ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8. Available from http://www.i6doc.com/en/.

Boosting face recognition via neural Super-Resolution

Guillaume Berger, Clément Peyrard and Moez Baccouche

Orange Labs - 4 rue du Clos Courtel, 35510 Cesson-Sévigné - France

Abstract. We propose a two-step neural approach for face Super-Resolution (SR) to improve face recognition performance. It consists in first performing generic SR on the entire image, based on Convolutional Neural Networks, followed by a specific local SR step for each facial component, using neural autoencoders. Obtained results on the LFW dataset for a $\times 4$ upscaling factor demonstrate that the method improves both image reconstruction ($+2.80$ dB in PSNR) and recognition performance ($+3.94$ points in mean accuracy), compared with $\times 4$ bicubic interpolation.

1 Introduction

Super-Resolution (SR) aims to produce high-resolution (HR) images from one or several low-resolution (LR) ones. We can distinguish specific SR methods which expect a particular type of input images from generic ones which do not make any assumption on the input.

Face SR, also known as Face Hallucination, refers to specific methods which perform SR on facial images. One of the main applications of face SR methods address recognition systems. Indeed, performance of face recognition algorithms is very dependent on image resolution, and drastically decreases with LR images. The face SR research topic is roughly composed of two main classes of methods. The first class consists in designing low and high resolution spaces where face SR can be performed more efficiently. Instead of working in the pixel domain, these approaches first project images into the designed spaces that incorporate a strong prior on face images, being most of the time aligned. These methods are generally used as a pre-processing step leading to a medium-resolution image which is then improved by a generic step [1, 2]. The other class of methods incorporates the face prior differently, in that they do not have a global face approach, but rather make use of facial component detection to focus on these specific parts [3, 4]. Matching this second class, the goal of the proposed method is to produce $\times 4$ SR images yielding the closest recognition scores to the ones obtained by HR.

The rest of this article is organized as follows: our two-step approach is presented in section 2. Then, the obtained results are described in section 3. Finally, we give a conclusion and perspectives in section 4.

2 A two-step neural approach for face Super-Resolution

This section describes the proposed approach (see Fig. 1). In the first step, a high-resolution image is generated using a generic SR approach. During the second step, localized SR is performed on facial components.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

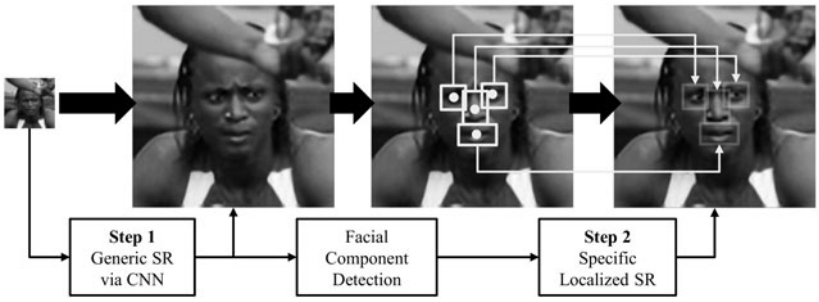

Fig. 1: The proposed two-step neural approach for face SR.

2.1 Generic SR using Convolutional Neural Networks

The first step of the proposed method aims to resample the initial LR image on a denser grid. Instead of using a simple interpolation scheme, we propose to take profit of the recent advances in Single Image SR (SISR). Learning based methods have shown good performance in accomplishing this task (such as neural models described in [5, 6]). These approaches take advantage of large sets of training examples, consisting most of the time in pairs of low and high resolution patches. During the learning process, they incorporate specific priors on both the LR-HR relationship and the nature of the images. Generally, natural images are used to train SR learning-based systems in order to fit general purpose application. However, if the systems are trained on specific image data, it can specialize on this kind of images and therefore boost the performance of further processing tasks (*e.g.* text recognition in [7]).

We train a generic Convolutional Neural Network (CNN) to perform the described SISR task, using a configuration similar to [6] adapted to the desired $\times 4$ scale factor. The network consists in several convolutional layers followed by neuron ones, and an output layer performing upsampling, as illustrated in Fig. 2. Non-linear activation functions (*tanh*) are applied to the output of each layer, except the last one which is linear to avoid saturation during the learning process.

The CNN weights are learned with standard backpropagation and mean squared error loss function, taking as input 2D LR patches and targeting pixel-wise difference between HR and bicubic patches (see Fig. 2). The latter corresponds to lost visual information and aim to compensate for artefacts such as blur or jagged edges. The variety of these artefacts, especially for high upscaling factors, makes the problem difficult and highly non-linear.

The 2D patches are extracted from face images taken in the wild, which are very close to natural images as they contain faces in various positions and a surrounding environment. We blindly sampled the low and high resolution pairs of patches without any knowledge on the location of the face. Therefore, the CNN is generic and mainly learns to remove natural interpolation artefacts, via a global optimisation over all example pairs.

However, face recognition involves *a priori* knowledge on the face structure and components (*e.g.* eyes, nose and mouth). As the present CNN is blind to these details,

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

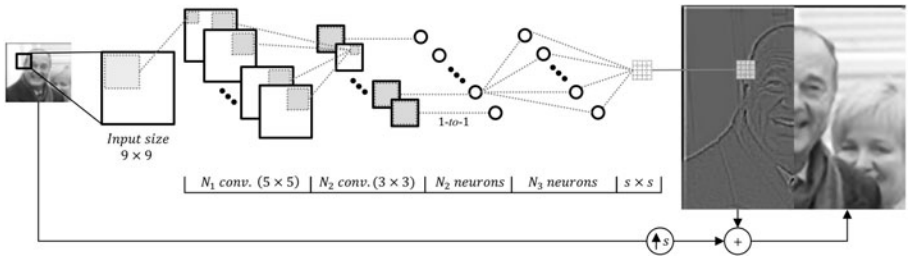

Fig. 2: CNN Architecture for the Generic SR step. We set $N_1 = 20$, $N_2 = 230$, $N_3 = 64$ after testing different configurations, and $s = 4$ for the experiments. Note that the input image is still sampled on the LR grid, while the output map is sampled on the HR grid using $s \times s$ linear output neurons, yielding a s^2 times larger image.

both by the variety of samples in the training set and the limited size of the input retina, a second step is proposed to specialize on the face components.

2.2 Localized SR using autoencoders

Whereas the first step is dealing with the entire image, the second one is focusing on crucial locations in terms of facial recognition. Indeed, eyes, mouth and nose regions are essential components for recognition. They cannot be blindly sampled in face images, especially taken in the wild, as they appear in specific locations. Therefore, we propose to exploit a detection algorithm (such as [8]) that would extract facial components from the output given by the first step and learn for each of them a specific model.

Once each of these components are extracted from the whole dataset, we may build a more compact representation via dictionary learning. These methods have been used in many ways to perform SR. They are generally patch-based and produce outputs as a weighted combination of atoms. In [2], coupled over-complete dictionaries are learned such that the HR and LR version of a same patch can be both represented by the same sparse code. In [9], several dictionaries are learned to perform both SR and expression normalization. While based on neural autoencoders, as can be seen on Fig. 3, our second step is inspired by those techniques.

For each facial component, a convolutional encoder projects the input patch into a N_D-dimensional hidden subspace, and the output patch is reconstructed from the obtained code through a fully connected one-layer decoder with linear activation. The output patch can then be written as a weighted linear combination of N_D atoms c_k: $o = \sum_{k=1}^{N_D} w_k \cdot c_k$ where o is the reconstructed output facial component, w_k are the components of the code in the hidden subspace, and c_k are the weights of the decoder layer associated to the k^{th} code component. Each atom is directly associated to one direction of the hidden subspace. In order to learn a meaningful representation, we add a sigmoid activation on the encoder output to make the code positive, and a non-negativity constraint on the decoder weights. As can be seen on Fig. 3, this constraint makes

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

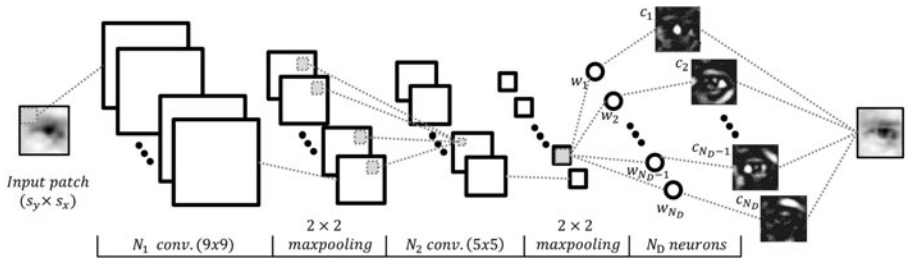

Fig. 3: Localized models for facial components. Left and right eyes, noses and mouths are extracted, and processed by distinct autoencoders. For eyes and nose patches: $s_x \times s_y = 36 \times 36$. For mouth patches: $s_x \times s_y = 48 \times 24$. Results of section 3 are given with $N_1 = 8$, $N_2 = 64$, $N_D = 128$.

atoms c_k become part-based and less noisy, just as in the case of non-negative matrix factorization [10].

In sum, our method first applies a generic SR model to the whole image. Then, the first step output is modified through localized models which project extracted facial components into hidden subspaces and reconstruct from these projections HR facial components as additive combinations of learned atoms.

3 Experimental results

3.1 Data and set-up

Experiments were carried out on the Labeled Faces in the Wild (LFW) dataset [11], widely used to evaluate facial recognition algorithms. Faces are present under different expressions, poses, expositions, illuminations and are sometimes partially occulted. We generate the LR images by blurring the original ones with a gaussian kernel of standard deviation $\sigma = 1.6$ and linearly downsampling them by a factor of 4. From the generated LR and HR image pairs, we randomly extract patches and train the first step CNN. Then, for the second step, we automatically extract facial components from the generic SR and HR images with an algorithm based on facial landmark detection [8].

To benchmark the proposed approach, we use a face recognition system inspired by the simile classifier described in [12]. Note that the performance of this recognition engine is not competitive with current state of the art approaches. However, our goal is to evaluate the benefit of the proposed SR algorithm in terms of face recognition, which can be outlined with any reasonably performing method.

3.2 Evaluation

Although improving recognition performance is the main goal of this work, we first propose to evaluate our method from a reconstruction and visual point of view, as it is usually done to evaluate SR methods.

Fig. 4 shows the LR, bicubic and HR version of one LFW test image as well as the outputs after each step of our method. Visually, SR images look sharper than the bicu-

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

Fig. 4: Obtained results. a.: Left to right: LR, bicubic (PSNR = 28.84 dB), generic step only (PSNR = 32.28 dB), our method (PSNR = 31.64 dB), HR. b. to e.: Compelling test examples of left eye, right eye, nose and mouth for generic step only, our method, and HR.

bic ones. This visual impression is reflected in the PSNR measurements (see Fig. 4). Indeed, mean PSNRs obtained on the whole LFW testset indicate that SR ouputs are closer to HR images than the bicubic ones. Nevertheless, it should be noticed that the highest PSNR is obtained on the first generic step outputs: facial component models make the PSNR decrease. This diminution can be explained by the fact that positivity constraints added on the code and decoder weights make the reconstruction goal harder to fulfil: second step outputs have to be produced by adding a limited number of positive atoms. As a consequence, facial components given by localized models tend to differ from HR targets and make the PSNR drop. However, even if slightly different from the target, facial components given by localized models contain less visual artefacts which were still present after the first step. This could be mainly explained by the fact that second step outputs are reconstructed combining clean part-based atoms.

	Accuracy improvement
Bicubic	+4.21
Generic SR	+6.91
Our method	+8.15
HR	+11.85

Table 1: Obtained results in terms of recognition improvement, w.r.t. LR mean accuracy (74.70%).

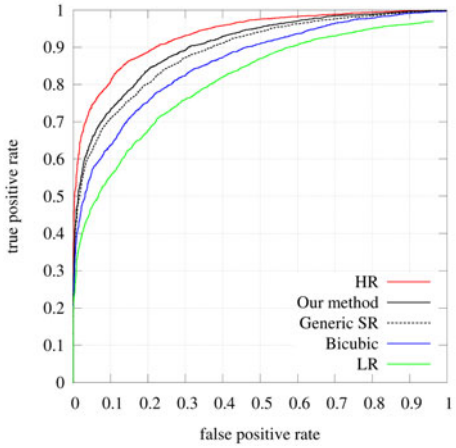

Fig. 5: ROC curves showing face recognition performance on the LFW dataset under different resolutions.

ESANN 2016 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 27-29 April 2016, i6doc.com publ., ISBN 978-287587027-8.
Available from http://www.i6doc.com/en/.

At the end, as long as processing the facial components of the first step by our localized models improves recognition performance, the PSNR decrease is a negligible counterpart. This is confirmed by the results reported in Tab. 1: if the generic step has already permitted to gain $+2.70$ points over the bicubic mean accuracy, localized models improve this score by additional $+1.24$ points. ROC curves of Fig. 5 show that the best performance is obtained when images are treated by our two-step method.

4 Conclusion and perspectives

In this article, we proposed a two-step super-resolution approach using two different neural architectures, that allows to produce a high-resolution face image from a low-resolution one without constraints on the pose, the illumination or the face alignment. Results on LFW dataset showed that recognition system performance benefits from both a higher resolution and reduced artefacts on facial components.

Still, the method is limited in some aspects. We could take profit from larger datasets and learn more complex neural networks. As far as the second step is concerned, a more complex encoder and the learning of an overcomplete dictionary could have led to better reconstruction performance. Finally, the proposed approach is highly dependent on the facial components detector. An interesting challenge would be to absorb the proposed steps into an end-to-end framework.

References

[1] C. Liu, H-Y. Shum, and W.T. Freeman. Face hallucination: Theory and practice. *International Journal of Computer Vision*, 2007.

[2] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *Transactions on Image Processing*, 2010.

[3] C.-Y. Yang, S. Liu, and M.-H. Yang. Structured face hallucination. In *Computer Vision and Pattern Recognition*, 2013.

[4] A. Choudhury and A. Segall. Channeling mr. potato head - face super-resolution using semantic components. In *Southwest Symposium on Image Analysis and Interpretation*, 2014.

[5] C. Dong, C.C. Loy, K. He, and X. Tang. Learning a deep convolutional network for image super-resolution. *European Conference on Computer Vision*, 2014.

[6] C. Peyrard, F. Mamalet, and C. Garcia. A comparison between multi-layer perceptrons and convolutional neural networks for text image super-resolution. In *International Conference on Computer VIsion Theory and Applications*, 2015.

[7] C. Peyrard, M. Baccouche, F. Mamalet, and C. Garcia. ICDAR2015 competition on text image super-resolution. In *International Conference on Document Analysis and Recognition*, 2015.

[8] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. *Transactions on Pattern Analysis and Machine Intelligence*, 2013.

[9] Y. Liang, X. Xie, and J.-H. Lai. Face hallucination based on morphological component analysis. *Signal Processing*, 2013.

[10] D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 1999.

[11] 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. Technical Report 07-49, University of Massachusetts, 2007.

[12] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar. Attribute and simile classifiers for face verification. In *International Conference on Computer Vision*, 2009.