Trainable Regularization for Multi-frame Superresolution

Teresa Klatzer[1], Daniel Soukup[2], Erich Kobler[1], Kerstin Hammernik[1] and Thomas Pock[1,2]

Institute of Computer Graphics and Vision, Graz University of Technology, Austria
Center for Vision, Automation, and Control, AIT, Austria

Abstract. In this paper, we present a novel method for multi-frame superresolution (SR). Our main goal is to improve the spatial resolution of a multi-line scan camera for an industrial inspection task. High resolution output images are reconstructed using our proposed SR algorithm for multi-channel data, which is based on the trainable reaction-diffusion model. As this is a supervised learning approach, we simulate ground truth data for a real imaging scenario. We show that learning a regularizer for the SR problem improves the reconstruction results compared to an iterative reconstruction algorithm using TV or TGV regularization. We test the learned regularizer, trained on simulated data, on images acquired with the real camera setup and achieve excellent results.

1 Introduction

In this paper, we investigate the problem of multi-frame superresolution (SR) on an exemplary industrial inspection task. To speed up image acquisition, we acquire multiple low resolution (LR) images using the lines of a multi-line scan camera with planar objects being moved under the sensor. To reduce redundancy and to improve the sampling pattern, the sensor is tilted. In such a setup, we can vary the resolution not only in transport direction by controlling the transport speed of the imaged object, but also in lateral direction by varying the tilting angle of the camera. This is visualized in Figure 1. The acquisition using different lines of the camera can be interpreted as a multi-camera setup, therefore we solve a multi-frame SR problem as a post-processing step. A similar idea has been used for reducing data transfer for a remote sensing application in satellite imaging [10] where several sub-pixel translated cameras were used to acquire images in half the desired resolution.

The multi-line camera setup is reflected in the forward model. Assuming registered images according to the forward model, our problem reduces to estimate a deblurred HR image from blurry but registered measurements. In our work, we view the SR problem as a variational image reconstruction problem which is one of the most popular approaches tackling SR according to the recent review article [15]. We propose a trainable regularizer based on the trainable reaction-diffusion model [5] and its extension to color images [8] applied to our

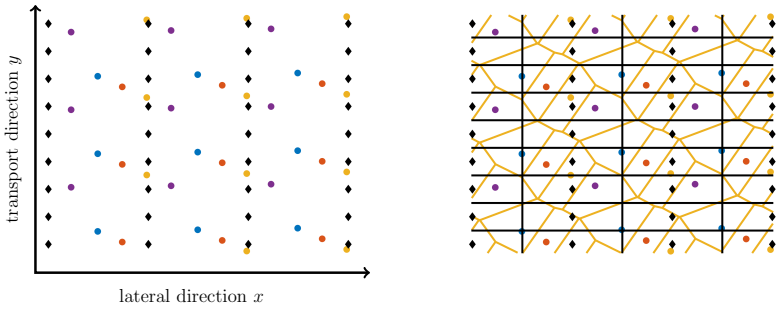


Fig. 1. Left: Black diamonds depict a regular upsampling pattern, the colored dots depict the sampling pattern with the suggested 4 line setup using a tilted multi-line scan camera, projected to HR space. With the proposed setup, the resolution is increased in x and y direction. Right: The corresponding Voronoi tesselation shows better coverage of the HR space with the suggested setup (yellow) vs. the regular (black) setup.

multi-frame SR problem. During inference, the trained regularizer does neither require parameter search nor the selection of stopping criteria and has constant run time for the reconstruction which depends on the amount of parameters and processed image data. In this sense, our approach could be seen as learning an optimal SR algorithm tailored for our task. With our approach, we show a successful application example where machine learning can improve image quality in a setup where ground truth is hard to obtain. If the specific camera setup is known the presented SR method is very effective and can recover fine details that are lost with common reconstruction techniques.

2 Multi-frame Superresolution

We introduce the forward model for the multi-frame SR setup

$$f_l = BW_l u_{gt} + \eta_l \tag{1}$$

which describes the degradation of a HR image $u_{gt} \in \mathbb{R}^{MNC}$ through the acquisition process plus some additive Gaussian noise η_l, MN defines the number of pixels and C the number of image channels. The result is a degraded LR image $f_l \in \mathbb{R}^{mnC}$, $l = 1 \dots L$, with L being the total number of observations (or readout lines). The degradation is modeled with the matrices $W_l \in \mathbb{R}^{mnC \times MNC}$ and $B \in \mathbb{R}^{mnC \times mnC}$. The matrix W_l encodes the warping from the HR space to the LR space, including downsampling, with possible shear and translation between the observed images, and interpolation to the pixel grid of the respective HR and LR coordinate system. The blur matrix B describes the point spread function (PSF) of the camera.

Based on the forward model (1) *regularization-based SR reconstruction approaches* aim to reconstruct a HR image from a set of L LR images. SR is an

ill-posed inverse problem, so we formulate the following minimization problem

$$\operatorname*{argmin}_u \sum_{c\in\mathcal{C}} \sum_{l=1}^{L} \psi(B_c W_l \mathcal{I}_c u - f_{l,c}) + \lambda \mathcal{R}(u) := \sum_{c\in\mathcal{C}} \sum_{l=1}^{L} \mathcal{D}(u, f_{l,c}) + \lambda \mathcal{R}(u) \quad (2)$$

where we estimate $u \in \mathbb{R}^{MNC}$ from observations $f_{l,c} \in \mathbb{R}^{mn}$ for each color channel $c \in \mathcal{C} := \{r, g, b\}$. The matrix $\mathcal{I}_c$ selects a single color channel from u. The left part is the data fidelity term $\mathcal{D}(\cdot)$, where $\psi(\cdot)$ is typically $\|\cdot\|_k^k$ with $k \in \{1, 2\}$. We assume that the blur matrix B_c is constant for all observations, but different for each color channel $c \in \mathcal{C}$. The right part defines the regularization term $\mathcal{R}(\cdot)$ which is added to make the reconstruction problem well-posed by imposing some prior knowledge about image structures. There have been many approaches tackling the SR problem based on model (2), also for the related task of video SR [15]. In the following we will describe two standard choices for the regularization term $\mathcal{R}(u)$.

Image Priors A very popular prior that has been used for regularization in SR [6] is the total variation (TV) prior. The discrete version of the TV image prior can be written as $\mathcal{R}(u) = \mathrm{TV}(u) = \|\nabla u\|_{2,1}$, with $\nabla \in \mathbb{R}^{2MNC\times MNC}$ a finite differences approximation of the image gradient. This prior assumes that an image consists of a finite number of piecewise constant regions. This works very well for certain image types, but for general images this assumption does not hold and leads to the staircasing effect. However, a bilateral version of this prior has been exploited for robust multi-frame SR in [6]; Babacan et al. [1] use the TV prior in a Bayesian framework for multi-frame SR, Liu and Sun [11] for video SR, to name a few. A second-order extension of the TV prior is the Total Generalized Variation (TGV) [2] $\mathcal{R}(u) = \mathrm{TGV}_2(u) = \lambda_1\|\nabla u - v\|_{2,1} + \lambda_0\|Dv\|_{2,1}$, with $D \in \mathbb{R}^{4MNC\times 2MNC}$ and $v \in \mathbb{R}^{2MNC}$ which is able to get rid of staircasing effects in affine parts of images.

Learned Regularization In general, the structure of images is more complex than assumed by the previously described priors. Especially for the SR task, it would be beneficial to have a regularizer that can describe high frequency content, because this is especially hard to reconstruct from the LR images. Recently, Chen et al. proposed the trainable reaction-diffusion model [5] which can be interpreted as a generalization of regularization terms. The SR problem from (2) with these generalized regularization terms is embedded in a learning framework. This consists of unrolling a few steps of a simple projected gradient descent optimization algorithm and learning the whole reconstruction algorithm based on training data. A few advantages of this approach are: fast and efficient reconstruction, no parameter tuning at inference, as well as more expressive image priors. We will use this idea together with its extension to color images [8] to build our trained regularizer for the SR task.

The fundamental differences between image priors and the learned regularization is the dependence of the latter on available training data. This might be a drawback, but we solve this problem by careful design of the imaging setup and data simulation. In some settings, this will not be possible, at the expense

of image-per-image optimization using a fixed image prior and manual choice of regularization parameters.

There exist also a number of multi-frame SR approaches based on learned correspondences between LR and HR image pairs, disregarding the forward model (1), mostly based on sparse coding [12–14]. However, these methods are designed to reconstruct images patch by patch, which can cause artifacts when combining patches to form the final image. With our approach, we do not rely on patches, but rather reconstruct the whole image, independent of the input data size. In that sense, inference with our approach is similar to CNN models, which have been successfully applied to video superresolution [7, 9].

3 SR Method Description

We define the regularization term for the trained reconstruction algorithm as

$$\mathcal{R}(u;\theta) = \sum_{i=1}^{N_k} \sum_{p=1}^{MNC} \phi_i((K_i u)_p) \tag{3}$$

where the matrices K_i denote convolutions of the C channel image u with kernels $k_i \in \mathbb{R}^{h\times h\times C}$ as defined in [8]. N_k defines the number of activation function-kernel pairs ϕ_i and k_i. As a data term we use the model defined in (2) with different choices for $\psi(\cdot)$. The resulting minimization problem becomes

$$\min_{u\in\mathcal{U}} \mathcal{R}(u;\theta) + \lambda \sum_{c\in\mathcal{C}} \sum_{l=1}^{L} \mathcal{D}(u, f_l) \tag{4}$$

with λ weighting the influence of the data term. To obtain our reconstruction algorithm, we unroll a few projected gradient steps T of Problem (4)

$$u_{t+1} = \mathrm{proj}_{\mathcal{U}}(u_t - \nabla\mathcal{R}(u_t;\theta_t) - \lambda_t \sum_{c\in\mathcal{C}} \sum_{l=1}^{L} \nabla\mathcal{D}(u_t, f_l)) \tag{5}$$

and obtain the superresolved result u_T. Each step of the algorithm is parametrized by parameters θ_t. The projection onto the set $\mathcal{U}$ ensures that the result image lies in an admissible range of values, typically $\mathcal{U} = \{u \in \mathbb{R}^{MNC} : 0 \leq u_p \leq \xi, p = 1, \ldots, MNC\}$, with ξ being the maximal image intensity. The trainable activation functions are parametrized using N_w radial basis functions (RBFs) as

$$\phi_i'(z) = \sum_{j=1}^{N_w} w_{i,j} \exp\left(-\frac{(z-\mu_j)^2}{2\sigma^2}\right) \tag{6}$$

with equidistant means μ_j and fixed standard deviation σ for all components.

Parameters of the regularizer from (3) are summarized in the vector $\theta = \{w_{i,j,t}, k_{i,t}, \lambda_t\}_{i,j,t=1}^{N_k,N_w,T}$. These parameters comprise the step dependent weights

of the activation functions $w_{i,j,t}$, the convolution kernels $k_{i,t}$, and data term weights λ_t. Because we require that the convolution kernels are zero-mean and have norm one to ensure that the output of the convolution lies in the domain of the activation functions, we add constraints to ensure that the parameters θ lie in an admissible set $\mathcal{Y}$ (see supplemental). We train our algorithm based on a loss function comparing ground truth HR data with the output of (5)

$$\min_{\theta\in\mathcal{Y}} \mathcal{L}(\theta) := \frac{1}{2N_b} \sum_{b=1}^{N_b} \|u_{T,b}(\theta) - u_{gt,b}\|_2^2 \tag{7}$$

evaluated on a mini-batch of training data consisting of N_b samples. Training is performed using standard backpropagation. Optimization of (7) was performed using a stochastic inertial incremental proximal gradient (IIPG) optimization algorithm which accounts for the constraint $\theta \in \mathcal{Y}$ (see supplemental).

Parametrization of the data term As mentioned earlier, the function $\psi(\cdot)$ in the data term can be chosen in various ways. For our experiments, we consider a trained regularizer with following ℓ_2 data term as *type A*

$$\mathcal{D}(u, f_l) = \sum_{c\in\mathcal{C}} \sum_{l=1}^{L} \|B_c W_l \mathcal{I}_c u - f_{l,c}\|_2^2 \tag{8}$$

trained for a single color channel, and trained for 3 color channels as *type C*. Additionally, we use a data term

$$\mathcal{D}(u, f_l) = \sum_{c\in\mathcal{C}} \sum_{l=1}^{L} \sum_{j=1}^{N_d} \rho_j(\bar{K}_j B_c W_l \mathcal{I}_c u - f_{l,c}) \tag{9}$$

with N_d trainable filter-function pairs $\bar{K}_j$ and ρ_j which we refer to as *type B*. The parametrization of ρ_j is analogous to (6).

4 Data Acquisition

We designed our application such that prior knowledge about the geometry of the acquisition setup enables precise determination of the (affine) registration transformations between the individual views. As a result, the warping component W of the transformations' forward model (1) is constant for all acquisitions and can be specified accurately. As registration quality is a crucial part for successful multi-frame SR, reliable knowledge about the warping transformations is an advantage. Furthermore, it makes the comparison of different SR algorithms independent of adverse influences of registration inaccuracies.

Additionally to real acquisitions, a sufficient amount of data for training of the SR algorithm was simulated. In that process, we generated not only simulated acquisitions in the setup's resolution, but also required ground truth data in the targeted SR. For real acquisitions and for simulations, we used banknotes as they comprise fine-textured image structures that allow to point out improvements w.r.t. reconstruction quality.

4.1 Acquisition Setup

The acquisition setup comprises a camera with a multi-line scan sensor. Thereby, a selectable set of individual sensor lines can be read-out separately. While planar objects are transported orthogonal to those sensor lines, read-outs are done at according time instances. In the course of that line-scan procedure, each utilized sensor line yields a separate image of the object, where all those images are slightly translated versions of each other. As we use only one sensor, accurate registration is possible in practice.

To achieve the SR requirement of multiple, possibly equally distributed samples around each object point, the entire sensor is rotated slightly w.r.t. the transport direction. Thus the sensor lines are not exactly perpendicular to the transport direction anymore, while the sensor plane remains parallel to the object plane. Depending on the number and mutual distances between the utilized sensor lines, a rotation angle can be derived so that each object line is sampled by the total of pixels of all those sensor lines in an equally distributed manner at a higher sampling rate than a single sensor line could (see Fig. 2) - only distributed over different time instances. As a result of sensor rotation, the resulting images comprise an induced vertical shear, together with a slight scale compression along the sensor line directions, while they mutually are translated versions of each other.

Mere rotation of the sensor enables oversampling perpendicularly to the transport direction. To sample the objects also at a higher rate in transport direction, the transport speed has to be decelerated slightly, depending on the rotation angle, the number of utilized lines, and the mutual distances between the lines. This results in an equally distributed oversampling in transport direction by means of the set of sensor images. The speed reduction induces a further dilation in the individual sensor images which will be considered in the warping model, while they mutually still only differ by constant translations.

4.2 Real Acquisitions

Real acquisitions were conducted utilizing 4 sensor lines, resulting in a targeted SR upsampling factor of 2, although other configurations are possible. One sensor line has 2320 pixel, and the acquisition resolution was set to $100\,\mu$m per pixel. We used two different sensor angles $\alpha \sim 1.75^\circ$ and $\alpha \sim 1.15^\circ$, respectively, with accordingly reduced transportation speeds to achieve higher sampling rates in and perpendicular to the transport direction. From those acquisitions, the precise mutual image translations and rotation angles were measured and the corresponding warping transformations for the forward model (1) derived. The PSF of the camera system was measured for an acquisition of a calibration sheet to get an estimate for the blurring component B of the forward model (1) which is of size 21×21 in our experiments. By means of the slanted edge approach [3], the line spread functions in horizontal and vertical directions were used to derive a Gaussian PSF kernel with corresponding variances in both directions.

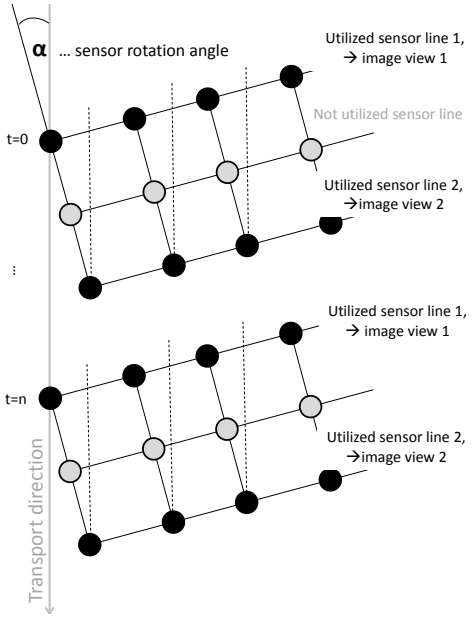


Fig. 2. The multi-line scan sensor is represented as grid of 12 pixels rotated by α w.r.t. the transport direction. Sensor lines with black pixel dots are actually utilized for line-scan acquisition. Those lines are read-out at time instances $t = 0, ...n$, thus each sensor line yields an image of $n+1$ image lines. Dashed vertical lines indicate that sensor line 2 samples the object space perpendicular to the transport direction at slightly translated positions w.r.t. sensor line 1, only at different time instances.

4.3 Simulated Acquisitions for Learning

Simulations were generated by mimicking the real acquisition process on 1200 dpi scans of three banknotes and one calibration sheet. Estimates of the forward warping model W, the corresponding PSF estimate B, and a downsampling factor of 5 were derived from real acquisitions and were applied at 1000 random image positions of each source scan. For training the SR algorithms, 4000 image quadruplets at 100 μm per pixel resolution were generated together with a corresponding ground truth image (single-line scan, unrotated) per quadruplet in double resolution, i.e. the SR target resolution.

5 Experiments and Results

We conducted experiments for two setups with angles $\alpha = 1.15°$ and $\alpha = 1.75°$ as described in Section 4. For each setup, three different SR algorithms were trained comprising of trained regularizers along with different data terms. The steps T were chosen to balance computation time and reconstruction accuracy.

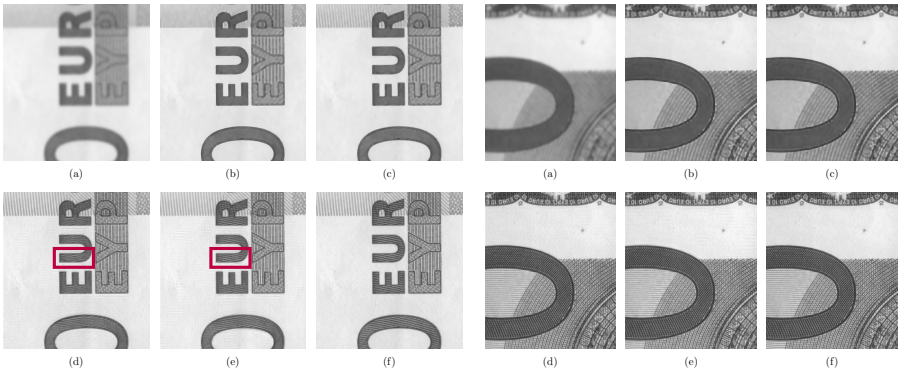


Fig. 3. Left: $\alpha = 1.15^\circ$, Right: $\alpha = 1.75^\circ$. (a) The average solution, (b) the TV reconstruction, (c) the TGV reconstruction, (d) the reconstruction with the trained regularizer *type A*, (e) trained regularizer *type B*, (f) ground truth.

The acquired data was split into distinct training and test sets. For all trained algorithms we used 400 images for training, and tested on 800 images. The resulting run time of the algorithm is 0.5s per MP on average using a current GPU. Further training details can be found in the supplemental.

Trained Type A This reconstruction algorithm is trained for a single channel (gray) and is optimized for $T = 10$ projected gradient steps. The convolution kernel size in the regularizer is 7×7, the data term is according to (8).

Trained Type B The setup of the regularizer is the same as *type A*, the data term is according to (9). The data term functions are initialized to ℓ_2 functions, and the data term convolution kernels are of size 5×5.

Trained Type C This reconstruction algorithm is trained for three color channels and is optimized for $T = 10$ projected gradient steps. The convolution kernel size in the regularizer is $5 \times 5 \times 3$, the data term is according to (8).

Results for both algorithms *type A* and *B* are shown in Fig. 3, and we observe that fine structures are nicely reconstructed by the trained regularizer where the TV and TGV regularized solutions fail. We also observe some hallucinated image structure in texture-less regions which can be seen both as a strength and limitation of our approach, because this effect is in general very helpful for reconstructing fine details. It is challenging to find an optimal trade-off between smoothing and enhancing fine image structures, which can be controlled by choosing the "right" training data. As the training data contains many oscillating patterns, there is a subtle bias towards those in the reconstructions. Comparing Fig. 3(d) and (e) we observe that the trainable data term function and kernel pairs help to reconstruct the fine stripe pattern marked in the images.

Qualitative results for the algorithm *type C* with the regularizer trained on color images are shown in Fig. 4 and 5. Again, fine details and text are nicely

(a) (b) (c)

Fig. 4. $\alpha = 1.15°$ (a) Average reconstruction, (b) the reconstruction with trained regularizer *type C*, (c) ground truth.

Table 1. Results for 800 images from the test set (simulated data)

Angle	$\alpha \approx 1.15°$		$\alpha \approx 1.75°$	
Reconstruction	PSNR	SSIM	PSNR	SSIM
Average	22.43	0.4915	22.35	0.4924
TV	25.00	0.5886	25.12	0.5906
TGV	25.63	0.6117	25.78	0.6146
Trained Type A	29.51	0.7900	29.36	0.7856
Trained Type B	29.30	0.8165	29.06	0.8184
Trained Type C	29.61	0.9121	29.56	0.9102

recovered. In the color setting hallucinating of structures in homogeneous areas is hardly visible compared to the single-channel case, which is apparent in Fig. 5(b). However, we observe some ringing artifacts which are due to over-enhancing little edges in the image.

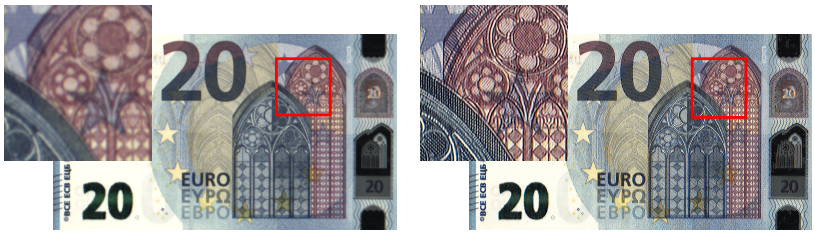
The SR results are evaluated in terms of Peak Signal to Noise Ratio (PSNR) and Structured Similarity Index (SSIM). The error measures are only evaluated in the image area where all observations overlap, because the reconstruction is not valid outside this area. As a baseline, we compare our algorithms with the solutions of TV and TGV regularized Problem (2) which were solved with a first-order primal-dual algorithm [4].

In Table 1 we summarize the performance of the different algorithms. It is remarkable that by using color data *type C* the SSIM index is much better compared to the results using only single-channel images, which is due to hallucinated structure in texture-less regions. PSNR values are similar in both cases. The regularizer *type A* yields better PSNR than *type B*, but the SSIM results are reversed. We believe this is due to the "invented" checker board patterns which are a little more pronounced for *type B*, as the results in general appear sharper.

We also tested our trained color image regularizer *type C* on acquired real data, see Fig. 6. Compared to the average solution, the proposed method leads to significant improvement in reconstructing high frequency content visible in

(a) (b) (c)

Fig. 5. $\alpha = 1.75°$ (a) Average reconstruction, (b) the reconstruction with trained regularizer *type C*, (c) ground truth.



(a) (b)

Fig. 6. Real acquisition data, $\alpha = 1.15°$ (a) Average reconstruction, (b) the reconstruction with trained regularizer *type C*.

the zoomed views. However, we also observe some overly enhanced edges and a few small artifacts, which stem from imperfections in the imaging setup.

6 Conclusion

In this paper, we proposed a fully learned variational model to improve the resolution of data acquired using a multi-line scan camera. We showed that the learned regularizers can successfully recover high frequency content which is especially apparent when inspecting fine textures. We showed that the capabilities of the reconstruction algorithm trained on simulated data also transfer to real data. The imaging setup together with the novel SR reconstruction algorithm enables faster, memory-efficient data acquisition together with increased image quality and near real-time reconstruction time.

Acknowledgements

We acknowledge grant support from the FWF START project BIVISION, No. Y729, the ERC starting grant HOMOVIS, No. 640156 and from the AIT and the Austrian Federal Ministry of Science under the HRSM programme BGBl. II Nr. 292/2012.

References

1. Babacan, S.D., Molina, R., Katsaggelos, A.K.: Variational Bayesian Super Resolution. IEEE Transactions on Image Processing 20(4), 984–999 (2011)
2. Bredies, K., Kunisch, K., Pock, T.: Total Generalized Variation. SIAM Journal on Imaging Sciences 3(3), 492–526 (2010)
3. Burns, P.D.: Slanted-Edge MTF for Digital Camera and Scanner Analysis. In: Proc. PICS Conference, IS&T. pp. 135–138 (2000)
4. Chambolle, A., Pock, T.: A First-order Primal-dual Algorithm for Convex Problems with Applications to Imaging. Journal of Mathematical Imaging and Vision 40(1) (2011)
5. Chen, Y., Yu, W., Pock, T.: On Learning Optimized Reaction Diffusion Processes for Effective Image Restoration. In: Computer Vision and Pattern Recognition. pp. 5261–5269 (2015)
6. 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)
7. Kappeler, A., Yoo, S., Dai, Q., Katsaggelos, A.K.: Video Super-Resolution with Convolutional Neural Networks. IEEE Transactions on Computational Imaging PP(99), 1–1 (2016)
8. Klatzer, T., Hammernik, K., Knöbelreiter, P., Pock, T.: Learning Joint Demosaicing and Denoising Based on Sequential Energy Minimization. In: IEEE International Conference on Computational Photography (2016)
9. Liao, R., Tao, X., Li, R., Ma, Z., Jia, J.: Video Super-resolution via Deep Draft-ensemble Learning. In: International Conference on Computer Vision (2015)
10. Lim, K.H., Kwoh, L.K.: Super-Resolution for Spot5 - Beyond Supermode. In: Asian Conference on Remote Sensing (2009)
11. Liu, C.S.: On Bayesian Adaptive Video Super Resolution. IEEE Transactions on Pattern Analysis and Machine Intelligence 36(2)(2), 346–360 (2014)
12. Song, B.C., Jeong, S.C., Choi, Y.: Video Super-resolution Algorithm Using Bi-directional Overlapped Block Motion Compensation and On-the-fly Dictionary Training. IEEE Transactions on Circuits and Systems for Video Technology 21(3), 274–285 (2011)
13. Yang, J., Wang, Z., Lin, Z., Cohen, S., Huang, T.: Coupled Dictionary Training for Image Super-resolution. IEEE Transactions on Image Processing 21(8), 3467–3478 (2012)
14. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image Super-Resolution Via Sparse Representation. IEEE Transactions on Image Processing 19(11), 2861–2873 (2010)
15. Yue, L., Shen, H., Li, J., Yuan, Q., Zhang, H., Zhang, L.: Image Super-resolution: The Techniques, Applications, and Future. Signal Processing 128, 389–408 (2016)