

DeepVisage: Making face recognition simple yet with powerful generalization skills

Abul Hasnat¹, Julien Bohné², Jonathan Milgram², Stéphane Gentric², and Liming Chen¹

¹Laboratoire LIRIS, École centrale de Lyon, 69134 Ecully, France.

²Safran Identity & Security, 92130 Issy-les-Moulineaux, France.

md-abul.hasnat@ec-lyon.fr, {julien.bohne, stephane.gentric, jonathan.milgram}@safrangroup.com,
liming.chen@ec-lyon.fr

Abstract

Face recognition (FR) methods report significant performance by adopting the convolutional neural network (CNN) based learning methods. Although CNNs are mostly trained by optimizing the softmax loss, the recent trend shows an improvement of accuracy with different strategies, such as task-specific CNN learning with different loss functions, fine-tuning on target dataset, metric learning and concatenating features from multiple CNNs. Incorporating these tasks obviously requires additional efforts. Moreover, it demotivates the discovery of efficient CNN models for FR which are trained only with identity labels. We focus on this fact and propose an easily trainable and single CNN based FR method. Our CNN model exploits the residual learning framework. Additionally, it uses normalized features to compute the loss. Our extensive experiments show excellent generalization on different datasets. We obtain very competitive and state-of-the-art results on the LFW, IJB-A, YouTube faces and CACD datasets.

1. Introduction

Face recognition (FR) is one of the most demanding computer vision tasks, due to its practical use in numerous applications, such as biometric, surveillance and human-machine interaction. The state-of-the-art FR methods [34, 29, 31, 24, 20] surpassed human performance (97.53%) and achieved significant accuracy on the standard labeled faces in the wild (LFW) [14] benchmark. These remarkable results are achieved by training the deep convolutional neural network (CNN) [10] with large databases [11, 24, 44, 2].

The facial image databases mostly provide the identity labels. These labels allow the CNN models to be easily trained with the softmax loss. FR methods generally use

the trained CNN model to extract facial features and then perform verification by computing distance or recognition with a classifier. However, from our extensive study (see Sect. 2), we observe that recent methods include different additional strategies to obtain better performance, such as:

1. *train CNN with different loss functions* [29, 31]: requires carefully preparing the image pairs/triplets by maintaining certain constraints [29], because arbitrary pairs/triplets do not contribute to the training. Online triplet generation requires a larger batch size (e.g., [29] used 1.8K images in a mini-batch with 40 images/identity), which is excessive for a limited resource machine. On the other hand, using offline triplets can be critical as many of them will be useless while training progresses. The joint optimization [31] with Softmax and Contrastive losses not only requires specific training data (with identity and pair labels) but also complicates the training procedure.
2. *fine-tune CNN*: requires training on each target dataset, which restricts the ability to generalize.
3. *metric learning* [28, 9]: requires particular form of training data (e.g., triplets). Moreover, it does not always guarantee to enhance performance [37].
4. *concatenating features from multiple CNNs* [31, 20]: requires additional training data of different forms and train CNNs for each form. Besides, it is necessary to explore the particular modalities that can contribute to enhance performance.

The use of the above strategies requires significant efforts in terms of data preparation or selection and computing resources. On the other hand, recent results on the ImageNet challenge [26] indicate that deeper CNNs enhance

performance of different computer vision tasks. These observations raise the following question - *can we achieve state-of-the-art results with a single CNN model which is trained only once with the identity labels?* Our research is motivated by this question and we aim to address it by developing a simple yet robust single-CNN based FR method. Moreover, we want that our once-trained single CNN based FR method generalizes well across different datasets.

In this research, our primary objective is to discover an efficient CNN architecture. We follow the recent findings, which suggest that deeper CNNs perform better on a number of computer vision tasks [13, 10]. We construct a deep CNN model with 27 convolutional and 1 fully connected (FC) layers, which incorporates the residual learning framework [13]. Moreover, we aim to find an efficient way to train our CNN only with the identity labels. Recently, [39] achieves high FR performance with a CNN trained from the identity labels. However, they perform joint optimization using the softmax and center loss [39] (CL). CL improves the features discrimination among different classes. It follows the principle that, features learned from a deep CNN should minimize the intra-class distances. Interestingly, we observe (see Fig 3) that an equivalent representation can be achieved by normalizing the CNN features before computing the loss. Therefore, we train our CNN using the softmax loss with the normalized features.

With our single CNN model, first we evaluate on the LFW [14] benchmark and observe that it obtains 99.62% accuracy. In order to demonstrate its effectiveness, we evaluated it on different challenging face verification tasks, such as face templates matching on the IJB-A [16] dataset, video faces matching on the YouTube Faces [40] (YTF) dataset and cross age face matching on the CACD [3] dataset. Our method achieves 82.4% TAR@FAR=0.001 on IJB-A [16], 96.24% accuracy on YTF [40] and 99.13% accuracy on CACD [3]. These results indicate that our method achieves very competitive and state-of-the-art results. Moreover, it generalizes very well across different datasets.

We summarize our contributions as follows: (a) conduct extensive study and provide (Sec 2) a review and methodological comparison of the state-of-the-art methods; (b) propose (Sect. 3) an efficient single CNN based FR method; (c) conduct (Sect. 4) extensive experiments on different datasets, which demonstrate that our method has excellent generalization ability; and (d) perform (Sect. 4.3) an in-depth analysis to identify the influences of different aspects.

In the remaining part of this paper, first we study and analyze the state-of-the-art FR methods in Section 2, describe our proposed method in Section 3, present experimental results, perform analysis of our method and discuss them in Section 4 and finally draw conclusions in Section 5.

2. Related work, state-of-the-art FR methods

Face recognition (FR) in unconstrained environment attracts significant interest from the community. Recent methods exploited deep CNN models and achieved remarkable results on the LFW [14] benchmark. Besides, numerous methods have been evaluated on the IJB-A [16] dataset. We study¹ and analyze these methods based on several key aspects: (a) details of the CNN model; (b) loss functions used; (c) incorporation of additional learning strategy; (d) number of CNNs and (e) the training database used.

Recent methods tend to learn CNN based features using a *deep architecture* (e.g., 10 or more layers). This is inspired from the extraordinary success on the ImageNet [26] challenge by famous CNN architectures [10], such as AlexNet, VGGNet, GoogleNet, ResNet, etc. The FR methods commonly use these architectures as their baseline model (directly or slightly modified). For example, AlexNet is used by [27, 28, 21, 25, 1, 22, 29], VGGNet is used by [24, 8, 23, 1, 22, 9, 33] and GoogleNet is used by [42, 29]. CASIA-Webface [44] proposed a simpler CNN model, which is used by [37, 5, 9]. Several methods, such as [32, 34, 35, 39, 31] use a model with lower depth but increase its complexity with locally connected convolutional layers. Besides, [46] use 4 parallel 10 layers CNNs to learn features from different facial regions. *We follow the ResNet [13] based deep CNN model.*

FR methods often train multiple CNNs and accumulate features from all of them to construct the final facial descriptors. It provides an additional boost to the performance. Different types of inputs are used to train these multiple CNNs: (a) [32, 31, 33, 37, 9, 20] used image-crops focused on certain facial regions (eyes, nose, lips, etc.); (b) [9, 1, 22, 34] used different modality of input images, such as 2D, 3D, frontalized and synthesized faces at different poses and (c) [35, 20] used different training databases with varying number of images. *We do not follow this approach and train only one CNN.*

The CNN model parameters are learned by optimizing loss functions, which are defined based on the given task (e.g., classification, regression) and the available information (e.g., class labels, price). The softmax loss [10] is a common choice for classification tasks. It is often used by the FR methods to create good face representation by training the CNN as an identity classifier. It requires only the identity labels. The contrastive loss [10, 7] is used by [32, 34, 33, 31, 42] for face verification and requires face image pairs and similarity labels. The triplet loss [29] is used by [29, 24, 8, 20] for face verification and requires the face triplets. Recently the center loss [39] is proposed to enhance feature discrimination, which uses the identity labels.

¹We consider only the CNN based methods. For the others, we refer readers to the recently published survey [17] for LFW and [16] for IJB-A.

We use the softmax loss.

Several methods use multiple loss functions and train CNN using joint optimization [32, 33, 31, 39, 25]. The other way is to use them sequentially [34, 24, 8, 20, 42], i.e., first train with the softmax and then train with the other loss. We observe that using multiple loss functions complicates the training data preparation task and the CNN training procedure. *Therefore, we avoid this type of strategies.*

Fine-tuning the CNN parameters is a particular form of transfer learning. It is commonly employed by several methods [37, 5, 27] on the IJB-A [16] dataset. It refines the CNN parameters from a previously learned model using a target specific training dataset. Several methods do not directly use the raw CNN features but employ an additional learning strategy. The *metric/distance learning* strategy based on the Joint Bayesian method [4] is a popular one and used by [32, 44, 37, 5, 33, 31, 9]. Recently, two different strategies [28, 28] have been proposed to learn feature embedding using face triplets. Another strategy, called template adaptation [8], exploits an additional SVM classifier. Apart from these, principal component analysis (PCA) is used by several methods [23, 1, 22] to learn a dataset specific projection matrix. [42] learns an aggregation module to compute scores among two videos. The above methods often need training data from the target datasets. Moreover, they [27, 28] may need to carefully prepare the training data, e.g., triplets. *We do not need any such learning strategies.*

The use of a large facial training dataset is important to achieve high FR accuracy [29, 46]. [46] provided an in-depth analysis and demonstrated the effect of the dataset size and the number of identities for FR. Following the high demand of a large FR dataset, several publicly available datasets have been released recently. Among them, CASIA-WebFace [44] is used by numerous methods [39, 27, 28, 21, 25, 44, 37, 5, 9, 41, 23, 1, 22]. Several researches [23, 1, 22] enlarge it by synthesizing facial images with different shapes and poses based on the 3D face models. Recently, the MSCeleb [11] dataset has been publicly released. It contains the largest collection of facial images and identities. *We exploit it to develop our FR method.*

3. Proposed Method

Our FR method, called *DeepVisage*, consists in pre-processing face image, learning CNN based facial features and computing similarity. Following the recent trend [34, 29, 31, 24, 44], we exploit the CNN as the core component. Our deep CNN model follows the residual learning framework [13]. Moreover, it intelligently exploits feature normalization, which is a crucial step, see Sect. 4.3. Our pre-processing stage consists in the detection of the face and facial landmarks, which are used to create a normalized face image. We compute the cosine similarity among the features of a pair of faces as the verification score. Below,

we describe these elements.

3.1. Building blocks and deep CNN architecture

Convolutional networks: We begin with the basic ideas of CNN [18]: (a) local receptive fields with identical weights via the convolution operation and (b) spatial sub-sampling via the pooling operation. At a particular layer l , the convolution of the input $f_{x,y}^{Op,l-1}$ to obtain the k^{th} output feature map $f_{x,y,k}^{C,l}$, can be expressed as:

$$f_{x,y,k}^{C,l} = \mathbf{w}_k^l {}^T f_{x,y}^{Op,l-1} + b_k^l \quad (1)$$

where, \mathbf{w}_k^l and b_k^l are the shared weights and bias. C denotes convolution and Op (for $l > 1$) denotes various tasks, such as convolution, sub-sampling or activation. For $l = 1$, Op represents the input image. Sub-sampling or pooling performs a simple local operation, such as computing the average or maximum value in a local spatial neighborhood followed by reducing spatial resolution. We apply max pooling for our CNN, which has the following form:

$$f_{x,y,k}^{P,l} = \max_{(m,n) \in \mathcal{N}_{x,y}} f_{m,n,k}^{Op,l-1} \quad (2)$$

where, $\mathcal{N}_{x,y}$ denotes the local spatial neighborhood of (x, y) coordinate and P denotes the pooling operation.

In order to ensure non-linearity of the network, the feature maps are passed through a non-linear activation function, e.g., the Rectified Linear Unit (ReLU) [10, 12]: $f_{x,y,k}^l = \max(f_{x,y,k}^{l-1}, 0)$. We apply the Parametric Rectified Linear Unit (PReLU) [12] as the activation function, which has the following form:

$$f_{x,y,k}^{A,l} = \max(f_{x,y,k}^{Op,l-1}, 0) + \lambda_k \min(f_{x,y,k}^{Op,l-1}, 0) \quad (3)$$

where, λ_k is a trainable parameter to control the slope of the linear function for the negative input values and A denotes activation operation.

At the basic level, a CNN is constructed by stacking series of convolution, activation and pooling layers, see LeNet-5 [18] for an example. Often a layer with full connections is placed at the end of the stacked layers, called the fully connected (FC) layer. It takes all points (neurons) from the previous layer as input and connects it to all points (neurons) of the output layer.

Residual learning framework [13]: A recent trend [10] on the ImageNet [26] challenge shows that deeper CNNs achieve better results. However, it increases the model complexity, which makes it harder to optimize the loss of the CNN model. Besides, they may generate higher training error than a shallower CNN [13]. The residual learning framework [13] provides a solution to these problems.

For a stack of a few layers, residual learning fits a mapping $\mathcal{F}(f) := \mathcal{H}(f) - f$ instead of fitting the underlying

Table 6: Study the influences from CNN related issues. All CNN models are trained with the CASIA [44] dataset. CL- center loss [39], FN- feature normalization. **CN-mod** modifies the *Cas-Net* [44] by replacing *Pool* layer with a *FC* layer of 512 neurons.

Settings	# params	Acc %	T@F 0.01
Base-CNN (<i>proposed</i>)	40.5M	99.00	0.988
Base-CNN - FN	40.5M	97.40	0.954
Base-CNN + CL - FN	44.8M	98.85	0.986
VGG-Net [24]	182M	95.15	0.883
Cas-Net [44]	6M	97.10	0.938
CN-mod	8M	97.50	0.956

ers. We optimize it using the softmax loss. Fig. 3 provides the illustration, from which we observe that: (a) FN provides a better feature discrimination in the normalized 2D space, see Fig. 3-b; (b) CL enforces the features towards its representative center and hence shows discrimination, see Fig. 3-c and (c) CL+FN does not provide much additional discrimination, see Fig. 3-b and Fig. 3-d. These observations reveal that, by exploiting the FN appropriately we can ensure feature discrimination and hence no additional loss function, e.g., CL, is necessary.

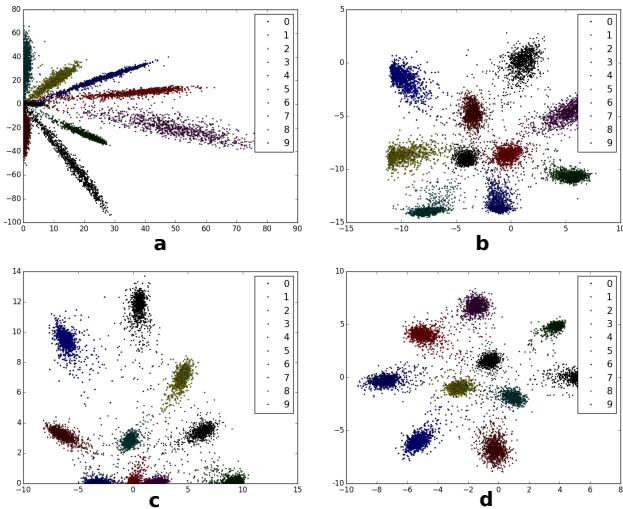


Figure 3: 2D visualization of the MNIST [18] digits features, which are obtained by using same baseline CNN model and training settings. CL [39] parameters are set to $\lambda = 0.003$ and $\alpha = 0.5$. **a.** CNN without FN and CL; **b.** CNN with FN; **c.** CNN with CL; and **d.** CNN with FN and CL.

Finally, we investigate the incorrect results by observing the face image pairs in which *DeepVisage* failed. Appendix A provides the illustrations of the false accept/reject cases from the different datasets. We observe that, on LFW it failed (11/20 error cases) when the eyes are occluded by glasses or a cap. Incorrect CACD results and higher false rejection rate indicate that our method (although provides best accuracy) encounters difficulties to recognize the same person from the images of different ages. Incorrect results

from YTF often suffers from high pose and perhaps low image resolution. IJB-A results reveal that our method needs to take care of the face images with extreme pose variations. Indeed, during the IJB-A experiments, we are forced to keep a large number of images as un-normalized due to the failure of landmarks detection for them. Based on empirical evidences, we believe that these un-normalized faces cause the degradation of our performance. Besides, the results from YTF and IJB-A indicate that we may need to use a better distance computation strategy.

5. Conclusion

In this paper we present a single-CNN based FR method which achieves state-of-the-art performance and exhibits excellent ability of generalize across different FR datasets. Our method, called *DeepVisage*, performs face verification based on a given pair of single images, templates and videos. It consists in a deep CNN model which is simple and straightforward to train. Overall, *DeepVisage* is very easy to implement, thanks to the residual learning framework, feature normalization, softmax loss and the simplest distance. It successfully demonstrates that, in order to achieve state-of-the-art results it is not necessary to develop a complicated FR method by using complex training data preparation and CNN learning procedure. We foresee several future perspectives of this work, such as: (a) train CNN with a larger and more balanced dataset, which can be constructed by combining multiple publicly available datasets or by adopting the face synthesizing strategy [23] with the existing one; (b) enhance FR performance by incorporating failure detection based technique [30], particularly for face and landmarks detection and (c) incorporate better distance computation method for the template and video comparison, e.g., use softmax based distance [23].

A. Analysis of the incorrect results

In this section, we provide examples of the incorrect results observed from the face verification experiments on different datasets.

LFW [14]: Figure 4 provides the examples of the failure cases on the LFW [14] benchmark. The ratio of false accept vs reject is **1:1.56**. Note that our method achieves 99.62% accuracy on LFW. In Figure 4(b) three pairs are marked with red colored rectangles. These pairs are erroneously labeled in the dataset, which means our method makes correct judgment on them and hence the accuracy further increases to 99.67% by considering them as correct match.

CACD-VS [3]: Figure 5 provides the examples of the failure cases on the CACD [14] dataset. The ratio of false accept vs reject is **1:6**.



(a)



(b)

Figure 4: Illustration of the false accepted/rejected image pairs from the LFW [14] benchmark. (a) false accepted pairs and (b) false rejected pairs. The red colored rectangles indicate the examples which were erroneously labeled in the dataset.

YTF [40]: Figure 6 provides few examples of the failure cases on the YTF [40] dataset. The ratio of false accept vs reject is **1:2.2**. In Figure 6, we only show top three mistakes (sorted based on their similarity score) in terms of false accept and reject.

IJB-A [16]: Figure 7 provides few examples of the IJB-A [16] failure cases. The ratio of false accept vs reject is **1:5.15**. In Figure 7, we only show top three mistakes (sorted based on their similarity score). From the falsely rejected template pairs, we observe that: (a) one pair has only one image in the template; (b) the pre-processor fails to detect face as well as landmarks and (c) the images in the template have very high pose and large occlusion which causes important face attributes to be absent.



(a)



(b)

Figure 5: Illustration of the false accepted/rejected image pairs from the CACD-VS [3] dataset. (a) false accepted pairs and (b) false rejected pairs.

References

- [1] W. AbdAlmageed, Y. Wu, S. Rawls, S. Harel, T. Hassner, I. Masi, J. Choi, J. Lekust, J. Kim, and P. Natarajan. Face



Figure 6: Illustration of the false accepted/rejected video pairs from the YTF [40] dataset. (a) false accepted pairs and (b) false rejected pairs.

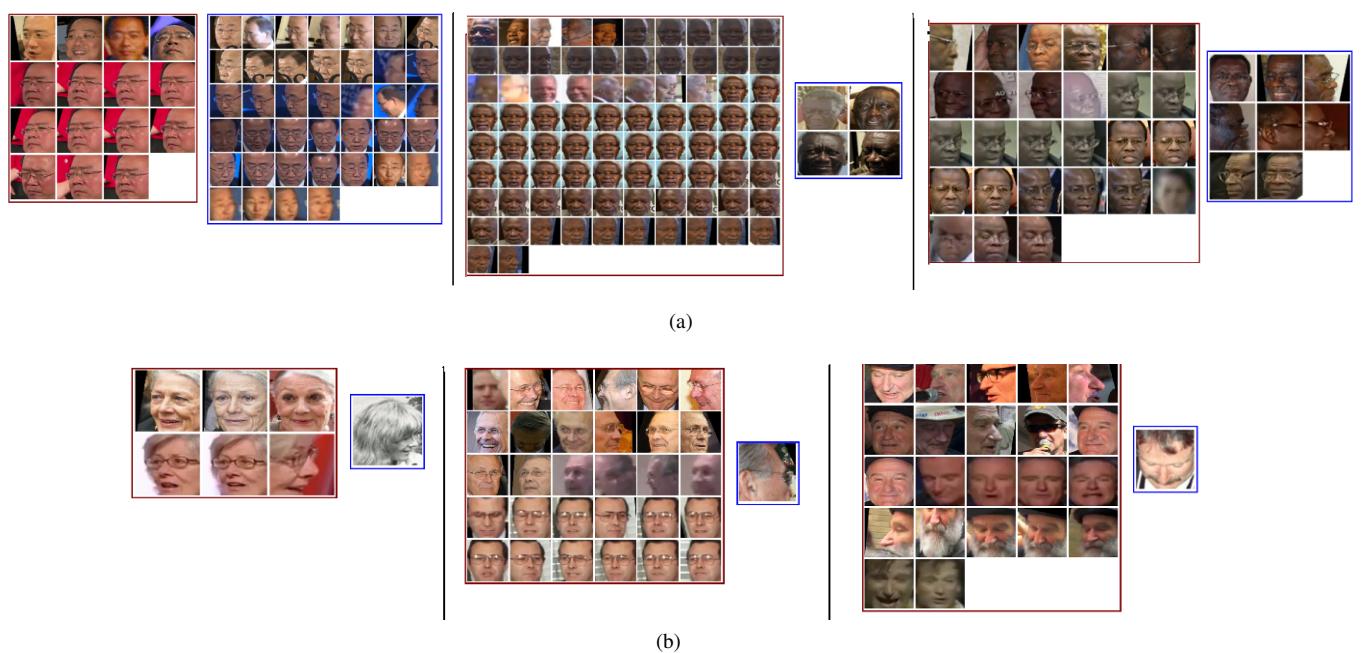


Figure 7: Illustration of the false accepted/rejected template pairs from the IJB-A [16] dataset. (a) false accepted pairs and (b) false rejected pairs.

- recognition using deep multi-pose representations. In *IEEE WACV*, pages 1–9, 2016.
- [2] A. Bansal, A. Nanduri, C. Castillo, R. Ranjan, and R. Chellappa. Umdfaces: An annotated face dataset for training deep networks. *arXiv:1611.01484*, 2016.

- [3] B.-C. Chen, C.-S. Chen, and W. H. Hsu. Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset. *IEEE Trans. on Multimedia*, 17(6):804–815, 2015.
- [4] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun. Bayesian

- [42] J. Yang, P. Ren, D. Chen, F. Wen, H. Li, and G. Hua. Neural aggregation network for video face recognition. *arXiv:1603.05474*, 2016.
- [43] H. Ye, W. Shao, H. Wang, J. Ma, L. Wang, Y. Zheng, and X. Xue. Face recognition via active annotation and learning. In *Proc. of ACM Multimedia*, pages 1058–1062. ACM, 2016.
- [44] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014.
- [45] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, Oct 2016.
- [46] E. Zhou, Z. Cao, and Q. Yin. Naive-deep face recognition: Touching the limit of Ifw benchmark or not? *arXiv:1501.04690*, 2015.