LOW-RESOLUTION FACE RECOGNITION BASED ON IDENTITY-PRESERVED FACE HALLUCINATION

Shun-Cheung Lai, Chen-Hang He, and Kin-Man Lam

Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Hong Kong

ABSTRACT

The state-of-the-art Convolutional Neural Network (CNN)-based methods have achieved promising recognition performance on human face images. However, the accuracy cannot be retained when face images are at very low resolution (LR). In this paper, we propose a novel loss function, called identity-preserved loss, which combines with the image-content loss to jointly supervise CNNs, for performing face hallucination and recognition simultaneously. Therefore, the trained network is able to perform face hallucination and identity preservation, even if the query face is of very low resolution. More importantly, experimental results show that our proposed method can preserve the identities for the LR images from unknown subjects, who are not included in the training set. The source code of our proposed method is available at: `https://github.com/johnnysclai/SR_LRFR`.

***Index Terms*—** Face hallucination, low-resolution face recognition, identity-preserved loss, deep learning.

1. INTRODUCTION

Because of the use of deep neural networks, performances of face-recognition (FR) algorithms have been significantly improved over last few years, and have surpassed that of humans [1, 2]. The state-of-the-art Convolutional Neural Network (CNN)-based FR methods [3, 4, 5] have achieved recognition rates of over 99% on the widely used Labeled Faces in the Wild (LFW) benchmark [6]. However, the accuracy cannot be retained in some real-world applications, such as video surveillance, where face images are usually of low resolution and poor quality. Matching the high-resolution (HR) gallery faces with low-resolution (LR) query faces is called low-resolution face recognition (LRFR).

To show the performance degradation of the existing CNN-based methods for LRFR, we used the author-released 20-layer CNN of [4] (SFace, for short), which was trained on CASIA-Webface [7], to conduct an experiment on the LFW face verification protocol (details in Section 4.3). The mean accuracy based on 10-fold cross-validation on 6,000 face pairs is reported in Fig. 1. In this experiment, the second face image of each subject was downsampled by a factor of N to form the LR query set, while the first images form the HR gallery set. The resolution of the HR face images is 112×96 pixels. Therefore, if the downsampling factor is 16, the resolution of the query images is 7×6 pixels. In order to use SFace for face recognition, all the LR faces were upsampled to 112×96 pixels by using bicubic interpolation, which is the required input size of SFace. As shown in Fig. 1, we can observe that the verification rate degrades gradually when the query faces are downsampled.

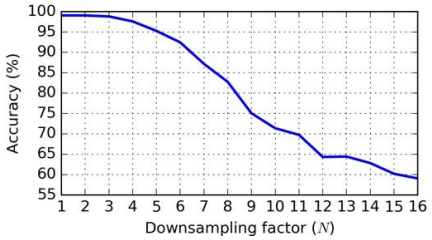

Fig. 1: Face verification accuracy (%) of SFace on LFW, with 6,000 pairs of face images at different query-face resolutions.

In this paper, we propose a novel loss function, called identity-preserved loss, which is combined with the image-content loss, for training a super-resolution network (SRNet). Therefore, SRNet, trained with the identity-preserved loss function, can perform face hallucination with identity preserved even if the face is of very low resolution.

2. RELATED WORKS

Face hallucination, also known as face super-resolution, was first proposed in [8]. It employed Bayesian formulation to estimate the gradient prior from the Gaussian and Laplacian pyramids of HR training images to reconstruct faces. Another early method, proposed in [9], used Principal Component Analysis (PCA) to reconstruct an input LR face by the weighted sum of the training face images. Tappen and Liu [10] used SIFT flow [11] to warp the training HR face candidates, followed by a Bayesian framework, to reconstruct the HR face images. These methods employ the global structural similarity of human faces. However, they fail to reconstruct an input image with pose variations. More importantly, they are in-

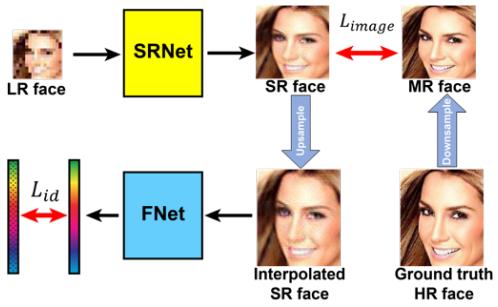

Fig. 2: An overview of the proposed method, where SRNet is jointly trained by using L_{image} and L_{id}. The resolution of a MR face is that of the output images from SRNet, while the resolution of an interpolated SR face is the required input size of FNet.

efficient for large-scale training sets, because these methods require iterations during testing.

For low-resolution face recognition, various approaches have been proposed, including multidimensional scaling [12, 13], facial feature super-resolution [14], simultaneous face hallucination and recognition [15, 16], etc. However, the features and the classifiers used are learned separately in these methods, so their performance is limited. Recently, Zhu *et al.* [17] attempted to train cascaded gated bi-networks to perform face SR and dense face corresponding field estimation simultaneously. Yu and Porikli [18, 19] employed a Generative Adversarial Network (GAN) [20] to perform face hallucination. However, these recent CNN-based methods are vision-oriented, and are not proposed for face recognition. Furthermore, if the upsampling factor is high, the identity of the super-resolved face images cannot be retained, *i.e.* a super-resolved image looks like another person.

3. PROPOSED METHOD

As shown in Fig. 2, our proposed method consists of two networks: a super-resolution network (SRNet) and a face recognition network (FNet), denoted as G and F, respectively. SRNet reconstructs an input LR face, I^{LR}, to a super-resolved face, I^{SR}, *i.e.* $I^{SR} = G(I^{LR})$. Then, I^{SR} is upsampled to the required input size of FNet by interpolation, denoted as I^{ISR}. Bilinear interpolation is used during training, so as to reduce the computational complexity, and the training will become more efficient. However, bicubic interpolation is used at testing, which can result in a slightly better performance. The deep feature of I^{ISR} is extracted by FNet, *i.e.* $\boldsymbol{y}^{ISR} = F(I^{ISR})$. The distance between $\boldsymbol{y}^{ISR}$ and the deep feature extracted from its corresponding HR face image, $\boldsymbol{y}^{HR}$, forms the identity-preserved loss.

In this paper, we use off-the-shelf SRNet and FNet. The recently proposed EDSR×4 model [21] and SFace [4] are adopted as our SRNet and FNet, respectively. The network architectures and the implementation details of EDSR and SFace can be found in their original papers. In our setting, SRNet is trained from scratch, while FNet is pre-trained on CASIA-Webface [7] with the Angular softmax (A-Softmax) loss [4]. The trainable parameters of FNet are not updated during training, *i.e.* they are frozen.

3.1. Loss function

To achieve face hallucination and identity preservation simultaneously, we train SRNet with both the image-content loss, L_{image}, and our proposed identity-preserved loss, L_{id}. Therefore, SRNet is jointly supervised by two types of signals, and an optimal set of parameters should satisfy both of the objectives.

Empirically, we found that the image-content loss is necessary to guide SRNet to generate visually convincing face images as we failed to train SRNet by using the identity-preserved loss only. Thus, we follow [21] to use the $L1$ loss as the image-content loss, which is defined as follows:

$$\begin{aligned} L_{image} &= \frac{1}{n}\sum_{i=1}^{n} \left\| I_i^{SR} - \downarrow I_i^{HR} \right\|_1 \\ &= \frac{1}{n}\sum_{i=1}^{n} \left\| I_i^{SR} - I_i^{MR} \right\|_1, \end{aligned} \tag{1}$$

where n is the number of training samples, $\downarrow$ is the bicubic downsampling operator, and I^{HR} is the ground-truth HR face image. The medium-resolution face $I^{MR} = \downarrow I^{HR}$, such that I^{MR} and I^{SR} are at the same resolution.

For face recognition, I^{SR} is interpolated to the required input size of FNet. Our proposed identity-preserved loss, L_{id}, encourages the reconstructed face image to have its feature similar to the deep feature of its HR counterpart. Therefore, this loss is defined as follows:

$$L_{id} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{y}_i^{ISR}, \boldsymbol{y}_i^{HR}), \tag{2}$$

where $\boldsymbol{y}^{ISR}$ and $\boldsymbol{y}^{HR}$ are the deep features extracted from I^{ISR} and I^{HR}, respectively. The cosine distance is used in $f(\cdot)$, because this distance is equivalent to the normalized Euclidean distance. More importantly, it has been shown that the identity information is only related to the angles of the deep features [4, 5, 22]. Therefore, $f(\cdot)$ is expressed as follows:

$$f(\boldsymbol{y}^{ISR}, \boldsymbol{y}^{HR}) = 1 - \frac{\boldsymbol{y}^{ISR} \cdot \boldsymbol{y}^{HR}}{\|\boldsymbol{y}^{ISR}\|_2 \, \|\boldsymbol{y}^{HR}\|_2}. \tag{3}$$

In this paper, the output of the FC1 layer of SFace [4] is considered as the deep feature of a face image. With the combined loss functions, SRNet is be able to super-resolve the LR face images and retain their identities simultaneously. The overall loss function, L, can be written as follows:

$$L = L_{image} + \lambda L_{id}, \tag{4}$$

where λ is a weighting factor for balancing the two losses.

4. EXPERIMENTS

4.1. Implementation

Data and preprocessing. The CelebA [23] dataset was used to train SRNet. It contains 202,599 images from 10,177 subjects. Although the dataset description states that the included identities and the identities in the LFW [6] database are mutually exclusive, as shown in Fig. 3, we have found that some of the face images are exactly the same, and some of the face identities are the same or very similar. Furthermore, some of the faces are incorrectly labelled, or are highly occluded, as shown in Fig. 4. After cleaning these images from the dataset, 200,315 images from 10,112 subjects were used to train our SRNet. We follow [4] to align all the faces based on the 5 given facial landmarks and resize them to 112×96 pixels to form HR face images.

Training details. During training, a HR face is randomly downsampled by a factor of 4, 8 or 16 (corresponding resolution 28×24, 14×12 or 7×6 pixels, respectively) to form an input LR face. To make use of parallelization of a GPU, the resolution of the input LR faces is the same in a mini-batch. The output SR faces are then upsampled to 112×96 pixels by using bilinear interpolation, followed by pixel normalization used in SFace. Training images are augmented by flipping them horizontally, with a 50% probability. The mini-batch size is set at 64, and Adam optimizer [24] with default parameters setting is used. The learning rate is initialized at 10^{-4}, and it is halved after 25K iterations. Training is finished after 50K iterations. We train the model with the PyTorch library [25], using two GTX 1080TI GPUs.

4.2. Evaluation

To measure the performance for super-resolution, the standard metrics, *i.e.* Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM), are not used, because they cannot quantitative justify if the super-resolved face images are beneficial for LRFR. To address this issue, we consider the face verification rate and the face identification rate as our evaluation metrics. We conducted experiments on the LFW [6] database. The face images were captured in uncontrolled environments with variations, including pose, expression, occlusion and lighting, so that they can demonstrate the effectiveness and generalization power of our proposed identity-preserved loss. MTCNN [26] is used to detect facial landmarks, and align the face images to form the ground-truth HR faces.

SRNet. We use the EDSR×4 model with different parameter settings to verify the effectiveness of the proposed identity-preserved loss. We denote $\text{EDSR}_{n,k,\lambda}$ as the EDSR×4 model, with n residual blocks and k filters. We trained SRNet with four parameter settings, with (*i.e.*, $\lambda = 0.5$) and without (*i.e.*, $\lambda = 0$) our proposed identity-preserved loss: $\text{EDSR}_{16,64,0/0.5}$ and $\text{EDSR}_{32,256,0/0.5}$. All the models are trained from scratch.

Fig. 3: Examples of overlapped face images in CelebA (left) and LFW (right), which are the same or have very similar identities.

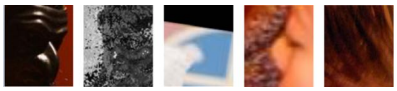

Fig. 4: Examples of noisy images in CelebA.

FNet. Apart from SFace, another pre-trained face recognition network, VGG-Face [27], was also used as FNet during testing. Thus, the SR face images are upsampled to the required input size of VGG-Face (*i.e.*, 224×224 pixels), followed by normalization used in VGG-Face. The output of the FC7 layer (before ReLU) is taken as the deep feature of a face image. It is worth noting that VGG-Face is not involved in the training pipeline. Thus, the use of VGG-Face can demonstrate the generalization capability of our proposed methods.

Following [4], in all the experiments, the final deep face feature vector is obtained by concatenating the features extracted from a face image and its horizontally flipped image. The cosine distance is used to compute the similarity of two feature vectors.

4.3. Face verification on LFW

We follow the LFW [6] unrestricted protocol to report the mean accuracy of 10-fold cross-validation on 6,000 face pairs. The second faces are downsampled to form LR query images, while the first faces are taken as the HR gallery images. The face recognition accuracy based on the HR query face images is also included for reference. We compare our proposed method to bicubic interpolation and a face hallucination method, TDAE [19]. We used the author-released model of TDAE[1] in our experiments, which was also trained on the CelebA dataset.

We measure the face-verification rate when the query images are downsampled by a factor of 4, 8 and 16 (*i.e.*, the corresponding image resolutions are 28×24, 14×12 and 7×6 pixels, respectively), which are the same resolutions used for training. Note that TDAE requires the input image size to be 16×16 pixels. Therefore, we also report the verification accuracy when the resolution is 16×16 pixels, for a fair comparison. All the verification rates are tabulated in Table 1, and some of the reconstructed faces, based on the different methods, are shown in Fig. 5.

From the face verification results, we have shown the ef-

[1] https://github.com/XinYuANU/TDAE (commit: 4676dc3)

FNet	Method/SRNet	7×6	14×12 (16×16)	28×24	112×96
VGG-Face	Bicubic	54.60	77.43 (82.53)	93.02	96.33
	TDAE [19]	-	- (73.67)	-	
	$EDSR_{16,64,0}$	71.42	83.62 (86.87)	93.88	
	$EDSR_{16,64,0.5}$	74.85	85.42 (86.88)	93.95	
	$EDSR_{32,256,0}$	77.47	85.87 (87.72)	93.95	
	$EDSR_{32,256,0.5}$	78.73	88.67 (88.90)	94.58	
SFace	Bicubic	59.03	82.75 (89.10)	97.60	99.07
	TDAE [19]	-	- (71.38)	-	
	$EDSR_{16,64,0}$	68.45	88.73 (92.18)	98.03	
	$EDSR_{16,64,0.5}$	79.67	92.07 (93.67)	98.22	
	$EDSR_{32,256,0}$	73.42	92.25 (94.32)	98.48	
	$EDSR_{32,256,0.5}$	**84.03**	**94.73 (95.10)**	**98.85**	

Table 1: Verification rates (%) based on LFW 6,000 pairs, following the LFW unrestricted setting.

FNet	Method/SRNet	7×6	14×12	16×14	18×16	112×96
-	SHI [16]	-	**66.19**	**68.05**	69.20	-
VGG-Face	Bicubic	0.23	8.04	13.78	23.01	83.20
	$EDSR_{32,256,0}$	5.13	28.68	32.59	36.84	
	$EDSR_{32,256,0.5}$	9.31	39.32	38.93	43.26	
SFace	Bicubic	0.55	11.52	22.99	40.95	97.73
	$EDSR_{32,256,0}$	5.91	50.78	56.19	65.89	
	$EDSR_{32,256,0.5}$	**14.91**	63.34	63.78	**71.96**	

Table 2: Rank-1 LFW identification rates (%), following the protocol used in [16].

Fig. 5: Examples of the reconstructed face images based on different methods. LFW face images are downsampled to 14×12 pixels to form the LR faces. The input faces for TDAE are downsampled to 16×16 pixels, which is the required input size. For the visualization purpose, all the reconstructed face images are resized to 112×96 pixels

fectiveness and generalization power of the identity-preserved loss. For both $EDSR_{16,64}$ and $EDSR_{32,256}$ models, when the model is trained with the identity-preserved loss, the LR face verification rates are improved significantly. From Fig. 5, we can obverse that even though TDAE is able to generate face images with fine details, the reconstructed face images seem to be of other identities. This is the reason why the face-verification results of TDAE are worse than those achieved by bicubic interpolation. However, we notice that there are some artifacts introduced in the reconstructed images when the identity-preserved loss is used. We have no conclusion to this observation, but we believe that if these artifacts are eliminated, the recognition rate will further be improved.

4.4. Face identification on LFW

To further demonstrate the ability of identity preservation by introducing the proposed identity-preserved loss for super-resolution, we conducted face identification experiments on the LFW database. We follow the protocol used in [16]. We select all the subjects, who consist of at least four face images per subject. One image is randomly selected as the gallery image, and the remaining images form the query images. The query images are downsampled to 18×16, 16×14, and 7×6 pixels. The accuracy based on the HR query face images is also measured for reference. All the reported results are the average rank-1 identification rates over 10 runs. The results of [16] are extracted from the original paper.

As shown in Table 2, we can see that the recognition rates of the two deep CNN-based methods, VGG-Face [27] and SFace [4], are promising, when the query face images are of high resolution. However, the performance of both methods drops dramatically, when LR query face images are used. The results reveal that the effectiveness of the proposed identity-preserved loss can be generalized to low-resolution face identification. Although the best results are just comparable with those reported in [16], we have not fine-tuned SRNet and FNet, or searched for an optimal set of hyper-parameters. In other words, our experiments completely follow the open-set face recognition setting, and our method is more straightforward and flexible. We would expect that the recognition rate for both LR face verification and LR face identification could be further improved, if a better or well-trained FNet was used.

5. CONCLUSIONS

In this paper, we have proposed a novel loss function, namely identity-preserved loss. Combining it with the image-content loss, the trained network is able to perform face hallucination with identity preservation even if the face images are of very low resolution. Experiments have shown that our proposed method can achieve superior face recognition performance, in terms of face verification and face identification. For our future work, we will consider real-world low-resolution face recognition, *i.e.*, surveillance images are used for recognition.

Acknowledgement: The work described in this paper was supported by the GRF Grant PolyU 152765/16E (project code: B-Q55J) of the Hong Kong SAR Government.

6. REFERENCES

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

[2] Y. Sun, X. Wang, and X. Tang, "Deeply learned face representations are sparse, selective, and robust," in *CVPR*, 2015.

[3] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016.

[4] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017.

[5] F. Wang, X. Xiang, J. Cheng, and A. L. Yuille, "Normface: L2 hypersphere embedding for face verification," in *ACM MM*, 2017.

[6] 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," Tech. Rep. 07-49, University of Massachusetts, Amherst, Oct 2007.

[7] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv: 1411.7923*, 2014.

[8] S. Baker and T. Kanade, "Hallucinating faces," in *FG*, 2000.

[9] X. Wang and X. Tang, "Hallucinating face by eigentransformation," *IEEE Trans. Syst. Man and Cybern. Part C*, vol. 35, no. 3, pp. 425–434, Aug 2005.

[10] M. F. Tappen and C. Liu, "A bayesian approach to alignment-based image hallucination," in *ECCV*, 2012.

[11] C. Liu, J. Yuen, and A. Torralba, "Sift flow: Dense correspondence across scenes and its applications," *IEEE Transactions on Pattern Analysis and MachineIntelligence*, vol. 33, no. 5, pp. 978–994, May 2011.

[12] S. Biswas, K. W. Bowyer, and P. J. Flynn, "Multidimensional scaling for matching low-resolution face images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 2019–2030, Oct 2012.

[13] M. Saad Shakeel and Kin-Man-Lam, "Recognition of low-resolution face images using sparse coding of local features," in *APSIPA*, 2016.

[14] K. H. Pong and K. M. Lam, "Multi-resolution feature fusion for face recognition," *Pattern Recognition*, vol. 47, no. 2, pp. 556–567, Feb 2014.

[15] P. H. Hennings-Yeomans, S. Baker, and B. V. K. V. Kumar, "Simultaneous super-resolution and feature extraction for recognition of low-resolution faces," in *CVPR*, 2008.

[16] M. Jian and K. M. Lam, "Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 11, pp. 1761–1772, Nov 2015.

[17] S. Zhu, S. Liu, C. C. Loy, and X. Tang, "Deep cascaded bi-network for face hallucination," in *ECCV*, 2016.

[18] X. Yu and F. Porikli, "Ultra-resolving face images by discriminative generative networks," in *ECCV*, 2016.

[19] X. Yu and F. Porikli, "Hallucinating very low-resolution unaligned and noisy face images by transformative discriminative autoencoders," in *CVPR*, 2017.

[20] I. Goodfellow et al., "Generative adversarial nets," in *NIPS*, 2014.

[21] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, "Enhanced deep residual networks for single image super-resolution," in *CVPRW*, 2017.

[22] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, July 2018.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *ICCV*, 2015.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2014.

[25] A. Paszke, , et al., "Automatic differentiation in pytorch," in *NIPS-W*, 2017.

[26] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[27] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *BMVC*, 2015.