SSPP-DAN: DEEP DOMAIN ADAPTATION NETWORK FOR FACE RECOGNITION WITH SINGLE SAMPLE PER PERSON

Sungeun Hong, Woobin Im, Jongbin Ryu, Hyun S. Yang

School of Computing, KAIST, Republic of Korea

ABSTRACT

Real-world face recognition using a single sample per person (SSPP) is a challenging task. The problem is exacerbated if the conditions under which the gallery image and the probe set are captured are completely different. To address these issues from the perspective of domain adaptation, we introduce an SSPP domain adaptation network (SSPP-DAN). In the proposed approach, domain adaptation, feature extraction, and classification are performed jointly using a deep architecture with domain-adversarial training. However, the SSPP characteristic of one training sample per class is insufficient to train the deep architecture. To overcome this shortage, we generate synthetic images with varying poses using a 3D face model. Experimental evaluations using a realistic SSPP dataset show that deep domain adaptation and image synthesis complement each other and dramatically improve accuracy. Experiments on a benchmark dataset using the proposed approach show state-of-the-art performance.

Index Terms— SSPP face recognition, Domain adaptation, Image synthesis, SSPP-DAN, Surveillance camera

1. INTRODUCTION

There are several examples of face recognition systems using a single sample per person (SSPP) in daily life, such as applications based on an ID card or e-passport [1]. Despite its importance in the real world, there are several unresolved issues associated with implementing systems based on SSPP. In this paper, we address two such difficulties and propose a deep domain adaptation with image synthesis to resolve these.

The first issue encountered while using SSPP is the heterogeneity of the shooting environment between the gallery and probe set [2]. In real-world scenarios, the photo used in an ID card or e-passport is captured in a very stable environment and is often used as a gallery image. On the other hand, probe images are captured in a highly unstable environment using equipment such as surveillance cameras. The resulting image includes noise, blur, arbitrary pose, and illumination, which makes recognition difficult.

(a) (b) (c)

Fig. 1: Examples of (a) a stable gallery image (source domain) (b) synthetic images generated to overcome the lack of gallery samples (source domain) (c) unstable probe images that include blur, noise, and pose variation (target domain)

To address this issue, we approach SSPP face recognition from the perspective of domain adaptation (DA). Generally, in DA, a mapping between the source domain and the target domain is constructed, such that the classifier learned for the source domain can also be applied to the target domain. Inspired by this, we assume stable shooting condition of a gallery set as the source domain and unstable shooting condition of a probe set as the target domain as shown in Fig. 1. To apply DA in the unified deep architecture, we use a deep neural network with domain-adversarial training, in a manner proposed in [3]. The benefit of this approach is that labels in the target domain are not required for training, i.e., the approach accommodates unsupervised learning.

The second challenge in using SSPP is in the shortage of training samples [4]. In general, the lack of training samples affects any learning system adversely, but it is more severe for deep learning approaches. To overcome this, we generate synthetic images with varying poses using a 3D face model [5] as shown in Fig. 1 (center). Unlike SSPP methods based on external datasets [4, 6, 7], we generate virtual samples from an SSPP gallery set. The proposed method also differs from conventional data augmentation methods that use crop, flip, and rotation [8, 9] in that it takes into account well-established techniques such as facial landmark detection and alignment that consider realistic facial geometric information. We propose a method SSPP-DAN that combines face image synthe-

This work was partly supported by Institute for Information & Communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) [R0124-16-0002, Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly] and the IT R&D program of MSIP/KEIT [10041610, The development of automatic user information(identification, behavior, location) extraction and recognition technology based on perception sensor network(PSN) under real environment for intelligent robot] The proprietary rights of EK-LFH introduced in this study belong to Electronics and Telecommunications Research Institute (ETRI).

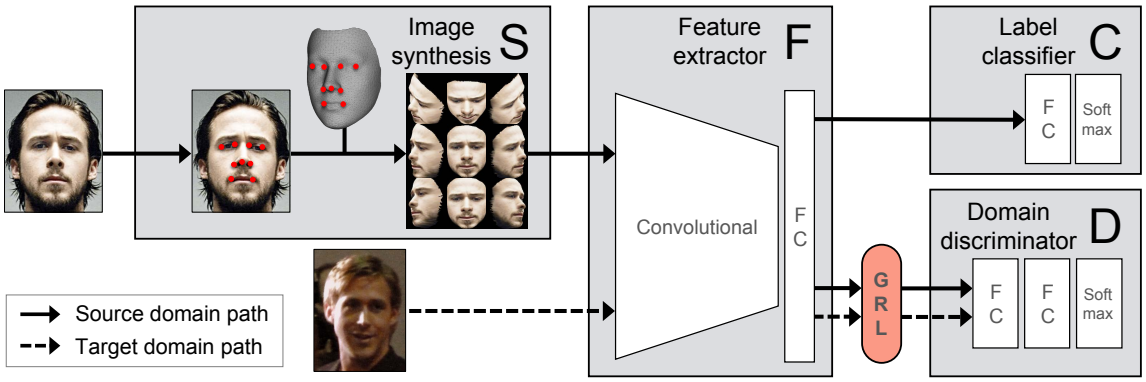


Fig. 2: Outline of the SSPP-DAN. Image synthesis is used to increase the number of samples in the source domain. The feature extractor and two classifiers are used to bridge the gap between source domain (i.e., stable images) and target domain (i.e., unstable images) by adversarial training with gradient reversal layer (GRL).

sis and DA network to enable realistic SSPP face recognition.

To validate the effectiveness of SSPP-DAN, we constructed a new SSPP dataset called ETRI-KAIST Labeled Faces in the Heterogeneous environment (EK-LFH). In this dataset, the gallery set was captured using a webcam in a stable environment, and the probe set was captured using surveillance cameras in an unconstrained environment. Using the experimental results, we validated that DA and image synthesis complement each other and eventually show a drastic 19.31 percentage points improvement over the baseline that does not use DA and image synthesis. Additionally, we performed experiments on the SSPP protocol of Labeled Faces in the Wild (LFW) benchmark [10] to demonstrate the generalization ability of the proposed approach and confirmed state-of-the-art performance.

The main contributions of this study are as follows: (i) We propose SSPP-DAN, a method that combines face synthesis and deep architecture with domain-adversarial training. (ii) To address the lack of realistic SSPP datasets, we construct a dataset whose gallery and probe sets are obtained from very different environments. (iii) We present a comparative analysis of the influence of DA with the face benchmark as well as with the EK-LFH dataset.

2. RELATED WORKS

A number of methods based on techniques such as image partitioning and generic learning have been proposed to address the shortage of training samples in SSPP face recognition. Image partitioning based methods augment samples by partitioning a face image into local patches [1, 11]. Although these techniques efficiently obtain many samples from a single subject, the geometric information of the local patch is usually ignored. There have been attempts to use external generic sets [4, 6, 7] by assuming that the generic set and the SSPP gallery set share some intra-class and inter-class information [12]. In this study, we augmented virtual samples from an SSPP gallery set instead of using an external set.

Several studies have proposed the application of DA for face recognition. Xie et al. [2] used DA and several descriptors like LBP, LPQ, and HOG to handle the scenario in which the gallery set consists of clear images and the probe set has blurred images. Banerjee et al. [13] proposed a technique for surveillance face recognition using DA and a bank of eight descriptors such as Eigenfaces, Fisherfaces, Gaborfaces, FV-SIFT, and so on. Unlike the above approaches, which apply DA after extracting the handcrafted-feature from the image, we jointly perform feature learning, DA, and classification in an integrated deep architecture. Moreover, we solve the SSPP problem and consider pose variations, unlike the abovementioned approaches that only use frontal images.

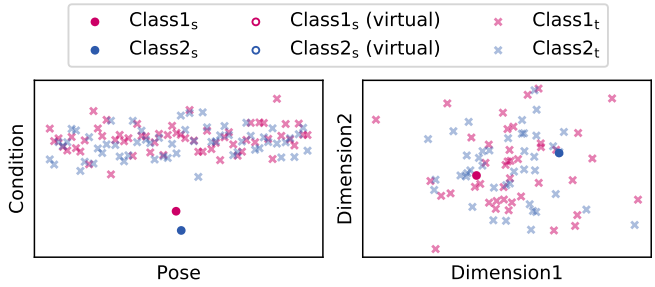
A face database using surveillance camera called SCface was proposed in [14]. In SCface, only one person appears in each image and they are photographed at a fixed location. In contrast, the images in ours were captured in an unconstrained scenario in which 30 people were walking in the room, which induced more noise, blur, and partial occlusions.

3. PROPOSED METHOD

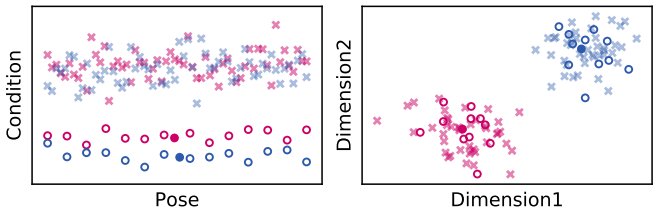
SSPP-DAN consists of two main components: virtual image synthesis and deep domain adaptation network (DAN) that consists of feature extractor and two classifiers. The overall flow of SSPP-DAN is illustrated in Fig. 2.

3.1. Virtual Image Synthesis

The basic assumption in DA is that samples are abundant in each domain and the sample distribution of each domain is similar but different (i.e., shifted from the source domain to the target domain [15]). However, in our problem under consideration, there are few samples in the source domain (i.e., SSPP). In such an extreme situation, it is difficult to apply DA directly and eventually, the mechanism will fail. To address this problem, we synthesize images with changes in pose, which improves the feature distribution obtained from the face images.



(a) DA fails to work because of the lack of samples in the source domain



(b) Virtual samples along the pose axis enable successful DA, resulting in a discriminative embedding space

Fig. 3: Facial feature space (left) and its embedding space after applying DA (right). The subscript "s" and "t" in the legend refer to the source and target domains, respectively.

For image synthesis, we first estimate nine facial landmark points from the source domain. We use the supervised descent method (SDM) [16] because it is robust to illumination changes and does not require a shape model in advance. We then estimate a transformation matrix between the detected 2D facial points and the landmark points in the 3D model [5, 17] using least-squares fit. Finally, we generate synthetic images in various poses, and these are added to the source domain as shown in Fig. 3.

3.2. Domain Adaptation Network

While the variations in pose between the distributions of the two domains can be made similar by image synthesis S, other variations such as blur, noise, partial occlusion, and facial expression remain. To resolve the remaining differences between the two domains using DA, we use a deep network that consists of feature extractor F, label classifier C, and domain discriminator D. Given an input sample, it is first mapped as a feature vector through F. There are two branches from the feature vector—the label (identity) is predicted by C and the domain (source or target) is predicted by D as shown in Fig. 2.

Our aim is to learn deep features that are discriminative on the labeled source domain during training. For this, we update the parameters of F and C, θ_F and θ_C, to minimize the label prediction loss. At the same time, we aim to transfer knowledge from the network trained on the labeled source domain to the unlabeled target domain (recall that we consider unsupervised DA). To obtain the domain-invariant features, we attempt to find a θ_F that maximizes the domain prediction loss, while simultaneously searching for parameters of D (θ_D) that minimize the domain prediction loss. Taking into consideration all these aspects, we set the loss functions as

$$\begin{aligned} L_C &= \sum_{i \in S} L_C^i && \text{when update } \theta_C \\ L_D &= \sum_{i \in S \cup T} L_D^i && \text{when update } \theta_D \\ L_F &= \sum_{i \in S} L_C^i - \lambda \sum_{i \in S \cup T} L_D^i && \text{when update } \theta_F \end{aligned} \tag{1}$$

where L_C^i and L_D^i represent the loss of label prediction and domain prediction evaluated in the i-th sample, respectively. Here, S and T denote a finite set of indices of samples corresponding to the source and target domains. The parameter λ is the most important aspect of this equation. A negative sign of λ leads to an adversarial relationship between F and D in terms of loss, and its size adjusts the trade-off between them. As a result, during minimization of the network loss L, the parameters of F converge at a compromise point that is discriminative and satisfies domain invariance.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

In all experiments, the face region was detected using the AdaBoost detector trained using Faces in the Wild [18]. For feature learning, we fine-tuned a pre-trained CNN model, VGG-Face [8], used it as the feature extractor F, and attached a shallow network as the label classifier C and domain discriminator D. The code for the SSPP-DAN and EK-LFH dataset are publicly available at our online repository (https://github.com/csehong/SSPP-DAN).

4.2. Evaluation on EK-LFH

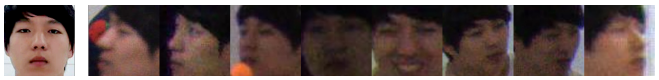
Owing to the lack of a dataset suitable for real-world SSPP, we constructed a EK-LFH dataset containing 15,930 images of 30 subjects. Table 1 shows the details of the dataset. The webcam set was used as the source domain for the training. In the surveillance set, 10,760 samples were used for training without labels in the target domain, and the rest were used for testing. Example images are shown in Fig. 4.

To demonstrate the effectiveness of the proposed method, we performed evaluations using several models as shown in Table 2 using the procedure followed in [3]. The source-only model was trained using samples in the source domain,

Table 1: Dataset specification

Domain	Source	Target
Set	webcam	surveillance
Subjects	30	30
Samples	30	15,900
Pose	frontal	various
Condition	stable	unstable (blur, noise, illumination)

(a) Shooting condition for the source (left) and target (center and right)



(b) Face regions from the source (leftmost) and target (the others)

Fig. 4: Sample images in EK-LFH

Table 2: Recognition rates (%) for different models and different training sets of the EK-LFH

Model	Training set	Accuracy
Source only	S	39.22
	$S + S_v$	37.15
DAN	$S + T$	31.11
SSPP-DAN	$S + S_v + T$	**58.53**
Semi DAN	$S + T + T_l$	67.28
Semi SSPP-DAN	$S + S_v + T + T_l$	**72.08**
Train on target	T_l	88.31

S: Labeled webcam T: Unlabeled surveillance
S_v: Virtual set from S T_l: Labeled surveillance

which revealed the theoretical lower bound on performance as 39.22%. The train-on-target model was trained on the target domain with known class labels. This revealed the upper performance bound as 88.31%. The unlabeled target domain as well as the labeled source domain were used in DAN and SSPP-DAN for unsupervised DA. Additionally, we evaluated the semi-supervised models using the same setting as DAN and SSPP-DAN, but by revealing only three labels per person in the target domain.

From Table 2, we clearly observe that SSPP-DAN with unsupervised as well as semi-supervised learning significantly improves accuracy. In particular, even when the labels of the target domain are not given, the accuracy of the proposed SSPP-DAN was 19.31 percentage points higher than that for source-only. The fourth and fifth rows validate the importance of image synthesis when applying unsupervised DA. Adding synthesized virtual images to the training set increased the performance by 27.42 percentage points. Interestingly, as shown in the third row, adding synthetic images to source-only degrades performance. This result indicates that image synthesis alone cannot solve the SSPP problem efficiently, instead DA and image synthesis operate complementarily in addressing the SSPP problem.

4.3. Evaluation on LFW for SSPP

In order to demonstrate the generalization ability of SSPP-DAN, we performed an additional experiment on the LFW

Table 3: Recognition rates (%) on LFW dataset for SSPP

Method	Accuracy	Method	Accuracy
DMMA [1]	17.8	RPR [20]	33.1
AGL [6]	19.2	DeepID [21]	70.7
SRC [4]	20.4	JCR-ACF [19]	86.0
ESRC [7]	27.3	VGG-Face [8]	96.43
LGR [22]	30.4	**Ours**	**97.91**

using the proposed SSPP method. For fair comparison with previous SSPP methods, we used LFW-a [10], and followed the experimental setup described in [19]. The LFW for SSPP included images from 158 subjects, each of which contained more than 10 samples, as well as the labels of all subjects. The first 50 subjects were used as probe and gallery, and the images of the remaining 108 subjects were used as a generic set. For the 50 subjects, the first image was used as the gallery set and the remaining images were used as the probe set.

Since the LFW did not consider DA originally, it made no distinction between source and target domain. Hence, we used the original generic set as the source domain and the synthetic images from the generic set as the target domain. We applied DA in a supervised manner to generate a discriminative embedding space. After training, we used the output of the last FC layer as the feature, and implemented prediction using the linear SVM. We also evaluated fine-tuned VGG-Face without image synthesis and DA. Experiments using the benchmark confirmed that VGG-face based methods including ours have superior discriminative power over other approaches as shown in Table 3. This indicates the generality of deep features from the VGG-Face trained on a large scale dataset. It is apparent from this table that, by comparing VGG-Face with the proposed method, the combination of image synthesis and DA shows promising results in the 'wild' dataset.

5. CONCLUSION

This paper proposed a method based on integrated domain adaptation and image synthesis for SSPP face recognition, especially for cases in which the shooting conditions for the gallery image and the probe set are completely different. Synthetic images generated in various poses were used to deal with the lack of samples in the SSPP. In addition, a deep architecture with domain-adversarial training was used to perform domain adaptation, feature extraction, and classification jointly. Experimental evaluations showed that the proposed SSPP-DAN had an accuracy 19.31 percentage points higher than that of the source-only baseline even when the labels of the target domain were not given. Our method also achieved state-of-the-art results on the challenging LFW for SSPP. In future work, we plan to expand our approach to a fully trainable architecture including image synthesis as well as domain adaptation using standard back-propagation.

6. REFERENCES

[1] Jiwen Lu, Yap-Peng Tan, and Gang Wang, "Discriminative multimanifold analysis for face recognition from a single training sample per person," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 39–51, 2013.

[2] Xiaokang Xie, Zhiguo Cao, Yang Xiao, Mengyu Zhu, and Hao Lu, "Blurred image recognition using domain adaptation," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 532–536.

[3] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, 2015, pp. 1180–1189.

[4] John Wright, Allen Y Yang, Arvind Ganesh, S Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[5] Li Zhang, Noah Snavely, Brian Curless, and Steven M Seitz, "Spacetime faces: High-resolution capture for modeling and animation," in *Data-Driven 3D Facial Animation*, pp. 248–276. Springer, 2008.

[6] Yu Su, Shiguang Shan, Xilin Chen, and Wen Gao, "Adaptive generic learning for face recognition from a single sample per person.," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 2699–2706.

[7] Weihong Deng, Jiani Hu, and Jun Guo, "Extended src: Undersampled face recognition via intraclass variant dictionary," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1864–1870, 2012.

[8] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman, "Deep face recognition.," in *BMVC*, 2015, p. 6.

[9] Kaihao Zhang, Yongzhen Huang, Ran He, Hong Wu, and Liang Wang, "Localize heavily occluded human faces via deep segmentation," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2311–2315.

[10] Lior Wolf, Tal Hassner, and Yaniv Taigman, "Effective unconstrained face recognition by combining multiple descriptors and learned background statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 10, pp. 1978–1990, 2011.

[11] Haibin Yan, Jiwen Lu, Xiuzhuang Zhou, and Yuanyuan Shang, "Multi-feature multi-manifold learning for single-sample face recognition," *Neurocomputing*, vol. 143, pp. 134–143, 2014.

[12] Tingwei Pei, Li Zhang, Bangjun Wang, Fanzhang Li, and Zhao Zhang, "Decision pyramid classifier for face recognition under complex variations using single sample per person," *Pattern Recognition*, vol. 64, pp. 305–313, 2017.

[13] Samik Banerjee and Sukhendu Das, "Domain adaptation with soft-margin multiple feature-kernel learning beats deep learning for surveillance face recognition," *arXiv preprint arXiv:1610.01374*, 2016.

[14] Mislav Grgic, Kresimir Delac, and Sonja Grgic, "Scface–surveillance cameras face database," *Multimedia tools and applications*, vol. 51, no. 3, pp. 863–879, 2011.

[15] Hidetoshi Shimodaira, "Improving predictive inference under covariate shift by weighting the log-likelihood function," *Journal of statistical planning and inference*, vol. 90, no. 2, pp. 227–244, 2000.

[16] Xuehan Xiong and Fernando De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[17] Jun-Yan Zhu, Aseem Agarwala, Alexei A Efros, Eli Shechtman, and Jue Wang, "Mirror mirror: Crowdsourcing better portraits," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 6, pp. 234, 2014.

[18] Tamara L Berg, Alexander C Berg, Jaety Edwards, and David A Forsyth, "Whos in the picture," *Advances in neural information processing systems*, vol. 17, pp. 137–144, 2004.

[19] Meng Yang, Xing Wang, Guohang Zeng, and Linlin Shen, "Joint and collaborative representation with local adaptive convolution feature for face recognition with single sample per person," *Pattern Recognition*, 2016.

[20] Shenghua Gao, Kui Jia, Liansheng Zhuang, and Yi Ma, "Neither global nor local: Regularized patch-based representation for single sample per person face recognition," *International Journal of Computer Vision*, vol. 111, no. 3, pp. 365–383, 2015.

[21] Yi Sun, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1891–1898.

[22] Pengfei Zhu, Meng Yang, Lei Zhang, and Il-Yong Lee, "Local generic representation for face recognition with single sample per person," in *Asian Conference on Computer Vision*. Springer, 2014, pp. 34–50.