2017 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, SEPT. 25–28, 2017, TOKYO, JAPAN

LEARNING GUIDED CONVOLUTIONAL NEURAL NETWORKS FOR CROSS-RESOLUTION FACE RECOGNITION

Tzu-Chien Fu[1]*, Wei-Chen Chiu*[2]*, and Yu-Chiang Frank Wang*[1]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2]National Chiao Tung University, Hsinchu, Taiwan

ABSTRACT

Cross-resolution face recognition tackles the problem of matching face images with different resolutions. Although state-of-the-art convolutional neural network (CNN) based methods have reported promising performances on standard face recognition problems, such models cannot sufficiently describe images with resolution different from those seen during training, and thus cannot solve the above task accordingly. In this paper, we propose Guided Convolutional Neural Network (Guided-CNN), which is a novel CNN architecture with parallel sub-CNN models as guide and learners. Unique loss functions are introduced, which would serve as joint supervision for images within and across resolutions. Our experiments not only verify the use of our model for cross-resolution recognition, but also its applicability of recognizing face images with different degrees of occlusion.

***Index Terms*—** Face recognition, deep learning, convolutional neural networks

1. INTRODUCTION

With the increasing use of video surveillance systems for applications in security and forensics, the demand for face recognition has been growing. However, recognizing faces using such systems in real-world scenarios not only requires one to deal with the facial image variations of pose, illumination and expression, but also those with insufficient resolution due to long distances between the subjects of interest and the camera sensors. For example, query images with low resolution (LR) need to be verified using gallery ones with high resolution (HR). As a result, how to matching images across different resolution would be a practical yet challenging task.

Face recognition has been benefited by the recent advances in deep learning, or particularly the evolvement of convolutional neural networks (CNN). Two representative CNN-based architectures for large-scale face recognition are DeepFace [1] and DeepID [2]. Viewing the last hidden layer as the extracted deep visual features, these models applied a final fully-connected softmax layer as classifiers. As an extension of DeepID, DeepID2 [3] took joint identification-verification information as supervision, which further improves the discriminating ability of the resulting deep features. Recently, FaceNet [4] introduced a triplet loss to minimize the difference between an anchor image and a positive one (i.e., with the same identity), while the distance between it and its negative one is to be maximized. In addition, Wen et al. [5] introduced a center loss function into existing CNN models, which also resulted in better recognition performance.

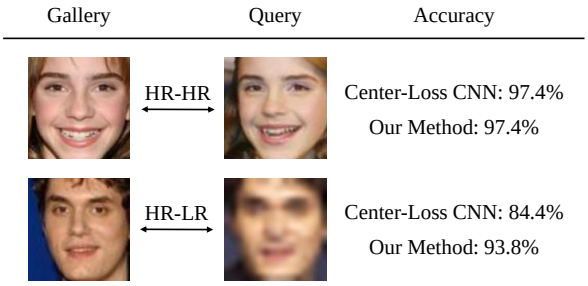

Fig. 1. Challenge of cross-resolution face recognition. While promising performance is reported for recent CNNs on the LFW dataset (e.g., Center-Loss CNN [5]), it does not generalize well if the query image is with insufficient resolution. Note that HR and LR denote high and low resolutions, respectively.

Although the above methods have reported promising results on challenging and large-scale benchmark datasets, these approaches typically assume that both the query and gallery images are with the same or similar resolution. In other words, as we verify later in the experiments, these CNN frameworks cannot be easily extended to cross-resolution recognition. Figure 1 shows examples of degraded performance of the CNN model of [5], while the resolution mismatch between training and test facial images is occurred.

In this paper, we propose a novel deep-learning based architecture of *Guided-CNN*, which can be applied for cross-domain face recognition and beyond. By utilizing an existing CNN-based face recognition model as a guide (e.g., [4] or [5]), we adapt and learn a parallel CNN model for dealing with face images with insufficient resolution. As a result, the proposed Guided-CNN can be viewed as a deep domain

978-1-5090-6341-3/17/$31.00

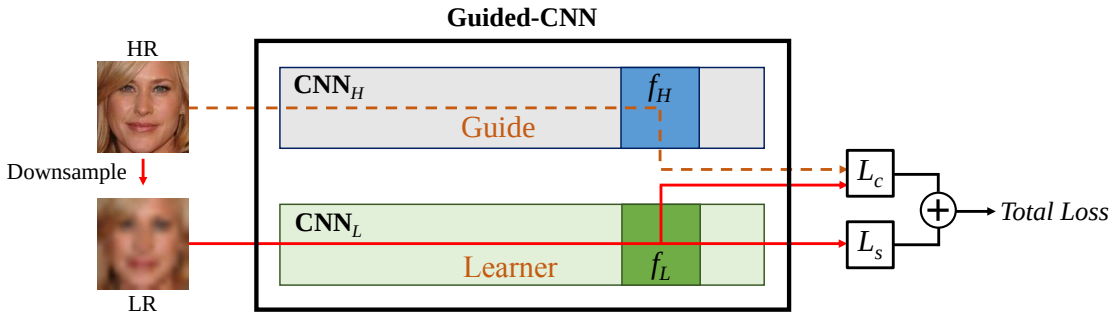

Fig. 2. The architecture of our proposed Guided-CNN for cross-resolution face recognition. During training, we downsample the HR images to be the LR inputs of $\mathbf{CNN}_L$. Softmax loss L_s is applied for learning the identification information of LR images and our unique cross-domain loss L_c associates the feature representations of LR images to the corresponding HR ones. In testing phase, we simply input HR/LR to $\mathbf{CNN}_H$/$\mathbf{CNN}_L$ respectively and calculate the similarities using the cosine distance of the associated features f_H/f_L.

adaptation model for relating HR and LR face images with recognition guarantees. Later in Section 2, we will detail our proposed model and explain why a unique loss function is required for our Guided-CNN to produce satisfactory performance on cross-resolution face recognition. Moreover, we will show that our Guided-CNN can be further applied to address robust face recognition in which the query face images are corrupted due to occlusion (up to 50% of the facial area).

The contributions of this paper are summarized as follow:

- We propose Guided-CNN, which consists of parallel CNN models (i.e., guide vs. learner) and unique loss functions for solving cross-resolution face recognition.
- By advancing existing CNN model as a guide, the parallel CNN model (i.e., learner) would be learned to handle within and cross-resolution face images.
- We show that, in addition to cross-resolution face recognition, our Guided-CNN can be applied to robust face recognition in which query images are corrupted due to occlusion.

2. GUIDED CONVOLUTIONAL NEURAL NETWORKS (GUIDED-CNN)

2.1. Architecture

To design CNN models for cross-resolution face recognition, we adopt the idea of learning similarity across domains [6] and propose to learn a unique model of Guided-CNN. As illustrated in Figure 2, our Guided-CNN consists of two sub-CNN models $\mathbf{CNN}_H$ and $\mathbf{CNN}_L$, dealing with the input HR and LR face images, respectively.

In our Guided-CNN architecture, the two sub-CNN models $\mathbf{CNN}_H$ and $\mathbf{CNN}_L$ have the identical structure but different configurations. Serving as a *guide*, $\mathbf{CNN}_H$ is pre-trained on HR face images (which we do not limit the use of any particular CNN architecture as). On the other hand, the input of $\mathbf{CNN}_L$ will be the LR images but upscaled to the same image size as the HR one. As a *learner*, this $\mathbf{CNN}_L$ is a parallel model with a different configuration, whose network parameters/weights will be learned from LR ones but share information across sub-CNN models.

It is worth noting that, our Guided-CNN is different form Siamese neural networks, which learn the same network configuration with shared parameters for two or more sub networks. Although with a similar goal of associating cross-domain data with comparable information, they require the training of the entire network architecture when a new cross-domain learning task is of interest. As for Guided-CNN, we can utilize any existing solution/model as the guide, and focus on the adaptation between the existing and the one of interest (e.g., images with different resolutions or corrupted regions). We also allow more than two sub-CNNs in our proposed architecture, which introduces additional robustness and flexibility in dealing with real-world recognition tasks.

2.2. Objectives

As noted in Section 2.1, our Guided-CNN applies existing CNN solutions as $\mathbf{CNN}_H$ and served as the guide. The other sub-CNN of $\mathbf{CNN}_L$ is to be learned by observing cross-resolution image data. Later in our experiments, we consider two recently proposed CNN models as $\mathbf{CNN}_H$ for verifying the effectiveness and robustness of our proposed architecture.

Given the same network structure as $\mathbf{CNN}_H$, we first consider the *softmax loss* function at the output layer of $\mathbf{CNN}_L$, which introduces the identification ability to $\mathbf{CNN}_L$ when observing LR image inputs. That is, for an input image x of class k, its softmax loss is calculated as:

$$L_s(x) = -log\frac{e^{y_k(x)}}{\sum_{j=1}^{n} e^{y_j(x)}}, \quad (1)$$

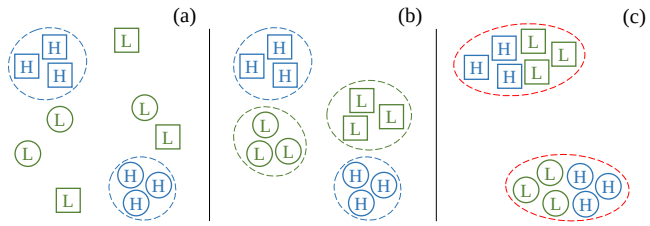

Fig. 3. Illustration of the derived deep feature spaces: (a) $\mathbf{CNN}_H$ only, (b) Guided-CNN using $\mathbf{CNN}_L$ with softmax loss only (i.e., $\lambda = 0$ in (3)), and (c) Guided-CNN with joint supervision of softmax and cross-domain losses. Note that instances in different shapes indicate images of different subjects, and H/L denote images in HR/LR.

Fig. 4. Examples of images of the LFW dataset with resolution and occlusion variations (from left to right: HR, LR, occlusion with 20% and 50%, respectively).

where $y_i(x)$ is the output for the i-th class for a given input x, and n denotes the number of classes.

Typically, standard CNN-based face recognition model takes the above loss function for learning identification loss. Take Figure 2 for example, the resulting representation f_L prior to the final fully connected layer would be taken as the feature to describe input LR images. However, such learning strategy would fail to incorporate the information retrieved from the associated HR images. This is the reason why we introduce an additional *cross-domain loss* function for bridging the knowledge across images with different resolutions.

In Guided-CNN, the cross-domain loss function calculates the Euclidean distance between the features of cross-resolution image pair, with the goal of enforcing the image features of the same identity yet with distinct resolutions to exhibit improved similarity. Thus, this loss function is defined as follows:

$$L_c(H_x, L_x) = \frac{1}{2}||f_{H_x} - f_{L_x}||_2^2, \tag{2}$$

where f_{H_x} and f_{L_x} represent the feature pair of input HR and LR images (i.e., H_x and L_x). Note that, l^2-normalization is performed for both f_{H_x} and f_{L_x} in (2), but we do not introduce additional variables for the sake of simplicity. By associating cross-resolution images of the same identity, the guided sub-CNN of $\mathbf{CNN}_L$ would exhibit improved ability in relating images across different resolutions, while the discriminative capability is preserved by observing the aforementioned softmax loss.

The idea of introducing the cross-domain loss for Guided-CNN is illustrated in Figure 3. Recall that, we apply state-of-the-art models as $\mathbf{CNN}_H$, which is pre-trained on HR images with promising data discriminating ability. During the training of $\mathbf{CNN}_L$ using $\mathbf{CNN}_H$ as a guide, our Guided-CNN not only preserves the separation between images with the same resolution but of different identities, the feature difference across resolutions is also jointly suppressed. As a result, the derived feature space can be expected to achieve satisfactory classification and (resolution) adaptation abilities.

It is worth noting that, our Guided-CNN does not incorporate the loss for enforcing the dissimilarity between cross-resolution image pairs with different identities. This is because that, adding such a loss function might affect our model for eliminating domain (resolution) differences between the image data of interest, and thus limit the performance of cross-resolution face recognition.

With both softmax and cross-domain losses introduced, the total loss for learning Guided-CNN is computed as:

$$\mathcal{L}(H_x, L_x) = L_s(L_x) + \lambda L_c(H_x, L_x), \tag{3}$$

where λ is the weight for regularizing the cross-domain loss.

3. EXPERIMENTS

To assess the performance of our proposed Guided-CNN, we first conduct experiments on cross-resolution face recognition in Section 3.1; in Section 3.2, we further consider robust face recognition in which gallery/query images are corrupted due to occlusion (see example images in Figure 4).

3.1. Cross-Resolution Face Recognition

To evaluate the performance of cross-resolution face recognition, we apply the CASIA-WebFace dataset [7] for training, and the LFW dataset [8] for testing. The CASIA-WebFace dataset contains 493,456 face images of 10,575 identities collected from the Internet, and the LFW dataset has 13,233 unconstrained face images of 5,749 identities.

Recall that, we do not require and limit the use of particular CNN models in our Guided-CNN. In our experiments, we utilzie two state-of-the-art CNN solutions for face recognition, Light CNN [9] and Center-loss CNN [5], as $\mathbf{CNN}_H$ in Guided-CNN. Pre-trained on HR images, the above $\mathbf{CNN}_H$ would serve as the guide for learning the parallel model $\mathbf{CNN}_L$, which has the same architecture as that of $\mathbf{CNN}_H$.

Prior to the training/testing of Guided-CNN, all face images are aligned with respect to the locations of the eyes and the mouth as the same pipeline of [9]. The input size of Light CNN and Center-Loss CNN are 128×128 and 120×96 pixels, respectively. For LR images, we follow the settings of [10, 11] to set the resolution of such images as 16×16

Table 1. Cross-resolution face verification using Light CNN [9] as the guide. Note that Ours* refers to our method with the softmax loss only (i.e., $\lambda = 0$ in (3)).

Method	Training Data	Testing (Gallery-Query)		
		HR-HR	LR-LR	HR-LR
$\mathbf{CNN}_H$ only	HR	**97.1**	-	84.5
$\mathbf{CNN}_L$ only	LR	-	92.7	86.9
[12]	HR & LR	91.5	91.8	89.0
Ours*	HR & LR	97.1	92.7	52.4
Ours	HR & LR	**97.1**	91.9	**93.7**

pixels. Thus, to synthesize LR images for training and testing purposes, we downsample HR images by a scaling factor of 8 and then upscale it by the same factor via bicubic interpolation.

We now compare our proposed method with three baseline/recent approaches. The baseline approach adopts the above state-of-the-art CNN models as $\mathbf{CNN}_H$, and the trained model is applied for performing HR and cross-resolution face recognition. We repeat the above process using LR images to train the associated model (denoted as $\mathbf{CNN}_L$ in Tables 1 and 2) for LR and cross-resolution face recognition. We consider a recent approach of [12] which applies the same CNN model with staged-training for addressing the same task. To be specific, the model first pre-trains on HR images and then continues to fine-tune on LR images.

To evaluate the performances, we report the equal error rate (EER) verification accuracy on 6,000 face pairs of LFW. The similarities of query/gallery face images are computed by the cosine distance of the associated features (f_H or f_L). From the results in Table 1, we see that the Light CNN trained on HR or LR images only were not able to produce satisfactory performance for cross-resolution face recognition (i.e., HR-LR with 84.5% and 86.9% only). While the recent adaptation model of [12] reported an improved accuracy of 89%, our Guided-CNN was able to achieve the best result of 93.7%. It is worth noting that, comparable results on only HR or LR inputs were also produced by our Guided-CNN (i.e., 97.1% and 91.9%).

We also vary λ in (3) from 0 to 100 to investigate the sensitiveness of the hyperparameter. The accuracies of cross-resolution verification for these models are shown in Figure 5. It is clear that using the softmax loss only (i.e. $\lambda = 0$) is not sufficient for cross-resolution face verification. We also observe that our models remain stable performance across a wide range of λ.

In addition, we experiments a special testing scenario of verifying an unknown resolution query image with respect to a HR gallery image. The features of the query image will be extracted from both $\mathbf{CNN}_H$ and $\mathbf{CNN}_L$. Then, we take the one that achieves shorter cosine distance with the feature of the HR gallery image into account. In this situation, our Guided-CNN also remain a comparable result of 93.4%, which helps us to implicitly attest the consistency between

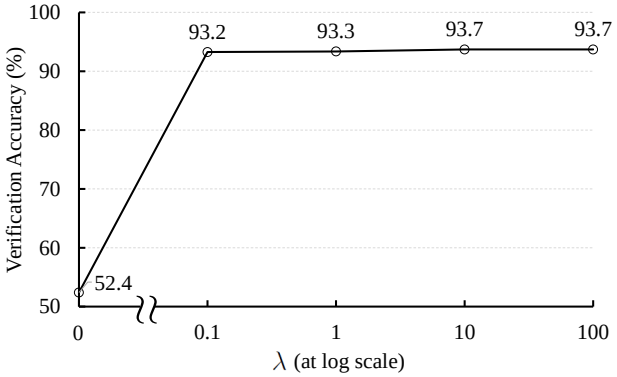

Fig. 5. Cross-resolution face verification using Light CNN [9] as the guide with different λ for regularizing the cross-domain loss.

Table 2. Cross-resolution face verification using Center-Loss CNN [5] as the guide. Note that Ours* refers to our method with the softmax loss only (i.e., $\lambda = 0$ in (3)).

Method	Training Data	Testing (Gallery-Query)		
		HR-HR	LR-LR	HR-LR
$\mathbf{CNN}_H$ only	HR	**97.4**	-	84.4
$\mathbf{CNN}_L$ only	LR	-	93.3	89.4
[12]	HR & LR	90.8	92.2	90.2
Ours*	HR & LR	97.4	93.3	48.8
Ours	HR & LR	**97.4**	92.1	**93.8**

HR and LR images in the deep feature space derived by cross-domain loss.

To verify that our Guided-CNN can utilize arbitrary existing CNN models as the guide, we consider the Center-Loss CNN as $\mathbf{CNN}_H$, and we repeat the above evaluation process. As shown in Table 2, our accuracy was 93.8%, which was higher than that of 90.2%$\sim$84.4% produced by baseline or recent approaches. Thus, the above results support the use of our Guided-CNN for cross-resolution face recognition, and its robustness of applying state-of-the-art CNN models in our proposed architecture.

3.2. Robust Face Recognition

In addition to cross-resolution face recognition, we further apply our method to address the task of robust face recognition, in which training/test images might be corrupted due to occlusion. To our tests, we randomly occlude 20% and 50% (see Figure 4) of the images for the sub-CNN $\mathbf{CNN}_L$ to handle. More specifically, we randomly blocked 20% contiguous areas of each HR images by setting their pixel values to 0; as for the 50% occlusion, we randomly blocked the upper or the bottom half of each HR images.

We follow all the settings (with Light CNN) in Section 3.1 for evaluation. That is, we simply replace LR images by the occluded ones, and we train Light CNNs on the occluded images for $\mathbf{CNN}_L$. The results are listed in Tables 3

Table 3. Robust face verification using Light CNN [9] as the guide. Note that $\mathbf{CNN}_L$ denotes the CNN trained on images with **20%** occlusion, and OC denotes occluded images, and Ours* refers to our method with the softmax loss only (i.e., $\lambda = 0$ in (3)).

Method	Training Data	Testing (Gallery-Query)		
		HR-HR	OC-OC	HR-OC
$\mathbf{CNN}_H$ only	HR	**97.1**	-	91.9
$\mathbf{CNN}_L$ only	OC	-	94.0	94.9
[12]	HR & OC	96.2	93.2	94.8
Ours*	HR & OC	97.1	94.0	52.0
Ours	HR & OC	**97.1**	94.1	**95.1**

Table 4. Robust face verification using Light CNN [9] as the guide. Note that $\mathbf{CNN}_L$ denotes the CNN trained on images with **50%** occlusion, and OC denotes occluded images, and Ours* refers to our method with the softmax loss only (i.e., $\lambda = 0$ in (3)).

Method	Training Data	Testing (Gallery-Query)		
		HR-HR	OC-OC	HR-OC
$\mathbf{CNN}_H$ only	HR	**97.1**	-	89.1
$\mathbf{CNN}_L$ only	OC	-	83.5	92.1
[12]	HR & OC	96.2	81.3	92.8
Ours*	HR & OC	97.1	83.5	50.1
Ours	HR & OC	**97.1**	84.1	**93.0**

and 4. From these two tables, we see that our method again performed favorably against baseline and state-of-the-art approaches. Thus, the effectiveness and robustness of our proposed Guided-CNN can be successfully verified.

4. CONCLUSION

We proposed Guided-CNN for solving cross-resolution face recognition problems. By advancing a CNN model pre-trained on HR images as a guide, our proposed architecture learns a parallel model on the LR ones with unique loss functions. The introduced loss functions allow us to jointly optimize the similarity for images within and across image resolutions. From our experiments, we confirmed that our method outperforms multiple baseline and recent approaches on cross-resolution face recognition, and the extension to robust face recognition was also successfully verified.

5. REFERENCES

[1] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[2] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[3] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in neural information processing systems*, 2014.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[5] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016.

[6] F. Wang, L. Kang, and Y. Li, "Sketch-based 3d shape retrieval using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

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

[8] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[9] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," *arXiv preprint arXiv:1511.02683*, 2015.

[10] W. W. Zou and P. C. Yuen, "Very low resolution face recognition problem," *IEEE Transactions on Image Processing*, 2012.

[11] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[12] X. Peng, J. Hoffman, X. Y. Stella, and K. Saenko, "Fine-to-coarse knowledge transfer for low-res image classification," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016.