# CNNs with Cross-Correlation Matching for Face Recognition in Video Surveillance Using a Single Training Sample Per Person

Mostafa Parchami[1], Saman Bashbaghi[2] and Eric Granger[2]
[1]Computer Science and Engineering Dept., University of Texas at Arlington, TX, USA
[2]École de technologie supérieure, Université du Québec, Montreal, Canada
mostafa.parchami@mavs.uta.edu, bashbaghi@livia.etsmtl.ca and eric.granger@etsmtl.ca

## Abstract

*In video surveillance, face recognition (FR) systems seek to detect individuals of interest appearing over a distributed network of cameras. Still-to-video FR systems match faces captured in videos under challenging conditions against facial models, often designed using one reference still per individual. Although CNNs can achieve among the highest levels of accuracy in many real-world FR applications, state-of-the-art CNNs that are suitable for still-to-video FR, like trunk-branch ensemble (TBE) CNNs, represent complex solutions for real-time applications. In this paper, an efficient CNN architecture is proposed for accurate still-to-video FR from a single reference still. The CCM-CNN is based on new cross-correlation matching (CCM) and triplet-loss optimization methods that provide discriminant face representations. The matching pipeline exploits a matrix Hadamard product followed by a fully connected layer inspired by adaptive weighted cross-correlation. A triplet-based training approach is proposed to optimize the CCM-CNN parameters such that the inter-class variations are increased, while enhancing robustness to intra-class variations. To further improve robustness, the network is fine-tuned using synthetically-generated faces based on still and videos of non-target individuals. Experiments on videos from the COX Face and Chokepoint datasets indicate that the CCM-CNN can achieve a high level of accuracy that is comparable to TBE-CNN and HaarNet, but with a significantly lower time and memory complexity. It may therefore represent the better trade-off between accuracy and complexity for real-time video surveillance applications.*

## 1. Introduction

FR is widely used in applications of law enforcement, forensics, biometrics and surveillance. In video surveillance applications, FR systems seek to recognize target individuals of interest appearing in unconstrained scenes based on their facial appearance [6, 12, 21]. Each face captured over distributed network of video cameras is segmented into a region of interest (ROI), and the pattern extracted from the ROI is matched against the facial models designed a priori for target individuals. When a person appears in a camera field of view, their face is initially detected and tracked over multiple frames, and the matching scores of each face model are integrated along a facial trajectory for robust spatio-temporal FR [1]. Captured in uncontrolled conditions, these faces may vary considerably according to pose, illumination, occlusion, blur, scale, resolution, expression, etc. [2, 5, 15]. The computational complexity is also an important consideration because of the growing number of cameras, and the processing time of state-of-the-art face detection, tracking and matching algorithms.

Watch-list screening is a common yet challenging application in video surveillance. In this case, still-to-video FR systems are employed to detect faces appearing in live or archived videos that correspond to the face of any person stored in a watch-list[1]. The facial model of target individuals are often designed using a single reference image or mugshot captured from a still camera under controlled conditions [2]. In pattern recognition literature, this challenging situation is referred to as a single sample per person (SSPP) problem [1]. Accordingly, the performance of still-to-video FR systems declines in complex real-world environments due to the limited robustness of facial models to intra-class variations [2, 5]. To improve robustness, several approaches have been proposed to generate synthetic faces, to extract multiple representations, and to exploit auxiliary data to enlarge the training set [6, 10, 11]. For instance, given the reference still of a target individual, morphing and 3D reconstruction methods have been proposed to generate synthetic facial images under various capture conditions [7, 14]. Local face matching systems, where patches from a facial ROI are extracted using different descriptors and

[1] In this paper, we focus on the face matching process in these FR systems, which is analogous to face verification. Future research will extend this process to spatio-temporal FR over consecutive frames.

978-1-5386-2939-0/17/$31.00 ©2017 IEEE

feature subspaces have also proposed to generate multiple diverse face representations [2, 1]. Sparse representation-based classifiers have also been proposed that train variational dictionaries to improve robustness [4, 16, 21].

Although the aforementioned methods can improve performance, current systems for still-to-video FR provide a low-level of accuracy in real-world watch-list screening [2, 10]. Recently, deep CNNs have been shown to achieve a high-level of accuracy in many FR applications, where effective facial representations are learned directly from large-scale datasets [5, 19]. For SSPP problems, triplet-based loss has recently been exploited in [5, 17, 18] to learn face embeddings, where the loss seeks to discriminate between the positive pair of matching facial ROIs from the negative non-matching facial ROI. In addition, CNNs like the Trunk-Branch Ensemble CNN [5] and HaarNet [17] can further improve robustness to variations in facial appearance. The trunk network extracts features from the global appearance of faces (holistic representation), while branch networks effectively embed asymmetrical and complex facial traits (local overlapping/non-overlapping patch representations) to handle variations in pose and occlusion. For instance, HaarNet employs 3 branch networks based on Haar-like features, along with a regularized triplet-loss function. However, these specialized CNNs represent complex solutions for real-time FR applications [3].

In this paper, an efficient CNN architecture is proposed for accurate still-to-video FR from a single reference facial ROI per target individual. Based on a novel pair-wise cross-correlation matching (CCM) and a robust facial representation learned through triplet-loss optimization, the proposed CCM-CNN architecture is a fast and compact network (requires few network branches, layers and parameters). The contributions of this paper are threefold. First, the matching pipeline exploits a matrix Hadamard product followed by a fully connected layer that simulates the adaptive weighted cross-correlation technique [8]. Second, a novel triplet-based optimization approach is proposed to learn discriminant facial representations based on triplets containing the positive, negative video ROIs and the corresponding still ROI. In particular, the similarity between the representations of positive video ROIs and the reference still ROI is enhanced, while the similarity between negative video ROIs and the both reference still and positive video ROIs is increased. Finally, to further improve robustness of facial models, the CCM-CNN fine-tuning process incorporates diverse knowledge by generating synthetic faces based on still and video ROIs of non-target individuals. The accuracy and complexity of the proposed CCM-CNN is compared with state-of-the-art FR systems on videos from the COX Face and Chokepoint datasets.

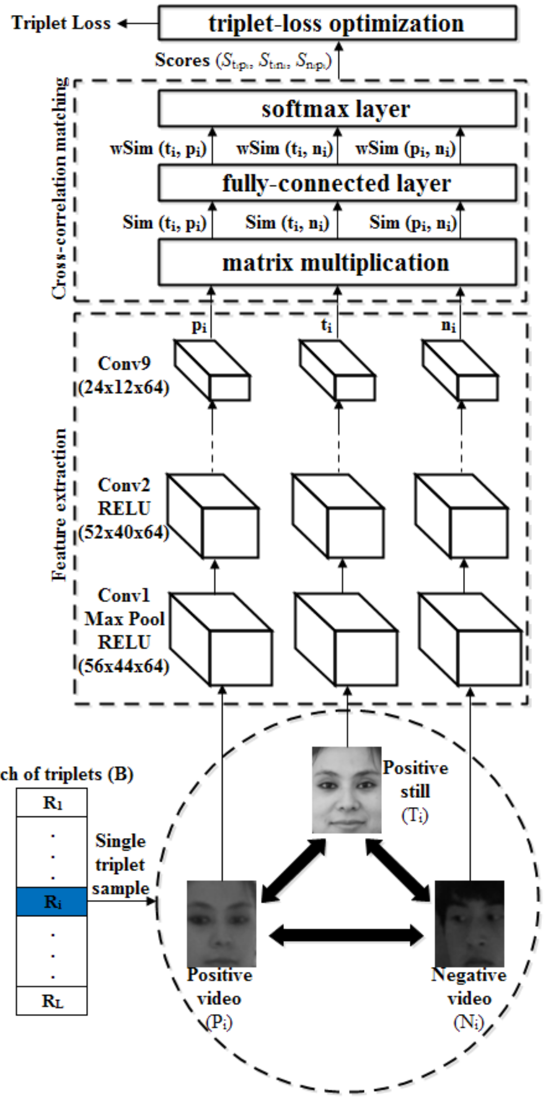


Figure 1: Training pipeline of the proposed CCM-CNN.

## 2. CNNs with Cross-Correlation Matching

As shown in Figure 1, the proposed CCM-CNN learns a robust facial representation by iterating over a batch of training triplets $B = \{R_1, R_2, ..., R_L\} = \{(T_1, P_1, N_1), \ldots, (T_L, P_L, N_L)\}$, where $L$ is the batch size, and each triplet $R_i$ contains a still ROI $T_i$ along with a corresponding positive ROI $P_i$ and a negative ROI $N_i$ from opertional videos. This architecture is inspired by Siamese networks containing identical subnetworks with the same configurations, parameters and weights. Therefore, fewer parameters are required for training that can avoid overfitting. The CCM-CNN consists of three main components – feature extraction, cross-correlation matching and triplet-loss optimization. The feature extraction pipeline extracts discriminative feature maps from ROIs that are similar for two images of the same person under different capture conditions (e.g., illumination and pose). The cross-correlation matching component inputs feature maps extracted from the ROIs and calculates the likelihood of the faces belonging to the same person. Finally, triplet-loss optimization computes a novel loss function to maximize similarity of the still ROIs and their respective positive samples in the batch, while minimizing similarity between still ROIs and their negative ROIs, as well as, positive and negative ROIs.

## 2.1. Feature Extraction Pipeline:

To obtain a discriminative matrix representation of facial ROIs and perform cross-correlation matching, a specialised version of the stereo matching pipeline in [13] is adopted. Despite differences in the domain between reference target still ROIs and target/non-target video ROIs, the proposed CCM-CNN can effectively extract discriminent features.

As shown in Figure 1, feature extraction is carried out by 3 identical subnetworks for still, positive and negative faces. These subnetworks process three input faces and the weights are shared across them. Each subnetwork consists of 9 convolutional layers each followed by a spatial batch normalization, drop-out, and RELU layers. Contrary to former convolutional layers, the last convolutional layer is not followed by a RELU in order to maintain the representativeness of the final feature map and to avoid losing informative data for the matching stage. Moreover, a single max-pooling layer is added after the first convolution layer to increase the robustness to small translation of faces in the ROI. Nevertheless, most of the state-of-the-art systems for SSPP rely on accurate face alignment and do not consider any possible displacement which highly affects the local matching [10].

In the proposed CCM-CNN, all three feature extraction pipelines (see Figure 1) share the same set of parameters. This ensures that the features extracted from target still ($\mathbf{t_i}$), positive ($\mathbf{p_i}$) and negative ($\mathbf{n_i}$) are consistent and comparable. Each convolutional layer has 64 filters of size 5x5 without padding. Thus, given the input size of 120x96, the output of each branch is of size $N_f$ = 24x12x64 features.

## 2.2. Cross-Correlation Matching:

After extracting features from the still and video ROIs, a pixel-based matching method is employed to effectively compare these feature maps and measure the matching similarity. The comparison in the proposed system has three stages: matrix Hadamard product, fully connected neural network, and finally a softmax. There are several approaches to join the two branches of the deep network. One basic approach is to concatenate the two feature vectors and form a single long vector as input to the fully connected network [6, 22]. However, instead of feature concatenation, the feature maps representing the ROIs are multiplied with each other to encode pixel-wise correlation between each pair of ROI in the triplet. This approach eliminates the complexity of matching by replacing the concatenation with a simple element-wise matrix multiplication and directly encodes similarity as opposed to let the network learn how to match input concatenated feature vectors.

The matrix Hadamard product is exploited to simulate cross-correlation, where Hadamard product of the two matrices provides a single feature map that represents the similarity of the two ROIs. For example, the similarity $Sim(\mathbf{t}_i, \mathbf{p}_i)$ and cross-correlation $\mathbf{w}Sim(\mathbf{t}_i, \mathbf{p}_i)$ of still $\mathbf{t}_i$ and positive $\mathbf{p}_i$ feature maps is computed as follows, respectively, using matrix Hadamard product:

$$Sim(\mathbf{t}_i, \mathbf{p}_i) = (\mathbf{t}_i \odot \mathbf{p}_j) \quad (1)$$

$$wSim(\mathbf{t}_i, \mathbf{p}_i) = \omega_m \cdot RELU(\omega_n \cdot Sim(\mathbf{t}_i, \mathbf{p}_i) + \mathbf{b}_n) + \mathbf{b}_m \quad (2)$$

where $\omega_m$, $\omega_n$, $\mathbf{b}_m$ and $\mathbf{b}_n$ are the weights and biases of the two fully-connected layers applied to the vectorized output of the matrix multiplication. Furthermore, a softmax layer is applied to obtain a probability-like similarity score for each of the two classes (match and non-match).

## 2.3. Triplet-Loss Optimization:

A multi-stage approach is considered to efficiently train the proposed CCM-CNN based on reference stills ROI per individual and operational videos. To that end, pre-training is performed using a large generic FR dataset, and a domain specific dataset for still-to-video FR is used for fine-tuning.

During pre-training, the CNN is mostly focused on optimizing the feature extraction pipeline. To obtain a high level of accuracy, the proposed network is trained using a set of triplets that are difficult to classify. To that end, a set of matching and non-matching images is selected from the Labeled Faces in the Wild (LFW) [9]. Images from this set are augmented to roughly 1.3M training triplets. In order to consistently update the set of training triplets, the on-line triplet sampling method [19] is used for 50 epochs.

Deep NNs are typically trained by back-propagating the loss calculated by comparing the network's output prediction with the ground-truth label. In contrast with FaceNet [19], we propose a pair-wise triplet-loss optimization function to effectively train the proposed network. In order to adapt the network for pairwise triplet-based optimization, it is modified by incorporating additional feature extraction branches.

Each batch contains several triplets, and, for each triplet, the network seeks to learn the correct classification. During the training, each branch of the feature extraction pipeline is assigned to a component of the triplet – the main branch is responsible for processing the reference still ROI, while the positive (negative) branch extracts features from the positive (negative) video ROI of the triplet. Moreover, the cross correlation matching pipeline is modified to benefit from the triplets by introducing an Euclidean loss layer followed by softmax which computes the similarity for each pair of ROIs in the triplet. The proposed loss layer is exploited to compute the overall loss of the network as follows:

$$\text{Triplet Loss} = \frac{1}{L} \sum_{R_i \in B} \sqrt{(1 - S_{t_i p_i})^2 + S^2_{t_i n_i} + S^2_{n_i p_i}} \quad (3)$$

where $S_{tp}$, $S_{tn}$, and $S_{np}$ are the similarity scores from cross-correlation matching between (1) the reference (positive) still ROI and positive video ROI, (2) still ROI and negative video ROI, and (3) negative and positive video ROIs of the triplet, respectively, computed using the aforementioned approach. During operations (once the network training is completed) the additional feature extraction branch (negative branch, N) is removed from the network, and only the still and the positive branches (P) are taken into account. Thus, the main branch (T) extracts features from a reference still ROIs, while the positive branch extracts features from the probe video ROI to determine whether they belong to the same person.

During fine-tuning, CCM-CNN acquires knowledge on the similarities and dissimilarities between the target individuals of interest enrolled to the system. So far, the network is pre-trained on facials ROIs that are not expected to be seen during operations. In order to improve the robustness of facial models intra-class variation, the network is fine-tuned with synthetic facial ROIs generated from the high-quality still ROIs that account for the operation domain. For each still image, a set of augmented images are generated using different transformations, such as shearing, mirroring, rotating and translating the original still image. Then, two levels of sub-sampling are applied to each of these images to obtain two images per transformation operation. While shearing, mirroring, rotating and translating increases the diversity in the viewpoint and facial appearance, sub-sampling encodes different distances from cameras, as well as, the quality of face ROI. After sub-sampling, all images are up-scaled to a common size, where the still ROIs resembles the ROIs extracted from operational videos.

As during pre-training, triplet-based optimization is also employed. The same sampling and training strategies are applied to effectively fine-tune the network. In contrast with the pre-training, the focus of the fine-tuning stage is to learn dissimilarities between the subjects of interest and thus the parameters of the feature extraction pipelines are fixed. In this case, fine-tuning on only a moderate number of epochs can significantly improve accuracy.

## 3. Experimental Methodology

The experiments are conducted using videos from two challenging datasets specifically designed for video-based FR: COX Face DB [10], Chokepoint [20] datasets. Both can emulate real-world still-to-video FR scenario, where their main characteristics are that they contain a high-quality still face images captured under controlled condition (with the same still camera), and low-quality surveillance videos for each subject captured under uncontrolled conditions (with surveillance cameras). Videos are captured over a distributed network of cameras that covers a range of variations (changes in, e.g., pose, illumination, blur, scale). The COX Face DB is constructed with participation of 1000 subjects. The dataset consists of one high quality still image and three uncontrolled video clips captured by three different off-the-shelf low-resolution cam-coders for each subject. The Chokeoint dataset contains stills and videos of 29 subjects walking through different portals. In total, the dataset contains 64,204 facoal ROIs extracted from 48 video sequences captured using three cameras locating above the portals and four different monthly sessions, with subjects entering and leaving the scene.

To compare the proposed network with state-of-the-art FR systems, standard experimental setups suggested by [2], [10] and [15] are followed. For COX Face DB experiments, the same training subjects are selected to train the feature extraction pipeline, where 300 subjects are considered for training and 700 subjects for testing over a course of 10 iterations with random selection of training and testing subjects for each iteration. During training, all the still and video face ROIs of the 300 subjects are adopted. On the other hand, the high-resolution still images from the remaining 700 subjects are used during testing as the gallery set and the probe set contains the face ROIs of the videos from the corresponding 700 subjects. Thus, each probe is compared against all the gallery images and rank-1 recognition is reported as the accuracy of still-to-video FR system. For fine-tuning, the still images of the 700 test subjects are used to perform augmentation of still faces according to video facial ROIs in videos. This allows the network to gain knowledge about the probable appearance of people and contextual information within the surveillance environments.

For Chokepoint experiments, 5 subjects of interest are randomly selected and thus the gallery set contains only the still ROIs of these 5 subjects. On the other hand, the probe set contains all video ROIs of these subjects along with videos of 10 unknown subjects that appeared in the operational environment. The pre-trained network that was already trained on COX Face DB is used to process Chokepoint videos. Apart from that, fine-tuning is performed using the stills and videos of remaining individuals in the Chokepoint dataset.

All the facial ROIs extracted from still and videos are scaled to a common size of 120x96 pixels. The proposed network is implemented using Torch 7.0 deep learning framework. The training is performed for 30 epochs using the training data gathered from the COX Face DB. Also, for the fine-tuning purpose on the COX Face DB, the network is trained for an additional 5 epochs on the augmented faces synthesized from the still images. In order to fine-tune the network for Chokepoint dataset, the network is trained for 3 epochs on faces generated from the still and video images from the same dataset. Rank-1 recognition accuracy and ROC curves of the proposed network is compared against Point-to-Set Correlation Learning (PSCL) [10], Learning Euclidean to Riemannian Metric (LERM) [11], Trunk-Branch Ensemble (TBE-CNN) [5] and HaarNet [17] on the COX Face DB data, and also against dynamic ensembles of SVMs (EoSVM) [2] and extended SRC with domain adaptation (ESRC-DA) [16] on the Chokepoint data. CCM-CNN is also compared with VGG-Face [18] on rank1 results. Given the imbalanced data, the area under precision-recall (AUPR) curve is used to measure the global performance, where it is defined by precision versus TPR as Recall.

## 4. Results and Discussion

Table 1 shows the rank-1 accuracy of the proposed CCM-CNN against state-of-the-art video-based FR system on COX Face DB videos. Rank-1 accuracy is computed based on the highest response in the reference still faces set for the given probe face ROI from a video. As shown in Table 1, CCM-CNN significantly outperforms PSCL and LERM (that exploit hand-crafted features), and provides a Rank-1 accuracy comparable to TBE-CNN and HaarNet. However, TBE-CNN and HaarNet employ an ensemble of CNNs to achieve a higher recognition accuracy. Despite the elegant TBE-CNN and HaarNet architectures, the proposed network can achieve competitive performance with a simpler design and training methodology.

Table 1: Rank-1 accuracy of FR systems on COX Face data.

| **FR systems** | **Camera 1** | **Camera 2** | **Camera 3** |
|---|---|---|---|
| **PSCL** [10] | 36.39 ± 1.6 | 30.87 ± 1.8 | 50.96 ± 1.4 |
| **LERM** [11] | 49.07 ± 1.5 | 44.16 ± 0.9 | 63.83 ± 1.6 |
| **VGG-Face** [18] | 69.61 ± 1.5 | 68.11 ± 0.9 | 76.01 ± 0.7 |
| **TBE-CNN** [5] | 88.24 ± 0.4 | 87.86 ± 0.8 | 95.74 ± 0.7 |
| **HaarNet** [17] | 89.31 ± 0.9 | 87.90 ± 0.6 | 97.01 ± 1.7 |
| **CCM-CNN** | 88.65 ± 1.1 | 87.82 ± 0.8 | 92.13 ± 0.9 |

Results confirm that the CCM-CNN can efficiently exploit reference still images in the gallery set to enhance robustness the intra-class variations, and to increase inter-class variations between target individuals. Moreover, the proposed face augmentation is shown to be effective at reducing false negative rates by learning the appearances of the face of subjects in the gallery set. Figure 2 shows ROC curves for each camera of COX Face DB for the proposed system, along with those of PSCL [10], LERM [11] and HaarNet [17]. As shown in the figure, the area under the ROC curve (AUC) of the proposed system outperforms PSCL and LERM, while it is slightly lower than HaarNet.

For comparison on the Chokepoint data, we use the CCM-CNN fine-tuned on the COX Face data without any modifications. The network is fine-tuned using facial ROIs extracted from videos augmented with still images of the subjects of interest. Table 2 shows the AUC accuracy of the proposed network compared to EoSVM [2] and HaarNet [17]. Results indicate that the proposed CCM-CNN can provide a level of accuracy that is comparable to that of the more complexe EoSVM and HaarNet systems. It is worth noting that, EoSVM implements a fast but somewhat complex individual-specific ensemble of classifiers for each subject of interest using multiple face representations.

The final experiment is conducted using a protocol similar to the one adopted by [15], where training is performed on a separate dataset (COX Face DB) and evaluated on the videos of Chokepoint dataset. All ROIs detected in Chokepoint videos are considered as probe ROIs and all reference still ROIs are stored as gallery. Transformation matrix using a stereo matching cost is computed between the high and low resolution faces for performing robust low resolution FR. The rank-1 accuracy of the CNN for low resolution FR in [15] on Chokepoint videos is 62.7%, whereas the proposed network can achieve an accuracy of 85.9%.

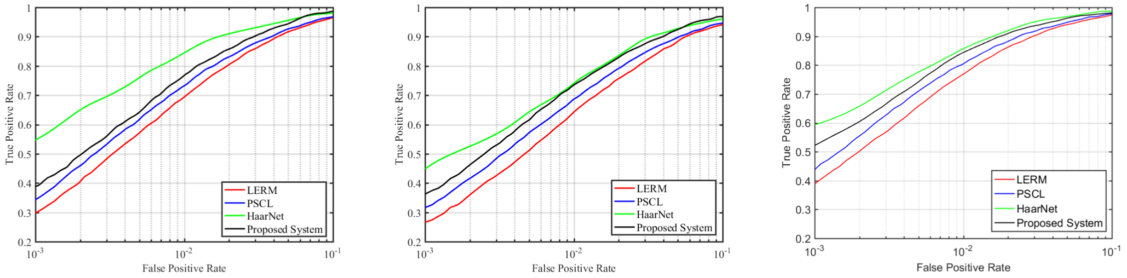
Table 2: Performance of FR systems on Chokepoint data.

| **FR systems** | **Accuracy** | **Complexity** | |
|---|---|---|---|
| | **AUPR** | **No. operations** | **No. parameters** |
| **ESRC-DA** [16] | 76.97±0.07 | 228M | 41.5M |
| **EoSVM** [2] | 99.24±0.38 | 2.3M | 230K |
| **TBE-CNN** [5] | N/A | 12.8B | 46.4M |
| **HaarNet** [17] | 99.36±0.59 | 3.5B | 13.1M |
| **CCM-CNN** | 98.87±0.63 | 33.3M | 2.4M |

Systems for FR in video surveillance are often required for real-time applications. In order to confirm the viability of CCM-CNN for such applications, Table 2 also compares the complexity of systems in terms of the number of parameters and operations (to match a video probe ROI to a reference still ROI) [3] using an Intel(R) Core(TM) i7-37700M (3.40GHz) PC with a GEFORCE GTX 1070 8GB. Since the proposed system has a significantly lower complexity, the implementations of the proposed CCM-CNN can provide fast and compact face matching solutions that nonetheless maintains a high-level of accuracy.

## 5. Conclusion

This paper presents a cost-effective CCM-CNN architecture that is specialized for still-to-video FR from a single reference still by simulating weighted CCM. In the pro-



(a) Camera1 (b) Camera2 (c) Camera3

Figure 2: ROC curves of the proposed CCM-CNN and baseline systems for videos of each camera in the COX Face dataset.

posed network, a cascade of convolutional layers effectively extracts discriminative feature maps from the still and video ROIs. A novel triplet-based loss optimization method allows to learn complex and non-linear facial representations that provide robustness across various real-world capture conditions. Finally, to overcome the limited robustness of facial models in such SSPP problems, CCM-CNN is fine-tuned using synthetically-generated faces from still ROIs of non-target individuals. Results obtained on the COX Face and Chokepoint videos indicate that the accuracy of CCM-CCN is comparable to state-of-the-art FR systems but with a significant reduction in complexity. As such, it is promising for the real-time FR in video surveillance applications, and can be extended for spatio-temporal recognition, where matching scores are combined over successive frames.

## References

[1] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau. Dynamic ensembles of exemplar-svms for still-to-video face recognition. *Pattern Recognition*, 69:61 – 81, 2017.

[2] 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*, 28(1):219–241, 2017.

[3] A. Canziani, A. Paszke, and E. Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

[4] W. Deng, J. Hu, and J. Guo. Extended src: Undersampled face recognition via intraclass variant dictionary. *IEEE Trans on PAMI*, 34(9):1864–1870, 2012.

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

[6] 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.

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

[8] Y. S. Heo, K. M. Lee, and S. U. Lee. Robust stereo matching using adaptive normalized cross-correlation. *IEEE Trans on PAMI*, 33(4):807–822, 2011.

[9] 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, October 2007.

[10] 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*, 24(12):5967–5981, 2015.

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

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

[13] W. Luo, A. G. Schwing, and R. Urtasun. Efficient deep learning for stereo matching. In *CVPR*, 2016.

[14] F. Mokhayeri, E. Granger, and G. Bilodeau. Synthetic face generation under various operational conditions in video surveillance. In *ICIP*, 2015.

[15] S. P. Mudunuri and S. Biswas. Low resolution face recognition across variations in pose and illumination. *IEEE Trans on PAMI*, 38(5):1034–1040, 2016.

[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] M. Parchami, S. Bashbaghi, and E. Granger. Video-based face recognition using ensemble of haar-like deep convolutional neural networks. In *IJCNN*, 2017.

[18] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

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

[20] 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.

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

[22] J. Zbontar and Y. LeCun. Computing the stereo matching cost with a convolutional neural network. In *CVPR*, 2015.