Improving Multi-Scale Face Recognition using VGGFace2

Fabio Valerio Massoli[0000−0001−6447−1301], Giuseppe Amato[0000−0003−0171−4315], Fabrizio Falchi[0000−0001−6258−5313], Claudio Gennaro[0000−0002−0967−5050], and Claudio Vairo[0000−0003−2740−4331]

ISTI-CNR, via G. Moruzzi 1, 56124 Pisa, Italy
{fabio.massoli, giuseppe.amato, fabrizio.falchi, claudio.gennaro, claudio.vairo}@isti.cnr.it

Abstract Convolutional neural networks have reached extremely high performances on the Face Recognition task. These models are commonly trained by using high-resolution images and for this reason, their discrimination ability is usually degraded when they are tested against low-resolution images. Thus, Low-Resolution Face Recognition remains an open challenge for deep learning models. Such a scenario is of particular interest for surveillance systems in which it usually happens that a low-resolution probe has to be matched with higher resolution galleries. This task can be especially hard to accomplish since the probe can have resolutions as low as 8, 16 and 24 pixels per side while the typical input of state-of-the-art neural network is 224. In this paper, we described the training campaign we used to fine-tune a ResNet-50 architecture, with Squeeze-and-Excitation blocks, on the tasks of very low and mixed resolutions face recognition. For the training process we used the VGGFace2 dataset and then we tested the performance of the final model on the IJB-B dataset; in particular, we tested the neural network on the 1:1 verification task. In our experiments we considered two different scenarios: 1) probe and gallery with same resolution; 2) probe and gallery with mixed resolutions.

Experimental results show that with our approach it is possible to improve upon state-of-the-art models performance on the low and mixed resolution face recognition tasks with a negligible loss at very high resolutions.

Keywords: Low-Resolution face recognition · Convolutional Neural Networks · Face verification.

1 Introduction

Face Recognition (FR) is nowadays among the hottest topic in computer vision. Thanks to the extremely high computational power reached by the modern

M. Cristani et al. (Eds.): ICIAP 2019 Workshops, LNCS 11808, pp. 21-29, 2019.
Final authenticated publication: https://doi.org/10.1007/978-3-030-30754-7_3

GPUs, the use of Deep Learning techniques [7], [6] has become state-of-the-art to solve FR task. Even though such algorithms perform well when tested against images taken under controlled conditions, e.g. high-resolution and frontal pose, a sudden drop in their performance has been observed when they are tested against images taken under uncontrolled conditions, e.g. low-resolution. For example, this situation occurs in the context of surveillance systems [12] which typically rely on cameras with limited resolution. Thus, a probe with variable low resolution has to be matched against high-resolution galleries.

Two common techniques used to deal with low-resolution face recognition tasks are the Super Resolution [11], [9] and the projection of the LR probe and the high-resolution (HR) gallery into a common space [4].

In our work we address the tasks of low and mixed resolutions face recognition by fine-tuning a Deep Convolutional Neural Network on very low-resolution images. In order to fine-tuned the original model, we used two random extractions to decide whether to down sample the image and at which resolution, comprised in the range of [8, 256] pixels. A schematic view of the procedure is shown in Figure 1.

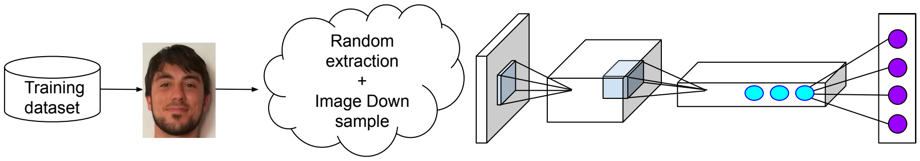

Figure 1. Schematic view of the training procedure.

After the training phase, we tested the model on the 1:1 verification protocol on the IARPA Janus Benchmark-B (IJB-B) dataset [8]. In particular, we considered the case of probe and gallery with the same resolution and the even more challenging scenario of probe and gallery with mixed resolution.

2 Related Works

In this section, we briefly mentioned some related works.

Super Resolution [9] (SR, or Face Hallucination) is a procedure in which a deep model is used to synthesize a high-resolution image starting from a low-resolution probe. A frequent objection to this technique is that the synthesis process is not optimized for discrimination tasks. Thus, the information about the identity of the person might be lost. In [11], the attempt is to address this issue by introducing an identity loss in order to impose the identity information during the training process. Hence, the aim is to recover the initial low-resolution image identity in the high-resolution one.

In [10], a two-branches model which learns nonlinear transformations in order to map high and low-resolution images into a common space has been proposed.

One branch accepts the HR images as inputs while the other one admits the LR images. Both branches produce features vectors that are then projected into a common space where their distance is evaluated. The error on the distance is then backprogated only through the bottom branch with the goal of minimizing the distance between features vectors of HR and LR images.
In [5], Shekhar et al. tackled the LR FR problem by means of a generative technique rather than a discriminative one, building their approach on learning class specific dictionaries that, moreover, resulted to be robust with respect to illumination variations.

3 Methodology

3.1 Datasets Description

We used two different datasets in order to train and test the performance of the model. Specifically, we used the VGGFace2 [1] dataset for training. It contains 3.31 million images from 9131 identities with a high variation in pose and ages of the subjects. Among all the identities, 500 of them were originally designed for test purposes. For this reason, we used only 8631 identities during our training. Instead, for the test procedure, we used the IJB-B [8] dataset that has been designed for test purposes only. It contains more than 76K images, 21K from still images and 55K frames from 7K videos, corresponding to 1845 different identities. We used it in order to measure the performance of our neural net on the 1:1 verification protocol. For this task, we needed to evaluate the similarity between face descriptors represented as L2-normalized 2048-d deep features vectors.
During the test phase, we used different down sampled versions of the IJB-B dataset in order to test the model on the 1:1 verification protocol with descriptors at the same and mixed resolutions.

3.2 Training Strategy

By using the VGGFace2 training dataset we fine-tuned the pre-trained model from [1]. It is a ResNet-50 [2] architecture equipped with Squeeze-and-Excitation blocks [3]. For the fine-tuning we made a first trial in which we kept the entire net frozen with the exception of the last fully connected layer. However, we obtained better results by fine-tuning the entire neural net. The intuition was that there were patterns in the new low-resolution images that the model needed to adjust for our goals. In the very first steps of our experiments, we tested various hyperparameters such as batch size, learning rate, momentum and weight decay for the optimizer. We obtained the best results setting the initial value for the learning rate at $5.e^{-4}$ and dropping its value by a factor of 5 every time the loss reached a plateau. We used a batch size of 256 and a weight decay of $1e^{-5}$.
During the fine-tuning of the model, for each input image, we used two random extractions. The two extracted numbers, a float and an integer, were uniformly distributed in the range [0, 1] and [3, 8], respectively. The first one was used

in order to decide if to down sample an image or not. Each image was down sampled if the the first extracted number was below a specific “down-sampling probability” that we fixed before training. Instead, the second one was used, as exponent of a power of two, in order to decide at which resolution to down sample the image. Specifically, the image was first down sampled so that the shortest side was equal to the extracted resolution, while keeping the original aspect ratio of the image, and then it was resized at the original dimensions. In both the down sampling and up sampling operations we used the bilinear interpolation algorithm from the PIL python library.
In Figure 2 an example of down sampled images at various resolutions is shown.

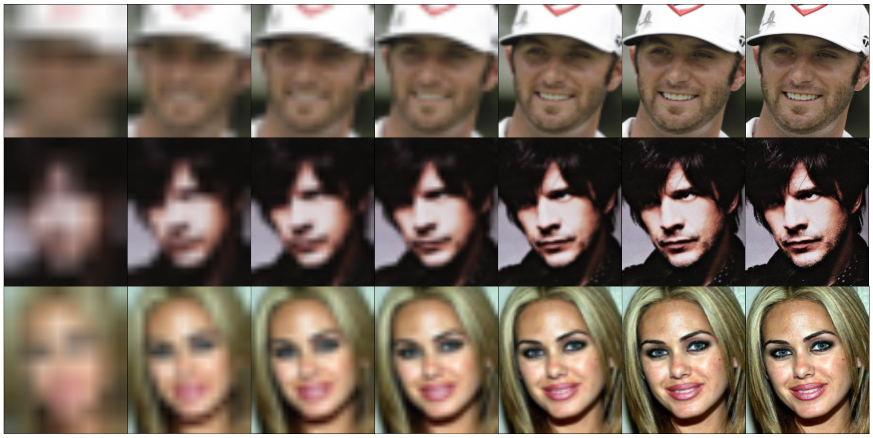

Figure 2. From left to right: example of images down sampled at resolution of 8, 16, 24, 32, 64, 128 and 256 pixels.

After this first preprocessing phase, the images were resized so that the size of the shortest side was 256 pixels, then a random crop was applied in order to select a 224x224 pixels region which matched the input of the network.
We split the training dataset into training set and validation set. Specifically, during the training phase, we used two versions of the latter split in order to monitor the performance of the model on both low and high-resolution domains. On one version, we down sampled all the images to a reference resolution of 24 pixels and on the other one we used them at full resolution. In both cases the images were then resized at 256 pixels and a center crop was applied.

4 1:1 Face Verification Results

In this section we present our experimental results obtained by using the 1:1 verification protocol on the IJB-B dataset.
As a base line, we chose the state-of-the-art model showed in Table VII [1] (penultimate row) and we first tried to reproduce that result. As a reference,

we considered the 1:1 verification True Acceptance Rate (TAR) value at a False Acceptance Rate (FAR) equals to 1.e^{-3}. We obtained a value of 0.898 which is 0.01 below the quoted 0.908 value. A difference of 1% can be attributed to a possible difference in the image crop and preprocessing phases. This difference is not fundamental to our analysis since we considered the relative improvement.
As a first result, in Figure 3 we show the Receiver Operating Characteristic (ROC) curves we obtained considering probe and gallery at the same resolution. Results at 128 pixels are not shown since they are similar to the ones at the original size.

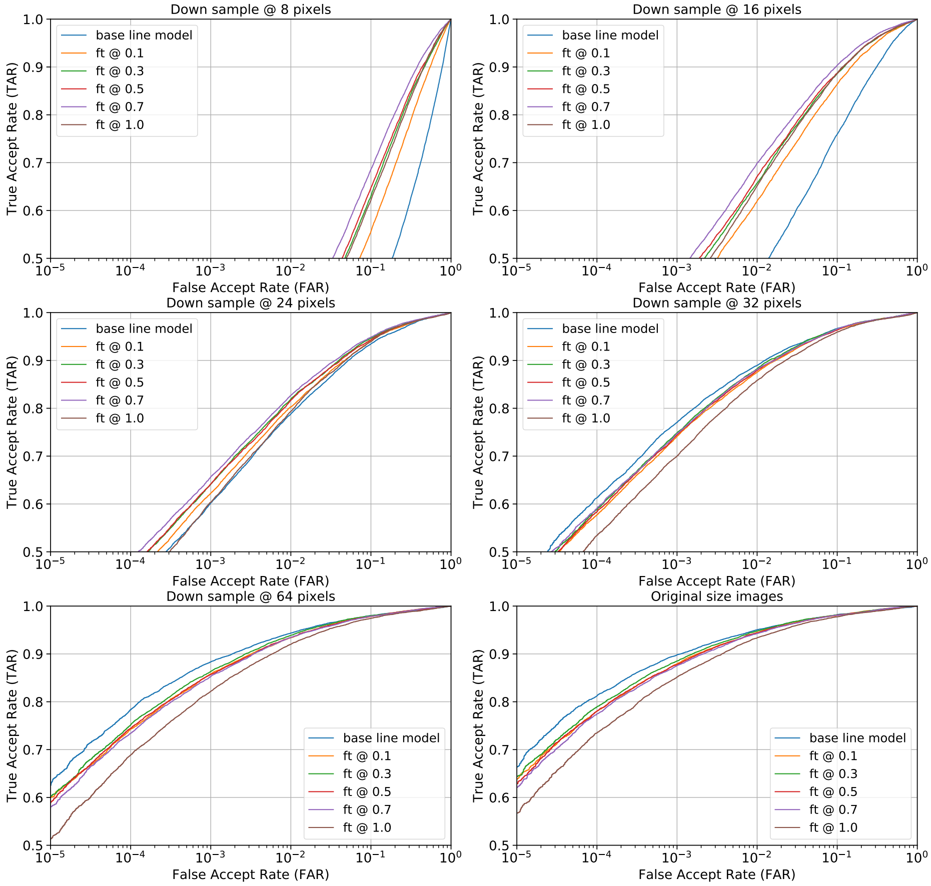

Figure 3. ROC curves, for the 1:1 verification protocol on the IJB-B dataset, for different values of the image resolution. The pre-trained model has been reported as "base line model" while the "ft" models are the fine tuned ones. The value after the "@" symbol in the legend, represents the probability we used, while training the model, in order to decide whether to reduce the input resolution or not. Results at 128 pixels are not shown since they are similar to the ones at the original size.

As we can see from Figure 3, the models display a notable improvement for resolution up to 24 pixels. In order to highlight these achievements with respect to the original model we considered a reference value for TAR at FAR = $1.e^{-3}$. The results are shown in Table 1. We especially acknowledged the largest margin of improvement at 16 pixels.

Table 1. True Acceptance Rate (TAR @ FAR = $1.e^{-3}$), for different values of the images resolution, from the 1:1 verification protocol on the IJB-B dataset. Face descriptors are considered at the same resolution. Given a specific image resolution, the various rows correspond to different values for the down sample probability (indicated in the second column) used in the training phase. The first row reports the results of the original pre-trained model.

Architecture	**Down-sampling**	**TAR (@ FAR = 1.e−3)**						
	probability	**8**	**16**	**24**	**32**	**64**	**128**	**256**
Se-ResNet-50 [1]		4.8	24.4	60.2	**77.0**	**88.3**	**89.5**	**89.8**
Se-ResNet-50_ft	0.1	8.5	38.6	62.2	74.1	85.9	87.7	88.0
Se-ResNet-50_ft	0.3	10.4	42.5	64.1	74.9	86.3	88.2	88.5
Se-ResNet-50_ft	0.5	10.7	44.2	64.2	74.4	85.7	87.5	87.8
Se-ResNet-50_ft	0.7	**13.9**	**46.7**	**65.7**	74.7	85.3	87.2	87.5
Se-ResNet-50_ft	1.0	9.4	40.8	60.4	70.0	82.2	84.7	85.1

Also, we carried out several experiments in which we undertook a "more linear" approach, i.e. we considered a fixed value for the input image resolution and kept it steady during the whole training run. The results for the 1:1 verification protocol we obtained in this scenario are reported in Table 2 from which it is clear that they are much worse than the ones shown in Table 1.

Table 2. TAR @ FAR = $1.e^{-3}$, for different values of the images resolution, from the 1:1 verification protocol on the IJB-B dataset. First row: results obtained using training images always down sampled at 24 pixels. Second row: results obtained considering a random down sample of the training images (the numbers correspond to the fifth row in Table 1).

Architecture	**TAR (@ FAR = 1.e−3)**						
	8	**16**	**24**	**32**	**64**	**128**	**256**
Se-ResNet-50_ft[1]	3.9	9.7	32.5	52.2	68.1	73.2	75.0
Se-ResNet-50_ft[2]	13.9	46.7	65.7	74.7	85.3	87.2	87.5

[1] Fixed resolution
[2] Randomly selected resolution

A possible justification for such a drop in the performance is that low-resolution images do not carry valuable knowledge, as high-resolution ones do, thus the model is not able to learn enough discriminative features.

Up to now, we have contemplated the case in which we evaluated the similarity between face descriptors at the same resolution. Our study became even more challenging when we analyzed the mixed scenario in which descriptors had different resolutions. Perhaps, these are the most interesting results regarding surveillance systems application. Table 3 reports the results of this study considering the fine tuned model from Table 1 for which we set the down sampling threshold to 0.7. The numbers represent the value of the TAR at FAR = 1.e^{-3} while in between brackets we have reported, as a reference, the results from the original model. As it is clear from Table 3, even though our training causes a small decrease in the performance at resolution levels higher than or equal to 32 pixels, the improvements for lower resolutions completely outweigh the mentioned loss.

Table 3. True Acceptance Rate (TAR @ FAR = 1.e^{-3}) for mix-resolution face verification with a model trained by using a down sample probability of 0.7. For comparison, between brackets we reported the value from the original model.

		Resolution (pixel)						
		8	**16**	**24**	**32**	**64**	**128**	**256**
Resolution (pixel)	**8**	**13.9** (4.8)						
	16	**11.1** (0.2)	**46.7** (24.4)					
	24	**8.5** (0.2)	**51.0** (18.3)	**65.7** (60.2)				
	32	**7.2** (0.2)	**50.0** (9.0)	**69.1** (65.4)	74.7 (77.0)			
	64	**5.6** (0.2)	**44.2** (2.9)	**69.3** (60.4)	78.1 (80.5)	85.3 (88.3)		
	128	**5.3** (0.3)	**41.9** (2.4)	**68.2** (57.9)	78.1 (80.1)	86.2 (88.9)	87.2 (89.7)	
	256	**5.2** (0.3)	**41.5** (2.3)	**68.0** (57.5)	78.2 (80.1)	86.3 (89.0)	87.4 (89.7)	87.5 (89.8)

5 Conclusions

In this paper we proposed a training approach to fine-tune a CNN architecture on the task of low and mixed resolution Face Recognition. We tested two different training strategies. In the first case, we let the image resolution be randomly extracted while in the second one we kept it fixed. According to our measurements, the first strategy gave us the best results for both low resolution and mixed resolution tasks. In both cases we have observed a drop within 3% percent in the model performance at resolutions strictly higher than 24 pixels but the decrease has been outweighed by the improvement we obtained in the low resolution regime. In particular, regarding the face verification task with both templates at the same low resolution, we measured a TAR value (at a reference value of the FAR = 1.e^{-3}) of 65.7%, 46.7% and 13.9% considering resolutions of 24, 16 and 8 pixels respectively.

The improvements we reached on the mixed resolution face verification task, i.e.

when probe and gallery have different resolutions, are even more interesting for surveillance system applications. To the best of our knowledge this is the first paper that investigate these topics.
As already said, in this case too we measured a drop of a few percent in the performance of the model for resolutions from 32 pixels and above, but the improvements at 24 pixels and below totally outweighed that decrease.
Specifically, we observed the higher improvements considering a down sample probability equal to 0.7.
To conclude, the measurements showed that our training campaign has been highly effective showing improvements in the low and mixed resolution face verification tasks up to a factor 30 w.t.r. to the base line model.

Acknowledgments. This publication is based upon work from COST Action 16101 “MULTI-modal Imaging of FOREnsic SciEnce Evidence” (MULTI-FORESEE), supported by COST (European Cooperation in Science and Technology).

References

1. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
3. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
4. Jian, M., Lam, K.M.: Simultaneous hallucination and recognition of low-resolution faces based on singular value decomposition. IEEE Transactions on Circuits and Systems for Video Technology **25**(11), 1761–1772 (2015)
5. Shekhar, S., Patel, V.M., Chellappa, R.: Synthesis-based robust low resolution face recognition. arXiv preprint arXiv:1707.02733 (2017)
6. Wang, F., Cheng, J., Liu, W., Liu, H.: Additive margin softmax for face verification. IEEE Signal Processing Letters **25**(7), 926–930 (2018)
7. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European conference on computer vision. pp. 499–515. Springer (2016)
8. Whitelam, C., Taborsky, E., Blanton, A., Maze, B., Adams, J., Miller, T., Kalka, N., Jain, A.K., Duncan, J.A., Allen, K., et al.: Iarpa janus benchmark-b face dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 90–98 (2017)
9. Yu, X., Fernando, B., Ghanem, B., Porikli, F., Hartley, R.: Face super-resolution guided by facial component heatmaps. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 217–233 (2018)
10. Zangeneh, E., Rahmati, M., Mohsenzadeh, Y.: Low resolution face recognition using a two-branch deep convolutional neural network architecture. arXiv preprint arXiv:1706.06247 (2017)

11. Zhang, K., Zhang, Z., Cheng, C.W., Hsu, W.H., Qiao, Y., Liu, W., Zhang, T.: Super-identity convolutional neural network for face hallucination. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 183–198 (2018)
12. Zou, W.W., Yuen, P.C.: Very low resolution face recognition problem. IEEE Transactions on image processing **21**(1), 327–340 (2012)