CNN-based System for Low Resolution Face Recognition (Discussion Paper)

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 Since the publication of the AlexNet in 2012, Deep Convolutional Neural Network models became the most promising and powerful technique for image representation. Specifically, the ability of their inner layers to extract high level abstractions of the input images, called deep features vectors, has been employed. Such vectors live in a high dimensional space in which an inner product and thus a metric is defined. The latter allows to carry out similarity measurements among them. This property is particularly useful in order to accomplish tasks such as Face Recognition. Indeed, in order to identify a person it is possible to compare deep features, used as face descriptors, from different identities by means of their similarities. Surveillance systems, among others, utilize this technique. To be precise, deep features extracted from probe images are matched against a database of descriptors from known identities.
A critical point is that the database typically contains features extracted from high resolution images while the probes, taken by surveillance cameras, can be at a very low resolution. Therefore, it is mandatory to have a neural network which is able to extract deep features that are robust with respect to resolution variations.
In this paper we discuss a CNN-based pipeline that we built for the task of Face Recognition among images with different resolution. The entire system relies on the ability of a CNN to extract deep features that can be used to perform a similarity search in order to fulfill the face recognition task.

Keywords: Convolutional Neural Networks · Face Recognition · Ensemble Methods.

1 Introduction

Content based image retrieval (CBIR) is one of the most active research field in the computer vision community. In this context, commonly faced tasks are instance-level retrieval and class retrieval [16]. In the former, given a query image, the goal is to retrieve images that contain the same object regardless of image distortions such as different illumination, rotation or occlusion. Instead, in the latter the purpose is to retrieve all the available images that belong to the same class.
Before the advent of the Convolutional Neural Networks (CNN), the scale-invariant feature transform [12] (SIFT) based methods were among the most frequently used in order to extract global descriptors from the given images. A breakthrough occurred in 2012 when the AlexNet [9] was created and won the ImageNet Large Scale Visual Recognition Competition (ILSRVC) improving upon the state-of-the-art by a noticeable margin. Since then, CNN-based methods for image retrieval received considerably more attention from the scientific community [1], [15].
Under the hood, these methods rely on the ability of deep models to extract the so called deep features from given input images. From a theoretical perspective, the inner layers of a CNN realize an abstraction of the input that describes specific concepts contained inside the data. Moreover, due to the typical structure of deep models architecture, inner layers combine the information available from previous layers thus achieving a higher level of abstraction that summarizes the overall content of the input data. Based on this observation, deep features are usually adopted as global descriptors for input images. Thus, the deeper the layer from which we extract deep features is, the more descriptive of the input they are. It is common practice to extract them from the penultimate layer of a CNN.
As previously said, deep features are high dimensional vectors defined in a space on which it is defined an inner product and thus a metric. This property is fundamental since it allows to evaluate similarities among descriptors, extracted from different images, that can be used as indicators of the similarities of the content of the original data. An example of this principle is sketched in Figure 1.

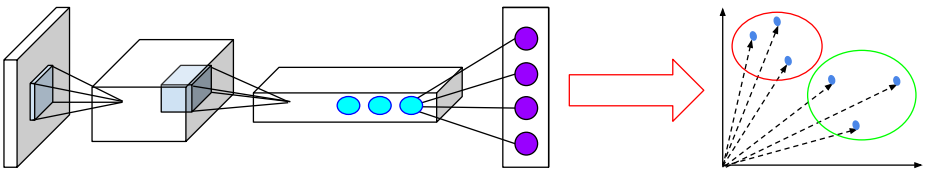

Figure 1. Schematic example of a CNN-based features extraction method for images that belong to two different classes. Vectors from same images tends to cluster in the arrival space.

This concept is typically applied in the context of surveillance systems.
Indeed, in a scenario where an input face image is acquired by a camera, its global descriptor is extracted and used in order to perform a similarity search on a database (db) containing a gallery of features vectors that belong to known identities. For example, the research in the db can be accomplished by evaluating the cosine similarity, or the Euclidean distance, among the probe and gallery vectors.
The similarity search becomes even more challenging when gallery and probe come from different resolution domains.
This is the background from which we started the study we present in this paper. Our final goal is to conceive a pipeline for face recognition based on neural networks. In order to extract deep features we used a pre-trained ResNet-50 [7] architecture, with Squeeze-and-Excitation blocks [8]. The extracted descriptors are then used to perform a similarity measurements.
The performance of the deep model used for features extraction has been evaluated on the 1:1 verification protocol on the IARPA Janus Benchmark-B (IJB-B) dataset [17].

1.1 Paper Organization

The remaining part of the paper is organized as follows. In Section 2 we briefly review some related works. In Section 3 we describe in detail the pipeline that we implemented. In Section 4 we present the experimental results. In Section 5 we conclude the paper with a summary of the main results and future perspectives.

2 Related Works

Deep learning techniques are currently experiencing a huge expansion in their field of application mainly as a result of the extremely high computational power reached by the modern GPUs. Moreover, the existence of big datasets [11], [18], [3], [2] and [6] has made it possible to train neural networks and to let them nearly reach human levels of performance when tested against tasks such as image classification [9], object detection [4] and face recognition [14].
Due to its wide range of applications, the task of Face Recognition (FR) is among the hottest topics in the computer vision community. In particular, FR plays a key role in the context of smart surveillance systems [10], [13]. In such systems, the case is usually that a low resolution face image, taken by a surveillance camera, has to be matched against a database containing deep features extracted from high resolution images.
To this end, several techniques have been developed in order to train deep models to deal with low resolution images. Some examples are Super Resolution and Common Space Projection techniques.
Super Resolution is a technique based on the ability of a neural network to synthesize a high resolution image starting from a low resolution one. The recognition task is later fulfilled in the high resolution domain [5].

One of the weaknesses of the super resolution technique is that the identity information can be lost. In [20] they developed a neural network that, together with the super resolution task, tries to recover the initial low resolution image identity in the high resolution one.
Instead, Common Space Projection techniques concern the ability of a neural networks to minimize the distance among deep features, in a common space, extracted from a low resolution image and its high resolution counterpart.
For example, in [19] they train a two-branch CNN to learn a mapping from high/low resolution domain to a common space. Specifically, given a low and high resolution image, the model extracts features vectors of size 2048 and their distance is evaluated and used as loss in order to lead the training towards the desired direction.

3 Pipeline

In this section we briefly describe the main modules of the pipeline we developed. A schematic view is shown in Figure 2.

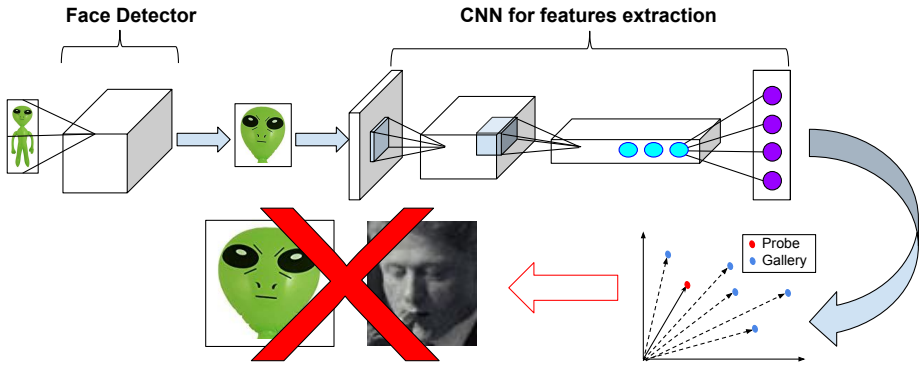

Figure 2. Sketch of the implemented pipeline.

There are three main components at the heart of the system: a face detector, a features extractor and a classifier. We will be focusing on the first two.
The face detection task is accomplished by means of a multi-stage architecture that, given an input image, delivers the coordinates of the bounding boxes that are centred around each face visible in the image. Specifically, we used the Multi-task Cascaded Convolutional Neural Networks (MTCNN) [21]. This step is performed once for each input frame. After all the faces have been identified from the picture, they are cropped, preprocessed and then used as input for the features extractor. The preprocess step includes the rescaling of the image with the shortest side resized at 256 pixels, the cropping of the squared central

region with side of 224 pixels and the normalization of the image
The features extractor module is made of a ResNet-50 [7] architecture, equipped with Squeeze-and-Excitation [8] blocks, that has been pretrained on the VGGFace2 dataset [3].
Features vectors are extracted before the classification layer, they have dimensionality equals to 2048 and they have been L_2-normalized before to evaluate any metric on them.

4 Experimental Tests

In order to test the performance of the features extractor we used the IJB-B dataset [17]. In particular, we tested the model against the 1:1 verification protocol aiming to estimate its ability to extract discriminative features. Due to the low resolution requirements for a surveillance system, we conducted the performance evaluation using the dataset with different resolutions. In Figure 3, we show an example of the various resolution versions of the test images. The first column contains the full resolution test images while from the second to the last column down sampled images in the [8, 256] pixels range are shown.

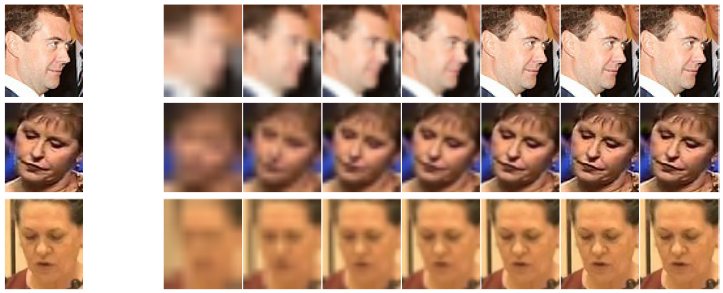

Figure 3. Example of images down sampled at various resolutions. First column: images at full resolutions. Second - Last columns: down sampled images in [8x, 256x] pixels.

The resulting Receiver Operating Characteristic (ROC) for the 1:1 verification task is shown in Figure 4.

In order to evaluate the similarity among the different features vectors we measured the cosine similarity among them. As it is clear from Figure 4 the performance of the features extractor are degraded for lower resolution, especially below 32 pixels. Moreover in Table 1 we reported the True Acceptance Rate (TAR) at a reference value for the False Acceptance Rate (FAR) of $1.e^{-3}$.

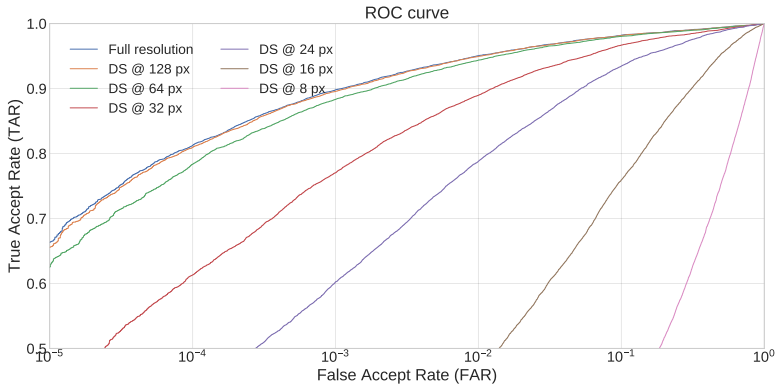

Figure 4. Feature extractor Receiver Operating Characteristic (ROC) for different values of the image resolution (DS stands for "Down Sampled").

Table 1. True Acceptance Rate (TAR @ FAR = $1.e^{-3}$) for different values of the images resolution.

Architecture	**TAR (@ FAR = 1.e−3)**						
	8	**16**	**24**	**32**	**64**	**128**	**256**
Se-ResNet-50	4.8	24.4	60.2	77.0	88.3	89.5	89.8

Up until now, we have considered the case in which the images had the same resolution. In a real scenario it usually happens that probe and gallery have different resolutions. For example, in the case of surveillance systems it is common that the gallery database is populated with high resolution images descriptors while the probe has a lower resolution.
In Table 2 we reported the value for the TAR @ FAR equals to $1.e^{-3}$ for the cross resolution 1:1 verification task.
It is clear from Table 2 that the cross resolution face recognition task is very challenging for a deep neural network model, especially when the images have very low resolutions.

5 Conclusion and Future Experiments

Surveillance systems require high performance from CNN-based systems on the face recognition task. Precisely, the deep models have to be robust with respect to variations of the input image resolution since it is usually the case that low resolution images, from surveillance cameras, have to be matched against a database containing deep features extracted from high resolution images. In fact, the descriptor extracted for each human face have to be robust with respect to resolution variations otherwise any kind of similarity search across the identities

Table 2. True Acceptance Rate (TAR @ FAR = $1.e^{-3}$) for 1:1 verification protocol considering images with mixed resolutions.

		Image Resolution (pixel)						
		8	**16**	**24**	**32**	**64**	**128**	**256**
Image Resolution (pixel)	**8**	**4.8**						
	16	0.2	**24.4**					
	24	0.2	18.6	**60.2**				
	32	0.2	9.0	65.3	**77.0**			
	64	0.2	2.9	60.4	80.5	**88.3**		
	128	0.2	2.4	57.9	80.1	88.9	**89.5**	
	256	0.2	2.3	57.5	80.1	90.0	89.7	**89.8**

database will fail. We have seen that, even though there are models that perform well on the FR task, their performances suddenly drop when we used low resolution images.

Although we have shown the feasibility of a pipeline for FR based on deep models, we need to improve upon its performance especially in the case of mixed resolutions. Thus, we are planning a new training campaign focused on the low resolution domain below 32 pixels. Finally, we will pay particular attention to the case of FR in which probe and gallery have different resolutions. What we expect from such a campaign is that, even if we might obtain a small drop in the performance at high resolution, the improvement at low and mixed resolutions should outweigh that drop.

References

1. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: European conference on computer vision. pp. 584–599. Springer (2014)
2. Bansal, A., Nanduri, A., Castillo, C.D., Ranjan, R., Chellappa, R.: Umdfaces: An annotated face dataset for training deep networks. In: 2017 IEEE International Joint Conference on Biometrics (IJCB). pp. 464–473. IEEE (2017)
3. 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)
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence **38**(1), 142–158 (2016)
5. Grm, K., Pernuš, M., Cluzel, L., Scheirer, W., Dobrišek, S., Štruc, V.: Face hallucination revisited: An exploratory study on dataset bias. arXiv preprint arXiv:1812.09010 (2018)
6. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision. pp. 87–102. Springer (2016)
7. 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)

8. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arxiv (2017)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
10. Lavi, B., Serj, M.F., Ullah, I.: Survey on deep learning techniques for person re-identification task. arXiv preprint arXiv:1807.05284 (2018)
11. Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G.: Labeled faces in the wild: A survey. In: Advances in face detection and facial image analysis, pp. 189–248. Springer (2016)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
13. Nikouei, S.Y., Chen, Y., Song, S., Xu, R., Choi, B.Y., Faughnan, T.: Smart surveillance as an edge network service: From harr-cascade, svm to a lightweight cnn. In: 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC). pp. 256–265. IEEE (2018)
14. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: bmvc. vol. 1, p. 6 (2015)
15. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of cnn activations. arXiv preprint arXiv:1511.05879 (2015)
16. Torresani, L., Szummer, M., Fitzgibbon, A.: Efficient object category recognition using classemes. In: European conference on computer vision. pp. 776–789. Springer (2010)
17. 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)
18. Wolf, L., Hassner, T., Maoz, I.: Face recognition in unconstrained videos with matched background similarity. IEEE (2011)
19. 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)
20. 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)
21. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)