Using Deep Autoencoders to Learn Robust Domain-Invariant Representations for Still-to-Video Face Recognition

Mostafa Parchami[1], Saman Bashbaghi[2], Eric Granger[2] and Saif Sayed[1]
[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,
eric.granger@etsmtl.ca and saif.sayed@uta.edu

Abstract

Video-based face recognition (FR) is a challenging task in real-world applications. In still-to-video FR, probe facial regions of interest (ROIs) are typically captured with lower-quality video cameras under unconstrained conditions, where facial appearances vary according to pose, illumination, scale, expression, etc. These video ROIs are typically compared against facial models designed with high-quality reference still ROI of each target individual enrolled to the system. In this paper, an efficient Canonical Face Representation CNN (CFR-CNN) is proposed for accurate still-to-video FR from a single sample per person, where still and video ROIs are captured in different conditions. Given a facial ROI captured under unconstrained video conditions, the CRF-CNN reconstructs it as a high-quality canonical ROI for matching that corresponds to the conditons of reference still ROIs (e.g., well-illuminated, sharp, frontal views with neutral expression). A deep autoencoder network is trained using a novel weighted loss function that can robustly generate similar face embeddings for the same subjects. Then, during operations, those face embeddings belonging to pairs of still and video ROIs from a target individual are accurately matched using a fully-connected classification network. Experimental results obtained with the COX Face and Chokepoint datasets indicate that the proposed CFR-CNN can achieve convincing level of accuracy. The computational complexity (number of operations, network parameters and layers) is significantly lower than state-of-the-art CNNs for video FR, and suggests that the CFR-CNN represents a cost-effective solution for real-time applications.

1. Introduction

Systems for video-based FR in many security and surveillance applications (e.g., airport security and access control) attempt to detect the presence of target individuals appearing in the field of view of a video camera. These systems often require accurate and real-time FR over a network of video cameras in unconstrained scenes [3, 13]. In still-to-video FR (needed for, e.g., watch-list screening), facial ROIs captured from lower-resolution video cameras are compared with facial models of target individuals designed with a limited number of facial ROIs captured a priori, using a high-quality still camera under controlled conditions [1, 3, 6]. The limited number of reference still ROIs can therefore adversely affect the robustness of facial models to intra-class variations, the performance of still-to-video FR systems [10, 17]. In unconstrained video scenes, the appearance of ROIs can also change significantly due to variations in pose, illumination, expression, occlusion and blur [1].

In real-world still-to-video FR, only a single sample (i.e., still ROI) per person (SSPP) is typically available during enrollment of a target individual to design a representative facial model [3]. There are different approaches in literature to address SSPP problems, including extracting multiple face representations, face synthesizing, and using auxiliary data [3, 9, 11, 12]. These approaches mainly seek to augment the number of target samples to compensate the limited diversity of views, and to enhance the facial model's robustness to intra-class variations. Despite of their ability to cope with SSPP problems, FR systems suffer from the significant performance gap compared to the human visual system [11, 20]. An important consideration is that faces are most accuratly matched when probe and reference faces have been capture under similar conditions[1].

To improve the performance of still-to-video FR wiþa SSPP, robust convolutional feature representations have been extracted in [23] by sampling and detecting facial anchor points using a CNN integrated with a joint and col-

[1] This focus on the face matching process of still-to-video 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

laborative sparse representation-based classification (SRC). Additionally, several recent techniques have addressed FR with SSPP from the perspective of domain adaptation (DA), with the reference gallery set as the source domain that contains a single labeled still ROIs under constrained capture conditions, and the probe ROIs as the operational domain that consists of unlabeled video ROIs under unconstrained capture conditions [1, 10, 16]. For example, in [16], an extended SRC with DA (ESRC-DA) was proposed to exploit a variation dictionary learned from unlabeled video ROIs. A pool of examplar-SVMs were trained to embed multiple face representation, and dynamic classifier ensemble selection was performed using DA methods in [2]. Finally, a deep DA network with synthetic pose-free generation of faces uses a 3D face model that has been introduced in [10] to address SSPP problems.

Recently, deep networks that can learn discriminant feature representations directly from facial images has received much attention [4, 6, 18, 17, 19]. In addition, using deep networks to producing facial images with pose- and illumination-invariant features have been extensively studied [24]. For instance, the facial component-based CNN in [26] allows to transform faces with different poses and illuminations to canonical frontal view and well-illuminated faces, where pose-robust features of the last hidden layer are employed for face representations. Several deep networks have also been proposed for multi-task learning, where faces with arbitrary poses and illuminations are rotated to target-pose faces, while preserving the identity [24, 25]. Finally, a general fully convolutional architecture was employed in [8] to encode a desired attribute and to combine it with the input image. This architecture generates images similar to the probe ROIs but with some altered attributes of facial appearance. However, these aforementioned methods cannot generate robust face embeddings that are suitable for applications in still-to-video FR, because the differences between the still and video domains are not fully considered.

Autoencoders are commonly used building blocks in deep learning architectures. The encoder module maps the input data to the hidden nodes, while the decoder returns the hidden nodes to the original input data space with minimal reconstruction error, using some deterministic mapping functions [7]. Inspired by the Denoising Autoencoder [21], several autoencoder networks have been proposed to extract robust features that remove the variances in face images [7, 14, 15]. These networks consider faces with different types of variations, like illumination, pose, etc., as noisy images. For instance, stacked progressive autoencoders (SPAE) composed of multiple shallow autoencoders was proposed in [14] to learn pose-invariat features by smoothly mapping faces to nearby frontal views. Moreover, supervised autoencoder networks has been proposed to enforce that facial variations be mapped to the canonical face (i.e., a well-illuminated frontal face with neutral expression) of target individuals in the SSPP scenario [7]. In contrast to standard autoencoders, this network was designed to extract similar features for facial ROIs corresponding to the same individual. This facilitates robust FR coupling with the conventional SRC in order to predict the labels of probe ROIs.

In this paper, an efficient Canonical Face Representation CNN (CFR-CNN) is proposed for accurate still-to-video FR from a SSPP, where still and video ROIs are captured under various conditions. This CNN is based on deep supervised autoencoder that can represent the divergence between source (still ROI) and target (video ROI) domains encountered in still-to-video FR. The autoencoder network is trained using a novel weighted pixel-wise loss function that is specialized for SSPP problems, and allows to reconstruct high-quality canonical ROIs (frontal, well-illuminated, less blurred faces with neutral expression) for matching that correspond to the conditions of reference still ROIs. In addition, the intermediate layers of the autoencoder are designed to generate discriminative face embeddings that are similar for the same individuals, and robust to variations typically observed in unconstrained real-world video scenes. A fully-connected classification network is also trained to perform face matching using the face embeddings extracted from the deep autoencoder, and accurately determine whether the pairs of still and video ROIs correspond to the same individual. The proposed CFR-CNN is compared against state-of-the-art systems for still-to-video FR using challenging Cox Face [11] and Chokepoint [22] datasets.

2. A Canonical Face Representation CNN

The proposed CFR-CNN consists of two major components: the autoencoder and classification networks. The autoencoder network allows to learn discriminant face embeddings, and to reconstruct a high-quality canonical ROIs (frontal, well-illuminated, less blurred faces with neutral expression) for matching based on a probe ROI captured in videos under various conditions. The classification network matches the face embeddings for a pair of reference still and probe video ROIs.

In order to normalize variation in face capture conditions from probe video ROIs to those in still reference ROIs, the CRF-CNN relies on autoencoders. These deep supervised autoencoders generate invariant face representations for face matching. The architecture of the proposed autoencoder is shown in Figure 1, where the input image is a probe video ROI captured using a surveillance camera, while the output is a reconstructed image. This network consists of (1) three convolutional layers each followed by a max-pooling layer to extract robust convolutional maps, and then (2) a two-layer fully-connected network that generates a 256-dimensional face embedding. The decoder reverses these operations by applying a fully-connected layer to gen-

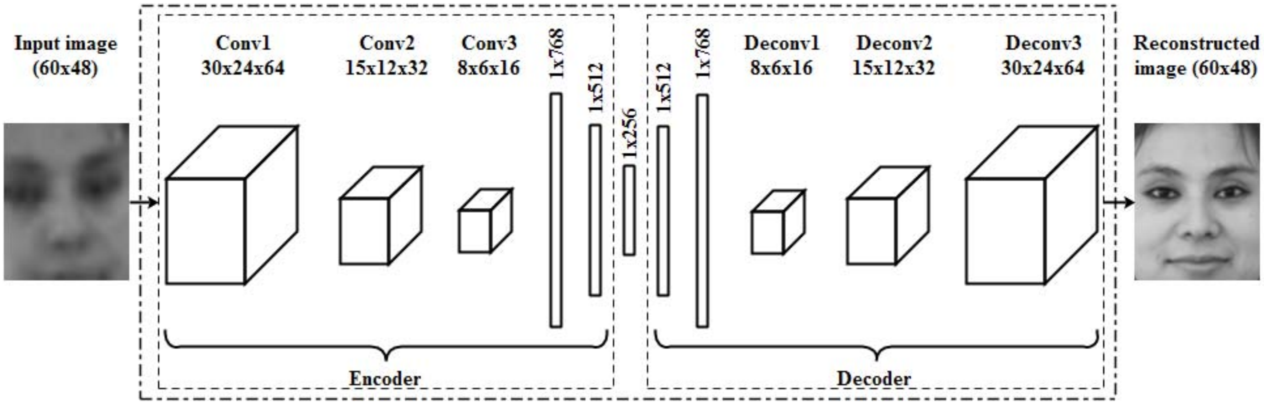


Figure 1: Block diagram of the proposed autoencoder network in the CFR-CNN.

erate the original vector and three deconvolutional layers, each one followed by un-pooling layers designed for generating the final reconstruction of the face. A fully-connected network is integrated with the deep convolutional autoencoder, and the output of the intermediate layer is then utilized as a face representation that is invariant to the different nuisance factors commonly encountered in unconstrained surveillance environments. Finally, face matching is performed using a fully-connected classification network as shown in Figure 3. This network is implemented to match the face representations of still and video ROIs.

2.1. Training the Autoencoder Network

A development set (assumed to be collected from unknown individuals captured during, e.g., camera calibration, in the operational domain) is employed for training of the deep autoencoder network. A batch of video ROIs are fed into the network where still ROIs of the corresponding persons are used for facial reconstructions. Using higher-quality still images that are captured during enrollment under controlled conditions as target faces, the autoencoder network simultaneously learns invariant face embeddings to normalize the input video ROIs. The parameters of this autoencoder network are optimized by employing a novel weighted Mean Squared Error (MSE) criterion, where a T-shaped region (illustrated in Figure 2) suggested by [1] is considered to assign a higher importance to discriminant facial components like eyes, nose and mouth. This loss function of the proposed is formulate as:

$$L_{CFR-CNN} = \sum_{i \in rows} \sum_{j \in cols} \tau_{i,j} \left\| X^2 - \hat{X}^2 \right\| \qquad (1)$$

$$\tau_{i,j} = \begin{cases} \alpha & \text{if (i,j) belongs to T} \\ \beta & \text{if (i,j) otherwise} \end{cases}$$

where $rows \times cols$ is the size of ROIs, X is the target still ROI and $\hat{X}$ is the reconstructed ROI. The weight α is considered for the T region, while the weight β is considered for pixels outside the T region.

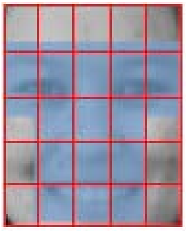


Figure 2: T-shaped weight mask used for the proposed CFR-CNN loss function.

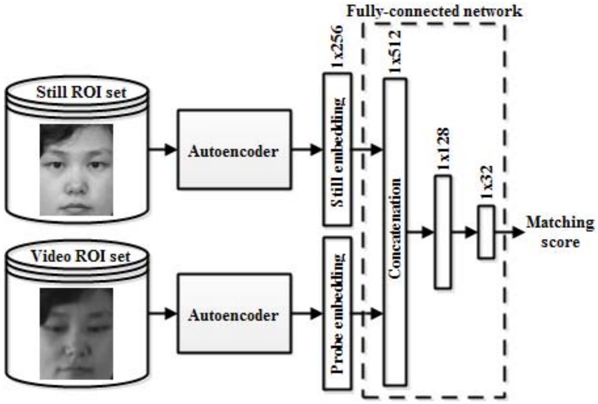


Figure 3: Block diagram of the proposed CFR-CNN.

In order to appropriately train the autoencoder network for domain adaptation, the COX Face DB was used to provide both low-quality video ROIs from the operational domain, and the corresponding high-quality still ROIs (ground truths) from enrollment for 1000 subjects. Following the training protocol suggested in [11], 200 of the subjects are randomly selected to train the autoencoder network. Their video ROIs are fed as input along with respective still ROIs (desired response) for 100 epochs, using the Adam optimization algorithm. Then, the trained network outputs a higher-quality reconstructed ROI, as well as a robust face embedding extracted from both ROIs. It thereby generates similar representations for the same identities. As shown in Figure 4, the network can successfully reconstruct the faces and subsequently, generate a neutral frontal image for each given video ROI. While these reconstructed faces may appear to be visually accurate, the face embeddings generated by the network can be utilized for robust still-to-video FR.

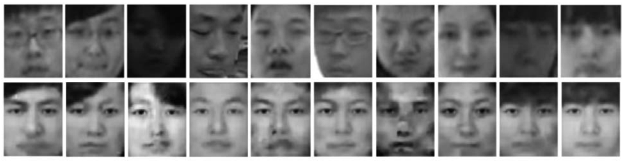


Figure 4: Examples of 10 reconstructed images from the autoencoder. The top row are the probe video ROIs, and the bottom are their corresponding canonical ROIs.

2.2. Training the Classification Network

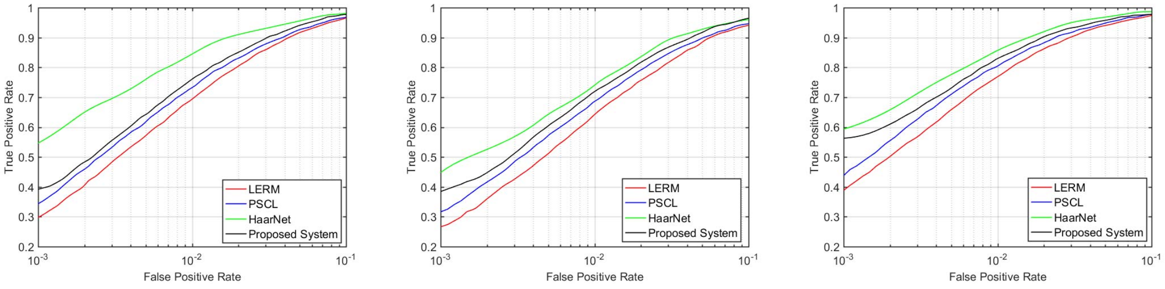
The fully-connected classification network is trained using a regular pairwise-matching scheme, where the face embeddings of the reference still and probe video ROIs are fed into the classification network. The network can thereby learn to classify each pair of still and video ROIs as either matching or non-matching. As in [11], ROIs belonging to a random selection of 100 subjects from the COX Face DB is used to train the classification network, so data from only 300 subjects are used for the overall training process. The training dataset is generated by pairing still and video ROIs and assigning matching or non-matching labels to each pair. Furthermore, the autoencoder is applied to each ROI to produce a face embedding, and each pair of still and video face embeddings are fed as input and label to the classification network. The network is trained for 20 epochs using the Adam algorithm with about 10,000 training samples by optimizing the cross-entropy criterion. This network has achieved 89.01% accuracy on validation data.

3. Experimental Methodology

The performance of the proposed CFR-CNN is evaluated and compared to the state-of-the-art systems for still-to-video FR using two publicly-available datasets for video FR – Cox Face DB [11] and ChokePoint [22]. These datasets are specifically constructed for video surveillance applications and are both composed of high-quality still faces captured with still cameras under controlled conditions and low-quality video faces captured with video cameras under uncontrolled conditions. More specifically, Cox Face DB consists of one still and three video sequences of 1000 subjects captured from different viewpoints. In contrast, the ChokePoint dataset is a benchmark for analysis of video surveillance techniques under real-world scenarios. This dataset consists of high-quality still images and videos of 25 subjects in portal one, and 29 subjects in portal 2. During four sessions, subjects walk through different portals, and videos were recorded using an array of three cameras located above the portals. In total, 64,204 facial ROIs are extracted from 48 video sequences captured while subjects enter and leave the scene from these two portals.

The proposed system is evaluated according to experimental protocols suggested in [3], [11] on different datasets. For COX Face DB, a set of 300 subjects are selected randomly for training the autoencoder and classification networks. Testing is performed using videos from the remaining 700 subjects. Average results are obtained from 10 independent replication, with random selection of training and testing subjects for each replication. Thus, high-resolution still ROIs from the 700 subjects are used as the reference gallery set, and facial ROIs of the videos from the corresponding 700 subjects are considered as the the probe set. Each probe video ROI is compared against all the still reference ROIs, and rank-1 recognition is reported as the FR accuracy. The same trained network is used without any further training for evaluation on ChokePoint videos, where 5 subjects of interest are randomly selected and their still ROIs used as the reference gallery ROI set. All video ROIs of these subjects along with 10 unknown subjects are selected at ramdom to appeared in the scene, are used as the probes. This process is replicated 5 times, each time with random selection of the subjects of interest.

All ROIs in both datasets were isolated in stills and video frames using the Viola-Jones algorithm, converted to greyscale, and then scaled to a common 60x48 pixels. Proposed CFR-CNN was implemented using Torch 7.0 deep learning framework [5]. Rank-1 recognition accuracy and ROC curves of the proposed network were compared to that of Point-to-Set Correlation Learning (PSCL) [11], Learning Euclidean to Riemannian Metric (LERM) [12], VGG-Face [18], Trunk-Branch Ensemble CNN (TBE-CNN) [6] and HaarNet [17] on the COX Face DB, and also ensemble of SVMs (EoSVMs) [3] and ESRC-DA [16] on the ChokePoint dataset. Additionally, the area under precision-recall curve (AUPR) is used to measure the global performance because of the imbalanced data of ChokePoint. The precision-Recall curve is defined by precision and true-positive-rate as recall, where precision is the ratio of true



(a) Camera1

(b) Camera2

(c) Camera3

Figure 5: ROC curves of the proposed method and baseline FR methods for videos of each camera in the Cox Face DB.

positives over the sum of true and false positives.

4. Results and Discussions

Table 1 shows the rank-1 accuracy of the proposed CFR-CNN compared against state-of-the-art FR systems using the COX Face DB. As observed in the table, PSCL, LERM and VGG-Face perform poorly, because they are not specifically designed for still-to-video FR. It is worth mentioning that, PSCL and LERM employed desciptors (hand-crafted features), while the CNNs that learn robust face representations. Amongst the CNN-based techniques that consider video-based FR, TBE-CNN and HaarNet provide the highest level of accuracy. Although the proposed light-weight CFR-CNN does not outperform these CNNs, but it can achieve its satisfactory accuracy with significantly lower complexity (see Table2).

Table 1: Average rank-1 accuracy of the proposed CFR-CNN and state-of-the-art FR systems on the COX Face DB.

FR system	**Camera 1**	**Camera 2**	**Camera 3**
PSCL [11]	36.39 ± 1.6	30.87 ± 1.8	50.96 ± 1.4
LERM [12]	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 [6]	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
CFR-CNN	85.32 ± 0.8	84.93 ± 1.2	91.52 ± 0.9

Figure 5 shows the ROC curves for PSCL, LERM, HaarNet and CFR-CNN for each video camera of COX Face DB. The CFR-CNN outperforms PSCL and LERM at transaction-level, while it can achieve comparable performance against HaarNet, specially over camera 2 and camera 3, where the facial appearence of ROIs is more distorted.

Table 2 presents the AUPR accuracy of the proposed CFR-CNN compared against the state-of-the-art FR systems on the Chokepoint dataset. Additionally, Table 2 shows the complexity of FR systems in terms of number of operations, network parameters and layers. As shown in the table, CFR-CNN outperforms ESRC-DA. Although CFR-CNN achieves slightly lower level of accuracy, it represents a significantly lighter solution than EoSVMs and HaarNet systems. Indeed, EoSVMs was designed using an individual-specific ensemble of classifiers for each subject of interest and HaarNet is a trunk-branch ensemble CNN.

Table 2: Average AUPR performance and complexity of CFR-CNN and state-of-the-art systems on Chokepoint data.

FR system	**AUPR**	**Computational Complexity**		
	Accuracy	**# operations**	**# parameters**	**# layers**
ESRC-DA [16]	76.97±0.07	228M	41.5M	N/A
E-SVMs [3]	99.24±0.38	2.3M	230K	N/A
TBE-CNN [6]	N/A	12.8B	46.4M	144
HaarNet [17]	99.36±0.59	3.5B	13.1M	56
CFR-CNN	96.47±0.86	3.75M	1.2M	7

Since still-to-video FR systems are often required to perform real-time processing in surveillance applications, the number of operations needed to process probe video ROIs is an important consideration. It can be seen in Table 2 that the proposed CFR-CNN needs significantly lower number of operations among other state-of-the-art FR systems. It confirms the viability of CFR-CNN to be for real-time applications. Moreover, the number of network parameters and layers are also a key factors in deep CNN designs that can greatly affect the training time. Considering these criteria, the proposed CFR-CNN has the lowest design complexity, and subsequently the shortest training time. Note that a complex triplet-based loss function was employed to train TBE-CNN and HaarNet to learn a face embedding, where it aims to discriminate between the positive pair of two matching ROIs and the negative non-matching ROI.

Meanwhile, training data is typically limited in many video-based FR applications, where gathering sufficient training data to train a large network is costly and time consuming. TBE-CNN and HaarNet trained their networks on 2.6M and 1.3M training data, respectively, while the CFR-CNN has been trained using only 136K training samples.

5. Conclusion

This paper presents an efficient deep learning architecture for accurate still-to-video FR from a SSPP, where the capturing conditions of still and video ROIs differ. The CFR-CNN employs a new supervised autoencoder network to generate canonical face representations from video ROIs that are robust to variations in appearance commonly found in the operational video scene, such as changes in illumination, pose, etc. In particular, prior to face matching, it allows to reconstruct faces from a low-quality video ROIs that correspond to capture conditions of reference still ROIs. During operations, these face representations are fed to a fully-connected classification network that accurately matches face embeddings belonging to pairs of still and video ROIs from a target individual. Experimental results obtained with the COX Face and Chokepoint datasets show that the proposed CFR-CNN is capable of learning discriminant face representations for matching, despite the simple configuration and small amounts of data needed for training. The proposed system can provide a high level of accuracy that is comparable to state-of-the-art CNNs for video FR, but with a significantly lower computational complexity. As such, it represents a cost-effective solution for real-time security and surveillance applications.

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. Dynamic selection of exemplar-svms for watch-list screening through domain adaptation. In *ICPRAM*, 2017.

[3] 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.

[4] 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*, abs/1601.07883, 2016.

[5] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *NIPS Workshops*, 2011.

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

[7] S. Gao, Y. Zhang, K. Jia, J. Lu, and Y. Zhang. Single sample face recognition via learning deep supervised autoencoders. *IEEE Transactions on Information Forensics and Security*, 10(10):2108–2118, 2015.

[8] A. Ghodrati, X. Jia, M. Pedersoli, and T. Tuytelaars. Towards automatic image editing: Learning to see another you. In *BMVC*. 2016.

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

[10] S. Hong, W. Im, J. Ryu, and H. S. Yang. Sspp-dan: Deep domain adaptation network for face recognition with single sample per person. *arXiv preprint arXiv:1702.04069*, 2017.

[11] 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.

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

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

[14] M. Kan, S. Shan, H. Chang, and X. Chen. Stacked progressive auto-encoders (spae) for face recognition across poses. In *CVPR*, 2014.

[15] Q. V. Le. Building high-level features using large scale unsupervised learning. In *ICASSP*, 2013.

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

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *JMLR*, 11:3371–3408, 2010.

[22] 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.

[23] M. Yang, X. Wang, G. Zeng, and L. Shen. Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person. *Pattern Recognition*, 66:117 – 128, 2017.

[24] J. Yim, H. Jung, B. Yoo, C. Choi, D. Park, and J. Kim. Rotating your face using multi-task deep neural network. In *CVPR*, 2015.

[25] Z. Zhu, P. Luo, X. Wang, and X. Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*. 2014.

[26] Z. Zhu, P. Luo, X. Wang, and X. Tang. Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543*, 2014.