A Hybrid Multiframe Superresolution Framework Using R-ELM Neural Networks and Maximum a Posteriori Estimation

Thais Pedruzzi do Nascimento[1] and Evandro Ottoni Teatini Salles[2]

Laboratório de Computação e Sistemas Neurais
Universidade Federal do Espírito Santo
Vitória, Espírito Santo, Brasil
thais.p.nascimento@aluno.ufes.br[1],evandro.salles@.ufes.br[2]

Abstract. A hybrid Superresolution framework is proposed, using an approach based on neural networks to improve the image estimated by superresolution algorithms based on maximum a posteriori (MAP). We refer to the residual image which represents, we believe, the missing information. We use a set of networks to learn the relationship between the residual image - available during the training step - and the image estimated by the MAP approach. We trained our networks under the Extreme Learning Machine paradigm, where a closed-form solution is applied when training a single layer feed-forward neural network. We tested our proposal over five image datasets: LIVE, Set5, Set14, Urban100 and B100. And we used another set, with 91 images, for the training step. Our results have shown improvements when comparing them with two multiframe superresolution methods, the Bilateral Total Variation (BTV) and Weighted Bilateral Total Variation (RWBTV) approaches.

Keywords: Multiframe Superresolution, Regularized Extreme Learning Machine, Neural Networks, Maximum a Posteriori, Prior Regularization, Bilateral Total Variation

1 Introduction

Superresolution (SR) methods aim to increase the spatial resolution of one image (singleframe SR) or a set of them (multiframe SR). This work is focused on multiframe SR. Among the many different approaches presented by state-of-the-art of superresoultion we emphasize the reconstruction-based ones and the example-based ones. Particularly, regarding the reconstruction methods, we stress the use of maximum a posteriori estimation (MAP) to solve the optimization problem that arises from the modeling. In turn, for the learning approaches, we refer to the use of neural networks.

Multiframe methods take advantage of the different information available from the sub-pixel displacement between the several LR input frames. Due to the modeling of the image formation process and the presence of random noise, such methods place the SR problem into the Bayesian paradigm. One of the

first relevant works to explore this framework was proposed by Elad and Feuer [6], where the solution is estimated via Maximum Likelihood (ML), MAP and Projection Onto Convex Sets (POCS), besides using prior information, calculated via ℓ_2 norm. On the other hand, the use of ℓ_1 norm in the prior term - also called "regularization term" - via Bilateral Total Variation (BTV) was originally proposed by Farsiu et al. [8] and later used by other prominent works such as the one from Villena et al [16]. The need of a regularization term is due to the ill-posedness of the inverse problem. The choice of ℓ_1 norm, in particular, regards the attempt to recover edge elements.

On the other hand, the main idea of example-based approaches is to map the relationship between the LR and original HR images, using a training dataset and machine learning techniques to learn this relationship. Later, such learned information is used to estimate the output image. Freeman et al. [9] were one of the firsts to apply a learning process via external database to the context of SR. In such work, patches are extracted from the LR image and for each patch, a nearest neighbor is searched among the patches from the external database. Then, the high frequency (HF) elements of the chosen patch are extracted and added to the estimated HR image. Deep learning approaches, such as [5], are among state-of-the-art methods, however, they demand a higher computational cost. Multiple-mapping algorithms [14] also present good results and are based on describing the HR-LR images relation via multiple regressors.

From the premise that the MAP-based methods do not reconstruct all relevant information, our proposal is to apply an example-based method to an image estimated by a reconstruction-based algorithm, in order to improve such result. The definition of the prior is made in a way to enhance edge content. However, we believe that other type of information may also be important. To test our hypothesis we chose two multiframe methods, one based on Bilateral Total Variation (BTV) regularization [8], which is a well-established method; and another based on Weighted Bilateral Total Variation (RWBTV) [12], which presents adaptive regularization.

The example-based singleframe method [3] aims to relate the filtered image (resulting from bicubic interpolation) and the relating high frequency content extracted from the original HR image, available during the training step. The neural networks are trained under the Regularized Extreme Learning Machine (R-ELM) [4] approach, which is a closed-form solution for the training problem when using a single layer feedforward network.

This paper is organized as follows: Section 2 briefly explains the MAP framework applied to SR algorithms, Section 3 explores the Neural Networks framework for superresolution, in Section 4 we explain our proposed framework, in Section 5 we show our experiments' methodology and results and finally, Section 6 explores our conclusions.

2 Maximum-a-Posteriori Framework

Superresolution can be defined as an inverse problem, where the LR input image (or the set of them) is seen as a blurred, warped, noisy and down-sampled version of the HR image [6]. Such model is formulated as

$$\mathbf{y}_k = \mathbf{W}_k\mathbf{D}\mathbf{B}\mathbf{x} + \eta = \mathbf{H}_k\mathbf{x} + \eta, \tag{1}$$

where $\mathbf{x}$ is the original HR image, $\mathbf{y}_k$ is one of the k LR images, $\mathbf{W}_k$ is the warping operator, $\mathbf{B}$ is the blurring operator, $\mathbf{D}$ represents the down-sampling operator and η is the additive noise.

Under the Bayesian perspective the sequence of LR images and the HR image are modeled as random variables [6]. Assuming η as additive White Gaussian Noise (AWGN), the image estimation problem can be written as

$$\hat{\mathbf{x}} = \arg\max_x \left[\sum_{k=1}^{K} \|\mathbf{H}_k\mathbf{x} - \mathbf{y}_k\| + p(\mathbf{x}) \right], \tag{2}$$

where $p(\mathbf{x})$ is a prior term and it works, from the optimization point of view, as a regularization parameter.

The prior is chosen in order to accentuate a particular characteristic such as edges elements or to perform noise removal, for instance. The Bilateral Total Variation (BTV) prior, proposed by [8], is known for removing noise while preserving edge components and was originally formulated as

$$p(\mathbf{x}) = \Gamma_{\mathrm{BTV}}(\mathbf{x}) = \sum_{m=-P}^{P} \sum_{n=-P}^{P} \alpha^{|m|+|n|} \|\mathbf{x} - \mathbf{S}_m^v\mathbf{S}_n^h\mathbf{x}\|_1, \tag{3}$$

with $P \geq 1$ and $\alpha \in]0, 1]$. $\mathbf{S}_m^v$ and $\mathbf{S}_n^h$ are the vertical and horizontal derivatives, respectively. P defines the window $(2P+1) \times (2P+1)$ where the derivatives are exploited and α is the scale factor.

However, the properties of the BTV prior are limited in terms of edge preservation. In this sense Kohler et al. [12], proposed a generalized version of such prior, formulated as

$$\Gamma_{\mathrm{RWBTV}} = \sum_{m=-P}^{P} \sum_{n=-P}^{P} \|\mathbf{A}(\mathbf{x} - \mathbf{S}_m^v\mathbf{S}_n^h\mathbf{x})\|_1, \tag{4}$$

where $A = \mathrm{diag}(\alpha_1, ..., \alpha_{N'})$ with $N' = (2P+1)^2N$, where $N = N_1 \times N_2$ is the dimension of the HR image. Note that, in Equation (3), only one number α weights the prior. On the other hand, in Equation (4), a weighting matrix A is applied. That way, it is possible to reconstruct the HR image locally. For example, a homogeneous region and a region with edges may be reconstructed differently.

3 Neural Networks Framework

Example-based approaches are usually applied to singleframe Superresolution methods. In such cases, an acquisition model is formulated as Equation (1), with only one LR image ($\mathbf{x}$) and no warping operator ($\mathbf{W_k}$). Also, the additive noise η is usually not assumed.

The goal of machine learning methods is to learn the relationship between $\mathbf{y}$ and $\mathbf{x}$ through a external dataset of examples. Each image of such dataset is deformed according to the acquisition model, forming then, the set of pairs of HR original images and corresponding LR images. However, instead of learning the direct relationship between the given LR image and the estimated HR image, one first estimation (e.g. bicubic interpolation) may be performed resulting in $\mathbf{Y}_0$. Such estimation is subtracted from the original HR image $\mathbf{X}$, resulting in $\mathbf{Z}$, which is believed to be constituted by high frequency components of the original HR image. Here, we refer to $\mathbf{Z}$ as *residual image*. The external database is then used to map $\mathbf{Y}_0$ to $\mathbf{Z}$.

Regarding the testing stage, the LR image is interpolated, resulting in $\hat{\mathbf{Y}}_0$, which feeds the previously trained set of artificial neural networks (ANNs) that estimates $\hat{\mathbf{Z}}$. Finally, such result is added to the first interpolated image, to obtain the final estimated image $\hat{\mathbf{X}}$.

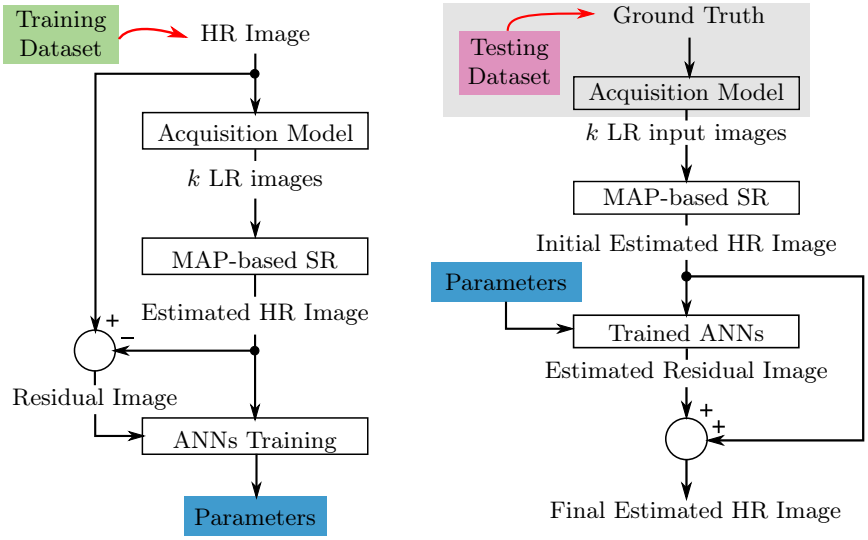
Cosmo et al. [3] consider patches of images instead of the whole data. The idea is to form K clusters, according to the geometric information contained in each patch. For each cluster one network is trained to map the relationship between the relating patch pairs, in a way that each network is more specialized on a type of information, given the cluster. Following [14], these clusters are formed using a set of three geometric information based on local gradient statistics, evaluated by eigenanalysis [18]: strength, coherence and orientation. The clusters are distributed in a 3D histogram, where each bin is defined by its respective coherence, strength and orientation. Therefore, the number of bins also defines the number of networks to be trained. In the testing stage, each patch is assigned to a cluster and then, set as input for the network associated with such cluster. The output patches from all networks are assembled forming the complete estimated image $\hat{\mathbf{Z}}$. For more details about the clustering scheme, the reader should refer to [18].

Once the patches are clustered, each network is trained using Regularized Extreme Learning Machine (R-ELM) [4], which results in a more generalized solution when compared with the original ELM [10]. Besides, it is also more robust to over-fitting than the latter. In such context the hidden nodes parameters are randomly chosen and the output weights are evaluated via Moore-Penrose Generalized Inverse, which is a closed-form solution based on the following equation,

$$\mathbf{B}_{opt} = \left(\frac{\mathbf{I}}{C} + \mathbf{L}^T\mathbf{L}\right)^{-1}\mathbf{L}^T\mathbf{T}, \tag{5}$$

where $\mathbf{L}$ is the matrix of output values of the hidden layer, C is the regularization parameter and $\mathbf{B}_{opt}$ is the weight matrix.

4 Hybrid Framework



(a) Training Stage (b) Testing Stage

Fig. 1: Hybrid Framework Proposed

Figure 1a shows a simplified scheme of the training stage. The original HR image - extracted from the external dataset - is deformed through the Acquisition Model, resulting on the set of k LR images. These images are firstly reconstructed by a multiframe MAP SR method. The MAP methods we used are based on BTV and RWBTV priors. The ANNs training includes the clustering and networks training. The clusterization scheme based on egeinalysis, as explained in Section 3. Finally, the networks are trained under the R-ELM approach. As for the testing stage, in Figure 1b, a different database is chosen. The Acquisition Model and MAP-based SR are the same used for training. The image estimated by the MAP-based SR is used as input for trained ANNs, obtaining the estimated residual image, which is added to the Initial Estimated imaged. Then, the final Estimated HR Image is compared with the Ground Truth, using PSNR and SSIM.

5 Experiments

We applied the algorithms on five dataset: LIVE [15], Set 14 [19], Set 5 [2], B100 [13] and Urban100 [11]. Four different setups were tested: the robust ℓ_1 norm minimization with BTV regularization proposed by [8] combined with the Image

Alignment Method proposed by [7], referred here as BTV; Our R-ELM trained networks applied to the image estimated by the BTV SR method; the Robust Multiframe Super-Resolution Employing Iteratively Re-Weigted Minimization, proposed by [12], referred here as RWBTV and Our R-ELM trained networks applied to the image estimated by RWBTV.

The acquisition model we are focused on is usually adopted by works based on reconstruction, such as [12], [16] and [8]. That is, we assume additive noise and a linear invariant point spread function (PSF). Thus, we compare our results with MAP-based works. It would not be fair to consider approaches that do not assume random noise, which is usually the case of singleframe example-based approaches [3], [1], [5]. Regarding color images, we consider the Y channel, from the YCbCr color space.

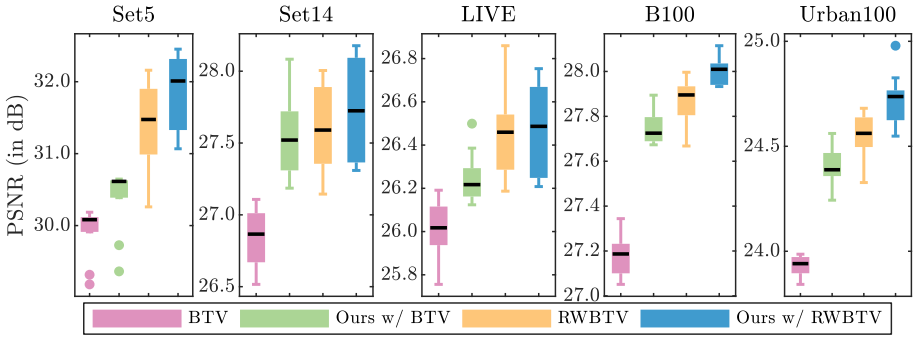
SSIM [17], PSNR and processing time were obtained. Each experiment was run ten times. All the experiments were implemented in MATLAB R2016b, using an Intel® Core™ i7-8700K processor and 32 GB of RAM memory. For each ground truth image we generated 8 LR images, simulating warping, down-sampling and blurring effects, as done in [12]. The LR images were displaced according to uniform distributed random translations in the range $[-3, +3]$ pixels and rotations in the range $[-1°, +1°]$. The blurring operator is based on a PSF approximated by an isotropic Gaussian kernel of size $6 \cdot \sigma_{\texttt{PSF}}$ (where $\sigma_{\texttt{PSF}} = 0.5$) and the down-sampling operator was modeled according to the magnification factor s. Moreover, each frame was disturbed by a white Guassian noise with standard deviation $\sigma = 0.025$. Regarding the reconstruction-based step, we set the parameters window size $P = 2$ and scale factor $\alpha = 0.6$. The regularization weight, in turn, was set to be automatically calculated over the iterations. The motion estimation was solved using Enhanced Correlation Coefficient (ECC) maximization [7], using an affine motion model.

The R-ELM networks were trained over the same dataset used by [5], which presents 91 images. From each image, 20.000 samples were randomly extracted and the patches were 5×5 sized. The 3D histogram was formed according to the following intervals: $[0, pi]$ with steps valued as $\pi/20$, for the orientation bins; $[0, 1/3, 2/3, 1]$ for the strength bins and $[0, 1/3, 2/3, 1]$ for the coherence bins. The ANNs were set with 1000 hidden neurons and sigmoid activation function. Finally, the regularization term was defined as $C = 2^8$.

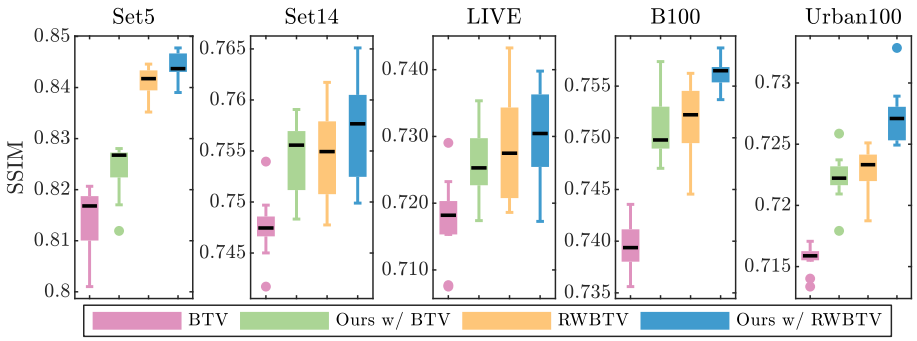
5.1 Results

Figures 2a and 2b show the quality results in boxplots. There is two charts for dataset (PSNR and SSIM), where each box refers to one method. Besides, each box represents the 10 realizations of the same experiment. In all graphs, the median of the blue box (that is, our method with RWBTV) is the highest one. In terms of PSNR, for Set5 our result is better in, at least, 25% of the cases, when comparing with pure RWBTV. For B100, the improvement increases to, at least, 75% of the cases and, for Urban100, 50% of them. For Set14, although our median is higher, the two boxes (blue and yellow) show almost the same results: ours over RWBTV from 27.3 dB to 28.2 dB and pure RWBTV from

27.1 dB to 28 dB. However, looking at Table 1, one can see that the RWBTV method is, on average, around 5 times slower than Ours with BTV. And for Set14, the green box (the one referring to our method with BTV) is within the range from 27.2 dB to 28 dB. That is, in this case, it is five times faster to apply our method with BTV than RWBVT and the quality is the same, according to the boxplot. Moreover, in Figure 3, the differences between the estimated images can be visually noted. The PSNR and SSIM values for Figure 3d are both lower than the ones related to Figure 3e. However, when observing the marked region, we can notice more loss on the edge components for the latter.



(a)

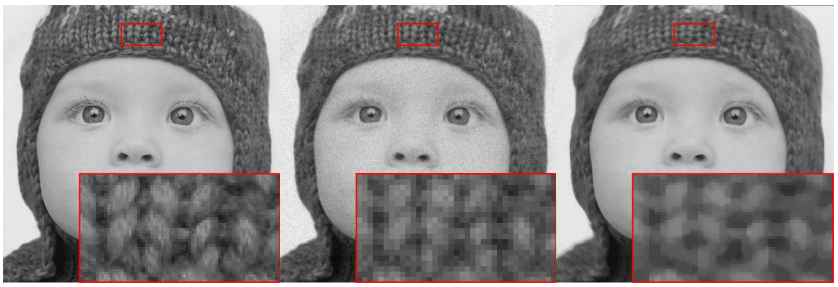


(b)

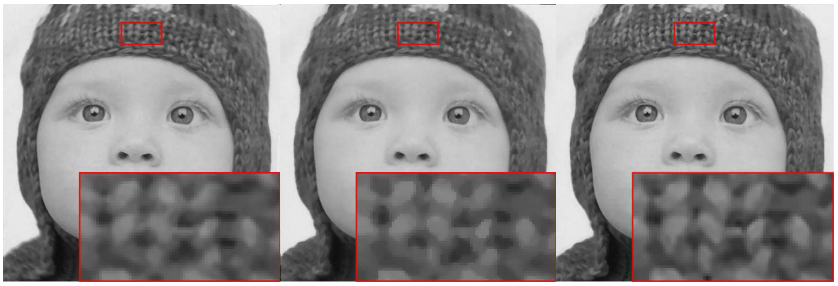
Fig. 2: Boxplots showing the results when running the experiments over the datasets Set14, LIVE, B100 and Urban100. In each plot, the boxes refer to the relating method. From the left to the right: pink box: BTV method; green box: Our method with BTV; yellow box: RWBTV method; and blue box: our method with RWBTV. The circles represent outliers and the black lines mark the median of each set. The results are presented in (a) PSNR and (b) SSIM.

	BTV	Ours w/ BTV	RWBTV	Ours w/ RWBTV
Set5	$\mathbf{3.97 \pm 0.08}$	6.92 ± 0.11	30.91 ± 0.76	32.82 ± 0.58
Set14	$\mathbf{8.73 \pm 0.17}$	13.05 ± 0.06	64.00 ± 1.27	67.19 ± 1.02
LIVE	$\mathbf{14.26 \pm 0.07}$	19.42 ± 0.06	99.16 ± 1.65	103.85 ± 1.25
B100	$\mathbf{5.74 \pm 0.04}$	9.08 ± 0.03	39.95 ± 0.17	43.28 ± 0.22
Urban100	$\mathbf{8.09 \pm 0.03}$	11.64 ± 0.08	53.23 ± 0.43	57.57 ± 0.26

Table 1: Average Processing Time.



(a) Original Image (b) LR Frame (c) BTV 28.9304 dB (0.8528)



(d) Ours w/ BTV 31.3250 dB (0.8615) (e) RWBTV 32.8469 dB (0.8900) (f) Ours w/ RWBTV 34.8138 dB (0.9103)

Fig. 3: Estimated images resultant from the SR methods (c) ℓ_1-BTV with IAT registration, (d) Our approach over ℓ-BTV, (e) RWBTVand (f) Our appoach over RWBTV applied to the ground truth image *baby*, from Set5, and $\times 2$ magnification factor.

6 Conclusion

In this paper we have proposed a hybrid superresolution technique, combining an approach based on MAP with one based on neural networks. Such networks were trained using R-ELM, an approach that is fast and robust to over-fitting. An external database and a clusterization based on eigenanalysis were used to feed the training process. Our motivation is due to the belief that the first approach do not estimate all the missing information from the LR frames and that such information could be learned and described by neural networks.

We ran each experiment 10 times applying the algorithms to 5 datasets and compared our proposal with one method based on BTV regularization and one based on Re-Weighted BTV (RWBTV). The results have shown the improvement accomplished by our proposal when applying it such MAP-based methods, both quantitatively (PSNR and SSIM) and qualitatively, via visual inspection.

References

1. An, L., Bhanu, B.: Image Super-Resolution By Extreme Learning Machine. Icip **1**(1), 2209–2212 (2012). DOI 10.1109/ICIP.2012.6467333
2. Bevilacqua, M., Roumy, A., Guillemot, C., Morel, M.L.A.: Low-Complexity Single-Image Super-Resolution based on Nonnegative Neighbor Embedding. Procedings of the British Machine Vision Conference 2012 (Ml), 135.1–135.10 (2012). DOI 10.5244/C.26.135. URL http://www.bmva.org/bmvc/2012/BMVC/paper135/index.html
3. Cosmo, D.L., Inaba, F.K., Salles, E.O.T.: Single Image Super-Resolution Using Multiple Extreme Learning Machine Regressors. 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI) pp. 397–404 (2017). DOI 10.1109/SIBGRAPI.2017.59. URL http://ieeexplore.ieee.org/document/8097339/
4. Deng, W., Zheng, Q., Chen, L.: Regularized Extreme Learning Machine. In: 2009 IEEE Symposium on Computational Intelligence and Data Mining, vol. 51, pp. 389–395. IEEE (2009). DOI 10.1109/CIDM.2009.4938676. URL http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=1443594http://ieeexplore.ieee.org/document/6319793/http://ieeexplore.ieee.org/document/4938676/
5. Dong, C., Loy, C.C., He, K., Tang, X.: Image Super-Resolution Using Deep Convolutional Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 295–307 (2016). DOI 10.1109/TPAMI.2015.2439281
6. Elad, M., Feuer, A.: Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. IEEE Transactions on Image Processing **6**(12), 1646–1658 (1997). DOI 10.1109/83.650118
7. Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(10), 1858–1865 (2008). DOI 10.1109/TPAMI.2008.113
8. Farsiu, S., Robinson, M.D., Elad, M., Milanfar, P.: Fast and robust multiframe super resolution. IEEE Transactions on Image Processing **13**(10), 1327–1344 (2004). DOI 10.1109/TIP.2004.834669
9. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-Based Super-Resolution. IEEE Computer Graphics and Applications **22**(March), 56–65 (2002)

10. Huang, G.b., Zhu, Q.y., Siew, C.k.: Extreme Learning Machine : A New Learning Scheme of Feedforward Neural Networks. IEEE International Joint Conference on Neural Networks **2**, 985–990 (2004). DOI 10.1109/IJCNN.2004.1380068. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1380068
11. Huang, J.B., Singh, A., Ahuja, N.: Single Image Super-Resolution From Transformed Self-Exemplars. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5197–5206 (2015). URL https://github.com/jbhuang0604/
12. Kohler, T., Huang, X., Schebesch, F., Aichert, A., Maier, A., Hornegger, J.: Robust Multiframe Super-Resolution Employing Iteratively Re-Weighted Minimization. IEEE Transactions on Computational Imaging **2**(1), 42–58 (2016). DOI 10.1109/TCI.2016.2516909. URL http://ieeexplore.ieee.org/document/7378486/
13. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to\nevaluating segmentation algorithms and measuring ecological statistics. Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001 **2**(July), 416–423 (2001). DOI 10.1109/ICCV.2001.937655
14. Romano, Y., Isidoro, J., Milanfar, P.: RAISR: Rapid and Accurate Image Super Resolution. IEEE Transactions on Computational Imaging **3**(1), 110–125 (2017). DOI 10.1109/TCI.2016.2629284
15. Sheikh, H.R., Bovik, A.C., Cormack, L., Wang, Z.: LIVE Image Quality Assessment Database Release 2 (2005). URL http://live.ece.utexas.edu/research/quality
16. Villena, S., Vega, M., Babacan, S.D., Molina, R., Katsaggelos, A.K.: Bayesian combination of sparse and non-sparse priors in image super resolution. Digital Signal Processing: A Review Journal **23**(2), 530–541 (2013). DOI 10.1016/j.dsp.2012.10.002. URL http://dx.doi.org/10.1016/j.dsp.2012.10.002
17. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). DOI 10.1109/TIP.2003.819861
18. XiaoGuang Feng, Milanfar, P.: Multiscale principal components analysis for image local orientation estimation. In: Conference Record of the Thirty-Sixth Asilomar Conference on Signals, Systems and Computers, 2002., vol. 1, pp. 478–482. IEEE (2002). DOI 10.1109/ACSSC.2002.1197228. URL http://ieeexplore.ieee.org/document/1197228/
19. Zeyde, R., Elad, M., Protter, M.: On Single Image Scale-Up using Sparse Representation. European Conference on Computer Vision (1), 1–20 (2010). DOI 10.1007/978-3-642-27413-8_47