, three branch networks effectively embed asymmetrical and complex facial features (local representations) based on Haar-like features. In order to increase the discriminativness of face representations, a novel regularized triplet-loss function is proposed that reduces the intra-class variations, while increasing the inter-class variations. Given the single reference still per target individual, the robustness of the proposed DCNN is further improved by fine-tuning the HaarNet with synthetically-generated facial still ROIs that emulate capture conditions found in operational environments. The proposed system is evaluated on stills and videos from the challenging COX Face and Chokepoint datasets according to accuracy and complexity. Experimental results indicate that the proposed method can significantly improve performance with respect to state-of-the-art systems for video-based FR.**

I. Introduction

Systems for video-based FR attempt to accurately recognize individuals appearing in the field of view of a video camera. Three distinct scenarios can be considered in video-based FR – still-to-video, video-to-still, and video-to-video FR [1]. For example, the still-to-video FR scenario is relevant in watch-list screening applications, where individuals of interest are enrolled to a video surveillance system using reference facial images captured a priori under controlled conditions using a still camera (i.e., mug-shots, passport or driver license photos). Then, facial ROIs extracted from video captured over a distributed network of surveillance cameras are matched against those still ROIs stored during enrollment [2], [3]. In contrast, the video-to-video FR scenario is relevant, for example, in person re-identification applications, where individuals of interest are enrolled to a video surveillance system using reference facial trajectories captured a priori in videos [4], [5], and then matched against facial ROIs extracted from video trajectories captured over a network of cameras.

Recognizing the face of an individual in unconstrained real-world videos remains a challenging task, due in large part variations of facial appearances caused by changes in ambient lighting, poses, expressions, occlusions, scale, blur, etc. [6], [7]. The performance of state-of-the-art systems for video-based FR also declines in real-world environments when a limited number of ROIs is available during enrollment to design a robust facial model [1], [2], [8], [9]. In literature, unified subspace FR methods such as PCA, LDA, Bayesian Face, and metric learning methods cannot simultaneously reduce the complex intra-class variations and enlarge the inter-class discrimination due to their linear nature or shallow structures [7], [10]. Recently, some techniques have been successful for generating representative facial models given a limited number fo reference ROIs, specifically in still-to-video and video-to-still FR [1], [3], [8], [11], [12]. These systems are typically proposed to compensate the lack of representative reference facial ROIs face using multiple face representations, synthetic face generation, and augmenting the target samples in order to enlarge the training set [13], [14]. In addition, recent sparse representation-based classification methods have provided a promising performance by learning additional auxiliary (variational) dictionaries for robust modeling of intra-class variability in video environments [5], [11], [15], [16].

Despite the improvements achieved through the above-mentioned methods, there still exists a significant gap compared to the human visual system [17]. In this paper, deep learning methods are considered to provide robust modeling of intra-class and inter-class variations, and accurate video-based FR [7], [18]. Deep learning methods have been shown to learn effective face representations directly from face images through their deep architecture and hierarchical nonlinear mapping [10], [19], [20], [21], [22]. In particular, to learn a face embedding that can suitably reduce the intra-class variations, as well as, increase the inter-class variations, a triplet-based loss has been utilized with FaceNet [22] in a compact Euclidean space in order to dissociate the negative

978-1-5090-6182-2/17/$31.00 ©2017 IEEE

facial ROIs of other identities from the positive pair of two faces corresponding to the same identity. Similarly, a Trunk-Branch Ensemble CNN (TBE-CNN) model has been proposed in [23] along with an improved triplet loss function to learn blur-insensitive face representations composed of both still face images and artificially blurred faces. This model is an end-to-end network that shares the early- and mid-layer convolutional layers between the trunk (to extract holistic features) and branch (to extract local features) networks to efficiently extract discriminative face representations. The main drawback of the TBE-CNN is that it requires the reliable detection of facial landmarks (that may fail due to occlusion), and thereby increase the complexity to perform in real-time applications.

In this paper, a novel end-to-end ensemble of DCNNs called HaarNet is proposed to efficiently learn robust and discriminative face representations for video-based FR applications. HaarNet consists of a trunk network with three diverging branch networks that are specifically designed to embed facial features, pose, and other distinctive features. The trunk network effectively learns a holistic representation of the face, whereas the branches learn more local and asymmetrical features related to pose or special facial features by means of Haar-like features. Furthermore, to increase the discriminative capabilities of the HaarNet, a second-order statistic regularized triplet-loss is proposed for an end-to-end training process. The proposed triplet-loss function takes advantage of the inter-class and intra-class variations existing in training data to learn more distinctive representations for subjects with similar faces. Finally, a fine-tuning stage is proposed to embed the correlation of facial ROIs stored during enrollment and improve recognition accuracy.

II. HaarNet Architecture

The overall architecture of the proposed HaarNet is presented in Fig. 2. Inspired by [23], this ensemble of deep convolutional neural networks (DCNNs) is composed of a global trunk network along with three branch networks that can effectively learn a representation that is robust to changing capture conditions. As shown in Fig. 2, the trunk is employed to learn the global appearance face representation, whereas three branches diverged from the trunk are designed to learn asymmetrical and more locally distinctive representations.

A. Face embedding:

Similar to [22] and [23], the face embedding is performed using a Haar-like deep neural network. In contrast with [23], instead of fusing the trunk and branch representations to obtain a final face representation using only one fully connected layer, we propose to concatenate the output of trunk and branches to obtain a final representation of the facial ROI. In particular, we propose to utilize three branch networks, where each branch computes one of the Haar-like features illustrated in Fig. 1. As outlined in [24] Haar features have been utilized for face detection to extract distinctive features from faces based on the symmetrical nature of facial components, and on contrast of intensity between adjacent components. In general, these features are calculated by subtracting sum of all pixels in the black areas from the sum of all pixels in the white areas. To avoid information loss, the Haar-like features are calculated by matrix summation, where black matrices are negated. Thus, instead of generating only one value, each Haar-like feature returns a matrix.

In the architecture (see Fig. 2), the trunk network and its three branches share the first two convolutional layers. Then, the first and second branches split the output of Conv2 into two sub-branches, and also apply two inception layers to each sub-branch. Subsequently, the two sub-branches are merged by a subtraction layer to obtain a Haar-like representation for each corresponding branch. Meanwhile, the third branch divides the output of Conv2 into four sub-branches and one inception layer is applied to each of the sub-branches. Eventually, a subtraction layer is exploited to combine those for sub-branches and feed to the fully connected layer. The final representation of the face is obtained by concatenating the output of the trunk and all three Haar-like features.

Fig. 1: Haar-like features used in branch networks.

As illustrated in Fig. 2, the first two convolutional layers (Conv1 and Conv2) extract low-level features representing local information [23]. These two layers share weights between all branches and the trunk. However, since the mid- and high-level features have different properties in each branch and the trunk, the corresponding layers don't share parameters.

The layers and specifications of the trunk network are presented in Table I. For the trunk network, the configuration of GoogLeNet [25] is employed with 18 layers. In order to have a consistent input, all the face images are scaled to 192x192 pixels for all datasets.

TABLE I: Specifications of the trunk network.

Features	Layer type	Kernel size/stride	Output size	Depth
Low-level features	Conv1	7x7/2	96x96x64	1
	Max pooling	2x2/2	48x48x64	0
	Conv2	3x3/1	48x48x192	2
	Max pooling	2x2/2	24x24x192	0
Mid-level features	Inception (3a)	-	24x24x256	2
	Inception (3b)	-	24x24x480	2
	Max pooling	2x2/2	12x12x480	0
High-level features	Inception (4a)	-	12x12x512	2
	Inception (4b)	-	12x12x512	2
	Inception (4c)	-	12x12x512	2
	Inception (4d)	-	12x12x528	2
	Inception (4e)	-	12x12x832	2
	Max pooling	2x2/2	6x6x832	0
	Inception (5a)	-	6x6x832	2
	Inception (5b)	-	6x6x1024	2
	Max pooling	2x2/2	3x3x1024	1
	Dropout	-	3x3x1024	1
	Fully connected	-	256	1

Table II presents the specification of the layers of the three branches of HaarNet, where each branch computes one of

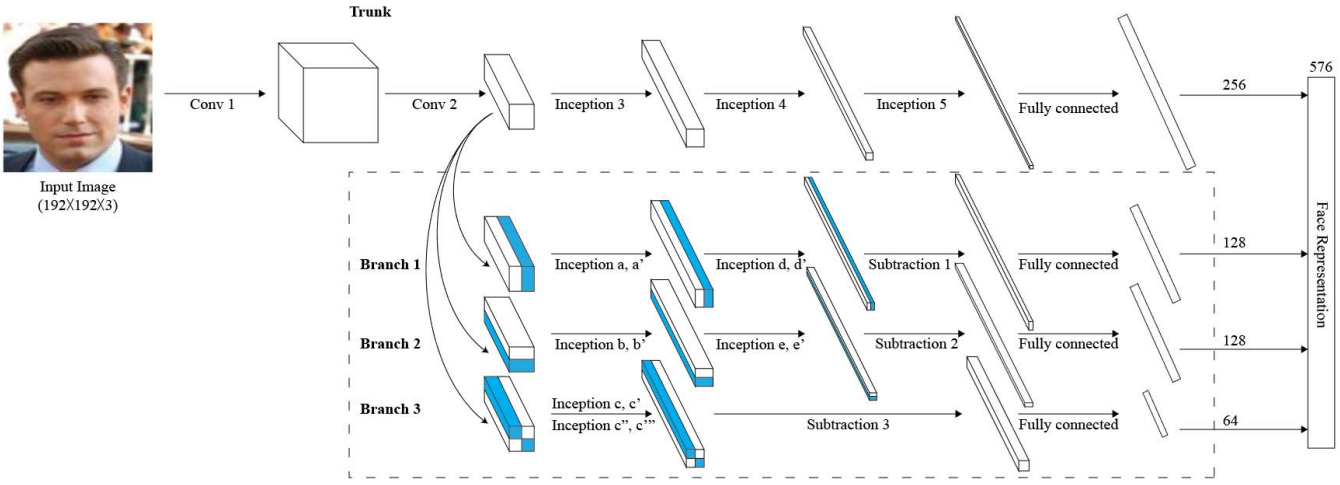


Fig. 2: HaarNet architecture for the trunk and three branches. (Max pooling layers after each inception and convolution layer are not shown for clarity).

the Haar-like features. The specifications of those branches without some layers are marked by a hyphen.

TABLE II: Specifications of the 3 branch networks.

Features	Layer type	Kernel size/stride	Branch1	Branch2	Branch3
Low-level features	Conv1	7x7/2	96x96x64	96x96x64	96x96x64
	Max pooling	2x2/2	48x48x64	48x48x64	48x48x64
	Conv2	3x3/1	48x48x192	48x48x192	48x48x192
	Max pooling	2x2/2	24x24x192	24x24x192	24x24x192
Mid-level features	Inception (a)	-	12x12x480	-	-
	Inception (a')	-	12x12x480	-	-
	Inception (b)	-	-	12x12x480	-
	Inception (b')	-	-	12x12x480	-
	Inception (c)	-	-	-	12x12x480
	Inception (c')	-	-	-	12x12x480
	Inception (c")	-	-	-	12x12x480
	Inception (c"')	-	-	-	12x12x480
	Max pooling	2x2/2	6x6x480	6x6x480	3x3x480
High-level features	Inception (d)	-	6x6x832	-	-
	Inception (d')	-	6x6x832	-	-
	Inception (e)	-	-	6x6x832	-
	Inception (e')	-	-	6x6x832	-
	Max pooling	2x2/2	3x3x832	3x3x832	-
	Dropout	-	3x3x832	3x3x832	3x3x480
	Subtraction 1	-	3x3x832	-	-
	Subtraction 2	-	-	3x3x832	-
	Subtraction 3	-	-	-	3x3x480
	Fully connected	-	128	128	64

B. Second-order statistics regularized loss function:

Recently, deep learning algorithms specialized for FR mostly utilize triplet-loss in order to train the deep architecture and thereby learning a discriminant face representation [22], [23], [26]. However, careful triplet sampling is a crucial step in order to achieve a faster convergence [22]. In addition, employing triplet-loss is challenging since the global distributions of the training samples are neglected in optimization process.

Ding and Tao [23] have shown that by adding a mean distance regularization term to the triplet-loss function, the distinctiveness of the face representation may improve. Fig. 4 illustrates the main idea of the proposed second-order statistics regularization term. In Fig 4 (a), triplet-loss function may suffer from nonuniform inter-class distances that leads to failure of using simple distance measures, such as Euclidean and cosine distances. In this regard (see Fig. 4 (b)), a mean distance regularization term can be added to increase the separation of class representations. On the other hand, representations of some facial ROIs may be confused with representation of the adjacent facial ROIs in the feature space due to high intra-class variations. Fig. 4 (c) shows such a configuration, where the mean representation of the classes are distant from each other but the standard deviations of classes are very high, leading to overlap among class representations. To address this issue, this paper introduces a new term in the loss function to examine the intra-class distribution of the training samples.

Fig. 3 illustrates the training process of the HaarNet using a triplet-loss concept, where a batch of triplets composed of <anchor, positive, negative> is input to the architecture is translated to a face representation.

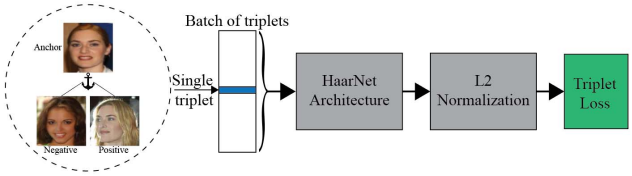


Fig. 3: Processing of triplets to compute the loss function. The network inputs a batch of triplets to the HaarNet architecture followed by an $L2$ Normalization.

As shown in Fig. 3, output of the HaarNet is then $L2$ normalized prior to feed into the triplet-loss function in order to represent faces on a unit hyper-sphere. Let's denote the $L2$ normalized representation of a facial ROI x as $f(x) \in R^d$ where d is the dimension of the face representation.

The triplet constraint can be expressed as a function of the representation of anchor, positive and negative samples as follows [22]:

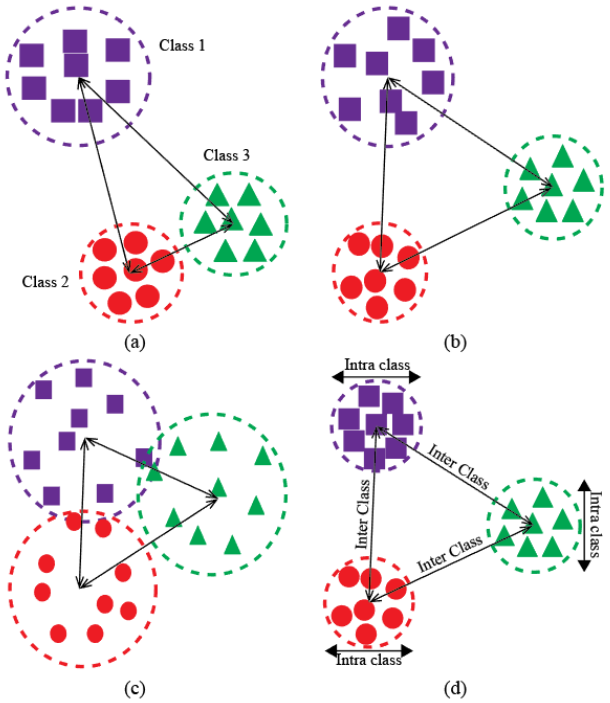


Fig. 4: Illustration of the regularized triple loss principle based on the mean and standard deviation of 3 classes, assuming a 2D representation of the facial ROIs.

$$\|f(x_i^a) - f(x_i^p)\|_2^2 + a < \|f(x_i^a) - f(x_i^n)\|_2^2 \quad (1)$$

where $f(x_i^a)$, $f(x_i^p)$, and $f(x_i^n)$ are the face representations of the anchor, positive, and negative, respectively. All the triplets sampled from the training set should satisfy the constraint. Thus, during training, HaarNet minimizes of the loss function:

$$L_{HaarNet} = \delta_1 L_{triplet} + \delta_2 L_{mean} + \delta_3 L_{std} \quad (2)$$

where δ_i denotes the weight for each term in the loss function. Furthermore, $L_{triplet}$ can be defined based on (1) as follows:

$$L_{triplet} = \frac{1}{2N} \sum_{i=1}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \quad (3)$$

Similar to [23], assuming that the mean distance constraint is $\beta < \|\hat{\mu}_c - \hat{\mu}_c^n\|_2^2$, we define L_{mean} as:

$$L_{mean} = \frac{1}{2P} \sum_{c=1}^{C} \max\left(0,\ \beta - \|\hat{\mu}_c - \hat{\mu}_c^n\|_2^2\right) \quad (4)$$

In addition, we define the standard deviation constraint to be $\sigma_c > \gamma$, where σ_c is the standard deviation of the class c. Therefore, L_{std} can be computed as follows:

$$L_{std} = \frac{1}{M} \sum_{c=1}^{C} \max(0,\ \gamma - \sigma_c) \quad (5)$$

where N, P, and M are the number of samples that violate the triplet, mean distance, and standard deviation constraints, respectively. Likewise, C is the number of subjects in the current batch and α, β,and γ are margins for triplet, mean distance, and standard deviation constraints, respectively. The loss function (2) can be optimized using the regular stochastic gradient descent with momentum similar to [23]. The gradient of loss w.r.t. the facial ROI representation of ith image for subject c (denoted as $f(x_{ci})$) is derived as follows:

$$\frac{\partial L_{std}}{\partial f(x_{ci})} = -\frac{1}{M} \sum_{c=1}^{C} \omega_c \frac{\partial \sigma_c}{\partial f(x_{ci})} \quad (6)$$

where ω_c equals to 1 if the standard deviation constraint is violated, and equals to 0 otherwise. Moreover, the derivative of L_{std} can be computed by applying the chain rule as follows:

$$\frac{\partial \sigma_c}{\partial f(x_{ci})} = \frac{\partial \sqrt{\frac{1}{N_c}\sum_{j=1}^{N_c} \|f(x_{cj}) - \mu_c\|_2^2}}{\partial f(x_{ci})} = \frac{\left[\sum_{j=1}^{N_c} \frac{1}{N_c} \|\mu_c - f(x_{cj})\|_2\right] - \|\mu_c - f(x_{ci})\|_2}{2\sqrt{\frac{1}{N_c}\sum_{j=1}^{N_c} \|f(x_{cj}) - \mu_c\|_2^2}} \quad (7)$$

As shown in Fig. 4 (d), the discriminating power of the face representations can be improved by setting margins such that $\gamma < \beta$. This ensures a high inter-class and a low intra-class variations to increase the overall classification accuracy.

C. Training phase:

Training a network with multiple branches followed by a triplet-loss is tricky and requires careful attention to the details. A multi-stage training approach is hereby proposed to effectively optimize the parameters of the proposed HaarNet. The first three stages are designed for initializing the parameters with a promising approximation prior to employ the triplet-loss function. Moreover, these three stages are beneficial to detect a set of hard triplets from the dataset in order to initiate the triplet-loss training.

In the first stage, the trunk network is trained using a softmax loss, because the softmax function converges much faster than triplet-loss function. During the second stage, each branch is trained separately by fixing the shared parameters and by only optimizing the rest of the parameters. Similar to the first stage, a softmax loss function is used to train each of the branches. Then, the complete network is constructed by assembling the trunk and the three branch networks. The third stage of the training is indeed a fine-tuning stage for the complete network in order to optimize these four components simultaneously. In order to consider the inter- and intra-class variations, the network is trained for several epochs using the hard triplets detected during the previous stages.

D. Recognition process:

The HaarNet generates a 576 dimensional face representations consisting of a 256 dimensional feature extracted from the whole image concatenated with 320 dimensional Haar-like features. This heterogeneous face representation is incompatible with regular distance metrics such as Euclidean or Cosine distances. In order to employ the HaarNet method in a FR setup, we propose to train a fully connected layer followed by a "softmax" which takes two face representations as input and outputs a similarity score between zero and one. This layer is trained on LFW dataset for several epochs after the feature extraction pipeline is completely trained and later is fine-tuned on COX Face DB training set images.

III. Experiments

In this section, several experimental results are shown to evaluate and comparing the performance of the proposed HaarNet against the state-of-the-art video FR systems.

A. Datasets:

Experiments are conducted using challenging datasets designed specifically to video-based FR, LFW, COX Face DB and ChokePoint datasets. Example faces of three datasets used in this paper are presented in Fig. 5 that shows variations in the video ROIs for a specific subject similar to surveillance environments. Noted that LFW dataset has been only employed to train the HaarNet and adjust the network parameters with a large number of faces.

Fig. 5: Examples of LFW (top row), Cox Face DB (middle row) and Chokepoint (bottom row) datasets, where they contain different variations in camera viewpoints, pose, expression, blurriness, and occlusion. The left most column represents high-quality frontal still faces (for Cox Face DB and Chokepoint datasets).

The COX Face DB [1] simulates real-world video surveillance data containing still and video images of 1000 subjects. For each subject, the dataset consists of one high-quality still image and three uncontrolled video clips recorded by low-resolution off-the-shelf cam-coders. These three videos are captured while subjects are walking roughly along an S-shaped path to emulate different poses and facial appearances similar to the real world scenarios. Moreover, these videos are taken from the subjects walking in a large gymnasium with high ceiling, thus, the environment and camera setup approximates the outdoor lighting conditions. In order to evaluate our proposed method on the COX Face DB, we adopted the still-to-video and video-to-still protocols introduced in [1].

The ChokePoint dataset [27] contains a collection of still images and videos for experiments in video-based FR that simulates the real-world surveillance conditions. The dataset contains still images of 25 subjects in portal 1 and 29 subjects in portal 2. In total, the dataset contains 64,204 facial ROIs accurately extracted from the images of 48 video sequences captured using three cameras, in two portals and with subjects entering and leaving the portals. For the comparison w.r.t. still-to-video scenario, we adopted the protocol proposed in [2] using a set of 5 randomly selected subjects of interest.

B. Protocol:

The main challenge in video-based FR is the lack of adequate amount of diversified training data to support training a deep model. Moreover, most of the video FR databases such as COX Face DB [1] and ChokePoint [27] contain a limited number of subjects and typically suffer from the lack of diversity in the video ROIs, specially diversity in different facial appearances with respect to the still ROIs. Following [23], we synthetically generate a video-like dataset from an existing dataset that contains a large number diverse subjects. In this paper, motion blur and out-of-focus blur are emulated by adding noise to the original image. We further augment the artificially generated dataset by applying several transformations. For each artificially generated video ROI, we construct a set of images through the following transformations: shearing, mirroring, rotating, translating. These transformations help to enrich the artificial video dataset by simulating different viewpoints. Moreover, for each transformation we generate two images by applying two different levels of down-sampling followed by an up-sampling. The subsampling emulates different scales (distance from the camera) and also helps to embed the low-quality nature of the video facial ROIs.

In our experiments, HaarNet is trained on the Labeled Faces in the Wild (LFW) dataset [28]. In order to emulate video ROIs, an artificial video dataset containing roughly 8.2 million video face ROIs is generated. Additionally, the network is fine-tuned using the COX Face DB in order to embed the camera field of view information in the network. So far, the network has no knowledge about the subjects of interest enrolled to the system. In order to embed this knowledge, final fine-tuning round is employed over still ROIs of subjects. During this phase, another artificial video dataset is generated using only the still face ROIs by following the aforementioned process. The objective of the fine-tuning process is to train the network in order to acquire knowledge about similarities and dissimilarities among the subjects of interest based on their still and synthesized video ROIs. Fig. 6 presents some of augmented images generated for the fine-tuning stage.

For COX Face DB, experiments conducted using the list of training and testing images provided as suggested by [1]. Thus, we used 300 subjects for training and 700 subjects for

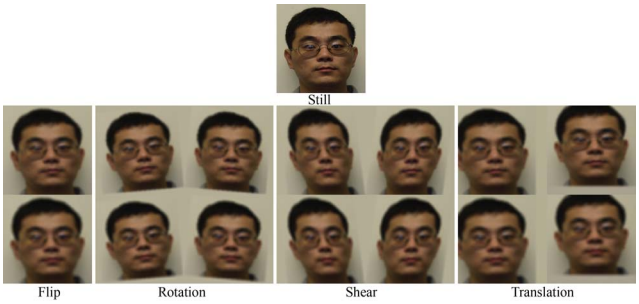


Fig. 6: Sample of augmented facial images generated from a Chokepoint still for the fine-tuning stage. The first row represents facial ROIs generated by one level of sub-sampling, while the second row represents images generate by two levels of subsampling followed by upsampling.

testing over a course of independent 10 replications, where the training and testing subjects are randomly selected for each replication. Training is performed using all the still and video facial ROIs of the 300 subjects, while for testing, the high resolution still images from the remaining 700 subjects are used for enrollment as the gallery set and the probe set contains the facial ROIs of the three video clips from the corresponding 700 subjects. Therefore, each probe is matched against all the gallery images and rank-1 recognition performance is reported for still-to-video FR scenario. Moreover, for video-to-still FR, the gallery set and probe sets are swapped and each still image is compared against all the video facial ROIs of those 700 subjects. Furthermore, for fine-tuning, we used the still images of the 700 test subjects to generate the artificial video facial ROI dataset. This allows the network to gain knowledge about the subjects of interest. However, for the sake of fairness in comparisons, we provide the results of our framework with and without this final fine-tuning stage.

In the experiment with ChokePoint dataset, the instructions from [2] is followed in order to perform still-to-video FR. In this experiment, 5 subjects are selected randomly to be enrolled in the system. On the other hand, the probe set contains all video ROIs of these subjects along with 10 unknown subjects appeared in the operational scene, while their still images are not included in the gallery. In the final experiment, all the video ROIs in the ChokePoint dataset are used as probe set and the gallery contains the 27 high-quality controlled images. This experiment is similar to the aforementioned experiment on the COX Face DB with more video facial ROIs per person.

In order to have a consistent neural network, we resized all the facial ROIs from these two datasets to 192x192 pixels. Moreover, the LFW dataset was used to train the network. First, the trunk was trained for 30 epochs using a softmax, then each branch is trained for 20 epochs using a softmax loss. Subsequently, the complete network is assembled and trained by adopting a softmax loss function for 15 epochs. Finally, HaarNet is trained using the proposed regularized triplet loss for extra 15 epochs. The similarity measure network is then trained using the face representations obtained from the HaarNet on the LFW dataset. Thereafter, the similarity measure network is trained on the 300 training subjects from the COX Face DB for another 5 epochs. On the top of all these training stages, there is an additional fine-tuning stage using the artificially simulated video images based on the 700 images of the gallery, where we only fine-tune the final classification layer. The network trained on COX Face DB was used to assess on the ChokePoint dataset with an exception that the fine-tuning is performed using the simulated images generated from the still images of the ChokePoint dataset. Noted that in this experiment, the network has no knowledge about the background subjects and thus, this experiment would be a more realistic challenge than the protocol suggested for experimenting on the COX Face DB. The parameters and their corresponding values of the proposed triplet loss function are presented in Table 3.

TABLE III: Parameters of the regularized triplet-loss function used during the training process.

Parameter	α	β	γ	δ_1	δ_2	δ_3
Value	0.2	0.3	0.2	0.5	0.3	0.2

C. Performance metrics:

In the experiments, rank-1 recognition is reported to compare the performance of the proposed HaarNet against the state-of-the-art video FR systems in a face identification scenario, while ROC curve is presented to perform a comparison under a face verification scenario. Rank-1 recognition is computed based on the highest response in the gallery (among enrolled subjects) for the given probe ROI. The rank-1 recognition and ROC curve of the HaarNet are compared against point-to-set correlation learning (PSCL) [1], learning euclidean-to-riemannian metric (LERM) [12], and TBE-CNN [23] on COX Face Database and ensemble-based method (EBM) [2], and [29] on the ChokePoint dataset.

Receiver Operating Characteristic (ROC) curve and the area under the ROC curve are more appropriate way for comparing methods in open-set authentication scenarios as found in video surveillance applications [1]. The ROC space is defined as False Negative Rate (FNR) along x-axis and True Positive Rate (TPR) along the y-axis. TPR is the ratio of correctly classified facial ROIs as a target subject in the gallery over number of all probes with a corresponding ROI in the gallery. On the other hand, FNR is the ratio of incorrectly labeled probes as one of the target subjects over the number of non-target probes. Area Under the ROC Curve (AUC) is a well-known global measure of detection performance and can be interpreted as the probability of the correct classification over the range of TPR and FPR [2].

D. Results

Table IV presents Rank-1 accuracy of the proposed HaarNet and baseline systems on the COX Face DB.

TABLE IV: Rank-1 accuracy for still-to-video FR over the COX Face DB.

FR systems	Video1	Video2	Video3
PSCL [1]	36.39±1.61	30.87±1.77	50.96±1.44
LERM [12]	49.07±1.53	44.16±0.94	63.83±1.58
TBE-CNN [23]	88.24±0.45	87.86±0.85	95.74±0.67
HaarNet	89.31±0.94	87.90±0.60	97.01±1.65
HaarNet + FT	98.86±0.37	97.58±0.77	98.97±0.15

As shown in Table IV, the proposed method significantly outperforms hand-crafted feature extraction methods. By exploiting Haar-like features along with the novel triplet-loss function, the HaarNet can provide higher level of performance compared with the existing deep learning methods. Table IV also shows an additional improvement in rank-1 accuracy after the proposed fine-tuning (HaarNet + FT). The presented results confirm that most of the existing still-to-video FR methods fail to convey the knowledge embedded in the still images. However, the proposed fine-tuning stage efficiently encodes the still images in the gallery to learn the similarities and dissimilarities among the subjects of interest. Moreover, by learning the facial appearance of the subjects of interest, the proposed data augmentation proved to be effective in reducing false negatives.

Table V shows the rank-1 accuracy for video-to-still FR in comparison with state-of-the-art FR methods. In this scenario, each still image is compared against all the video sequences. Due to existence of multiple video facial ROIs in the gallery, a higher accuracy than still-to-video FR scenario is expected. As shown in Table V, the proposed HaarNet with fine-tuning surpasses the state-of-the-art methods for video-to-still FR.

TABLE V: Rank-1 recognition for video-to-still FR over the COX Face DB.

FR systems	Video1	Video2	Video3
PSCL [1]	38.60±1.39	33.20±1.77	53.26±0.80
LERM [12]	45.71±2.05	42.80±1.86	58.37±3.31
TBE-CNN [23]	93.57±0.65	93.96±0.51	98.96±0.17
HaarNet	92.73±1.93	93.57±1.62	97.48±1.54
HaarNet + FT	98.26±0.49	95.27±0.12	99.26±0.69

Amongst the state-of-the-art methods, TBE-CNN is the most competitive one after the proposed HaarNet. However, as shown in Table VI, HaarNet has a significantly lower computational complexity. Since both the TBE-CNN and HaarNet are based on GoogLeNet, the trunk network requires 5,798K parameters, while HaarNet contains 3 branches and TBE-CNN considers 7 branches for each face landmark, respectively. Thus, the proposed HaarNet is more efficient in terms of the number of parameters.

Fig. 7 shows ROC curves for HaarNet, as well as, for PSCL [1] and LERM [12] for each camera, separately. As shown in this figure, the AUC accuracy for HaarNet is larger than others.

For evaluation on the Chokepoint dataset, we adapted the network trained on COX Face DB and tested it without any modifications on the Chokepoint dataset. Then, we fine-tuned the network using simulated video ROIs augmented from the still images of the five subjects of interest. The performance of the HaarNet against EBM [2] is presented in Table VII, where area under precision-recall (AUPR) curve is considered as the performance metric. AUPR is used to measure the performance under the imbalanced data circumstances, where the space is defined by TPR (recall) and precision. Precision is the ratio of true positives over the sum of true positives and false positives.

TABLE VI: The comparison of complexity (number of parameters that need to be estimated) by TBE-CNN and HaarNet architectures.

FR systems	Number of parameters		
	Trunk	Branch	Trunk + Branch
TBE-CNN [23]	5,798K	5,798K x 7	46.4M
HaarNet	5,798K	(3,338K x 2) + 654K	13.1M

TABLE VII: Average AUPR for videos of the Chokepoint.

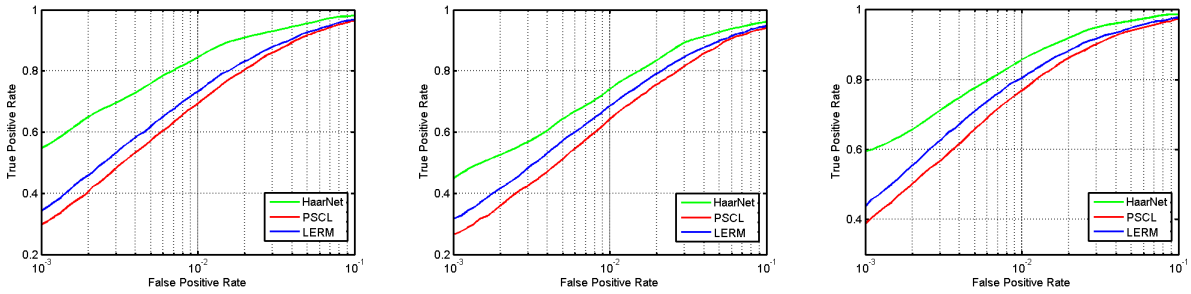
FR systems	EBM [2]	HaarNet	HaarNet + FT
AUPR	99.24±0.38	95.57±1.12	99.36±0.59

It is worth noting that, EBM [2] implements a complex individual-specific ensemble of classifiers for each subject of interest using multiple face representation, while HaarNet benefits from a deep specialized neural network.

The final experiment is performed using the the protocol adopted by [29], where the training is performed on a separate dataset (in our case, COX Face DB) and tested on all of the face images in the Chokepoint dataset. Therefore, all video ROIs are considered as probes and all still ROIs are registered in the gallery. However, the rank-1 accuracy rate documented in [29] for still-to-video FR is 62.7%, whereas we could reach up to 84.92% before fine-tuning. Moreover, by performing the aforementioned fine-tuning stage, HaarNet could achieve 96.12% rank-1 accuracy on all the probe images in the dataset.

IV. Conclusion

This paper presents a deep neural network that can learn face representations for each target individual for accurate video-based FR systems. The proposed HaarNet architecture employs an ensemble of DCNN in order to obtain a discriminative embedding of the facial ROI. In particular, the network utilizes a trunk that shares weights with branches and each branch is trained to compute features similar to Haar-like features. The trunk is specialized for matching the global appearance of the face, while the branches embed informative features, such as pose, and asymmetrical facial features of the subjects. In order to effectively train the proposed deep architecture, a novel regularized triplet-loss function was proposed to generate face embedding with high similarity among intra-class samples, while maximizing the inter-class variations. In order to address the single training sample issue, synthetic facial images were generated form still images of the subjects of interest using different transformations, such as shearing,



(a) camera 1 (b) camera 2 (c) camera 3

Fig. 7: ROC curves of HaarNet and baseline FR methods for videos of each camera in the Cox Face DB.

rotation, translation, and subsampling. Finally, the network was fine-tuned over the simulated video ROIs in order to utilize the knowledge existing in still images in the gallery set for higher recognition accuracy.

Several experiments were conducted to evaluate the performance of the proposed HaarNet under different real-world scenarios, such as still-to-video FR. The results obtained over COX Face DB and Chokepoint data indicate a convincingly higher level of accuracy of HaarNet, yet a lower complexity against state-of-the-art FR systems, even when the gallery set contains a large number of subjects. In order to achieve a higher level of performance, future research should focus on utilizing temporal information, where facial ROIs can be tracked over frames to accumulate the predictions over time. Thus, the combination of face detection, tracking, and classification in a unified deep learning-based network would lead to a robust spatio-temporal suitable for real-world video surveillance applications.

REFERENCES

[1] Z. Huang, S. Shan, R. Wang, H. Zhang, S. Lao, A. Kuerban, and X. Chen, "A benchmark and comparative study of video-based face recognition on cox face database," *IEEE Trans on Image Processing*, vol. 24, no. 12, pp. 5967–5981, 2015.

[2] S. Bashbaghi, E. Granger, R. Sabourin, and G. A. Bilodeau, "Watch-list screening using ensembles based on multiple face representations," in *ICPR*, 2014.

[3] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg, "Matchnet: Unifying feature and metric learning for patch-based matching," in *CVPR*, 2015.

[4] M. De-La-Torre, E. Granger, P. V. Radtke, R. Sabourin, and D. O. Gorodnichy, "Partially-supervised learning from facial trajectories for face recognition in video surveillance," *Information Fusion*, vol. 24, pp. 31 – 53, 2015.

[5] H. Xu, J. Zheng, A. Alavi, and R. Chellappa, "Learning a structured dictionary for video-based face recognition," in *WACV*, 2016.

[6] A. K. Jain, K. Nandakumar, and A. Ross, "50 years of biometric research: Accomplishments, challenges, and opportunities," *Pattern Recognition Letters*, vol. 79, pp. 80 – 105, 2016.

[7] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *NIPS*, 2014.

[8] S. Bashbaghi, E. Granger, R. Sabourin, and G. A. Bilodeau, "Ensembles of exemplar-svms for video face recognition from a single sample per person," in *AVSS*, 2015.

[9] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *CVPR*, 2008.

[10] Y. Sun, X. Wang, and X. Tang, "Hybrid deep learning for face verification," in *ICCV*, 2013.

[11] W. Deng, J. Hu, and J. Guo, "Extended src: Undersampled face recognition via intraclass variant dictionary," *IEEE Trans on PAMI*, vol. 34, no. 9, pp. 1864–1870, 2012.

[12] Z. Huang, R. Wang, S. Shan, and X. Chen, "Learning euclidean-to-riemannian metric for point-to-set classification," in *CVPR*, 2014.

[13] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *CVPR*, 2015.

[14] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau, "Robust watch-list screening using dynamic ensembles of svms based on multiple face representations," *Machine Vision and Applications*, vol. 28, no. 1, pp. 219–241, 2017.

[15] D. Zhang, Y. Xu, and W. Zuo, "Sparse representation-based methods for face recognition," in *Discriminative Learning in Biometrics*. Springer, 2016, pp. 199–214.

[16] F. Nourbakhsh, E. Granger, and G. Fumera, "An extended sparse classification framework for domain adaptation in video surveillance," in *ACCV, Workshop on Human Identification for Surveillance*, 2016.

[17] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*, 2014.

[18] S. Gao, Y. Zhang, K. Jia, J. Lu, and Y. Zhang, "Single sample face recognition via learning deep supervised autoencoders," *IEEE Trans on Information Forensics and Security*, vol. 10, pp. 2108–2118, 2015.

[19] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *CVPR*, 2014.

[20] R. Chellappa, J. Chen, R. Ranjan, S. Sankaranarayanan, A. Kumar, V. M. Patel, and C. D. Castillo, "Towards the design of an end-to-end automated system for image and video-based recognition," *CoRR*, vol. abs/1601.07883, 2016.

[21] G. B. Huang, H. Lee, and E. Learned-Miller, "Learning hierarchical representations for face verification with convolutional deep belief networks," in *CVPR*, 2012.

[22] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015.

[23] C. Ding and D. Tao, "Trunk-branch ensemble convolutional neural networks for video-based face recognition," *CoRR*, vol. abs/1607.05427, 2016.

[24] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, 2004.

[25] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[26] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *CVPR*, June 2014.

[27] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell, "Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition," in *CVPR WORKSHOPS*, 2011.

[28] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[29] S. P. Mudunuri and S. Biswas, "Low resolution face recognition across variations in pose and illumination," *IEEE Trans on PAMI*, vol. 38, no. 5, pp. 1034–1040, 2016.