Non-Local Kernel Regression for Image and Video Restoration

Haichao Zhang[1,2], Jianchao Yang[2], Yanning Zhang[1], and Thomas S. Huang[2]

[1] School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2] Beckman Institute, University of Illinois at Urbana-Champaign, USA
{hczhang,jyang29,huang}@ifp.uiuc.edu, ynzhang@nwpu.edu.cn

Abstract. This paper presents a non-local kernel regression (NL-KR) method for image and video restoration tasks, which exploits both the non-local self-similarity and local structural regularity in natural images. The non-local self-similarity is based on the observation that image patches tend to repeat themselves in natural images and videos; and the local structural regularity reveals that image patches have regular structures where accurate estimation of pixel values via regression is possible. Explicitly unifying both properties, the proposed non-local kernel regression framework is robust and applicable to various image and video restoration tasks. In this work, we are specifically interested in applying the NL-KR model to image and video super-resolution (SR) reconstruction. Extensive experimental results on both single images and realistic video sequences demonstrate the superiority of the proposed framework for SR tasks over previous works both qualitatively and quantitatively.

1 Introduction

One of the recent trends in image processing is to pursue the low-dimensional models for image representation and manipulation. Examples include the local structure based methods [1], sparse representation methods [2][3], manifold methods [4], etc. The success of such models is guaranteed by the low Degree of Freedom (DOF) of the local structures in natural images, represented as meaningful local structural regularity as well as self-similarity of local patterns.

Many conventional image processing algorithms are based on the assumption of local structural regularity, meaning that there are meaningful structures in the spatial space of natural images. Examples are structure tensor based methods [1][5][6]and bilateral filtering [7]. These methods utilize the local structural patterns to regularize the image processing procedure and are based on the assumption that images are locally smooth except at edges.

Another type of methods exploiting the self-similarity in natural images are recently emerging. The self-similarity property means that higher level patterns (e.g., texton and pixon) will repeat themselves in the image (possibly in different scales). This also indicates the DOF in one image is less than the DOF offered by the pixel-level representation. A representative work is the popular Non-Local Means (NL-Means) [8], which takes advantage of the redundancy of

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part III, LNCS 6313, pp. 566–579, 2010.

similar patches existing in the target image for denoising tasks. Later, this idea is generalized to handle multi-frame super-resolution tasks in [9]. Recently, this self-similarity property is thoroughly explored by Glasner *et. al* in [10] for addressing single image super-resolution problems. Gabriel Peyré *et. al* proposed a non-local regularization method for general inverse problems [11].

We propose in this paper a Non-Local Kernel Regression (NL-KR) method for image and video restoration (see Fig. 1 for a graphical illustration of the proposed model). We take advantage of both local structural regularity and non-local similarity in a unified framework for more reliable and robust estimation. The non-local similarity and local structural regularity are intimately related, and are also complimentary in the sense that non-local similar pattern fusion can be regularized by the structural regularity while the redundancy from similar patterns enables more accurate estimation for structural regression.

The rest of the paper is organized as follows. We first review and summarize related works in Section 2, then we propose our NL-KR model and discuss its relations to other algorithms in Section 3. The practical algorithm for SR based on NL-KR is described in Section 4. Experiments are carried out in Section 5 on both synthetic and real image sequences, and extensive comparisons are made with both classical as well as *state-of-the-art* methods. Section 6 provides some discussions and concludes our paper.

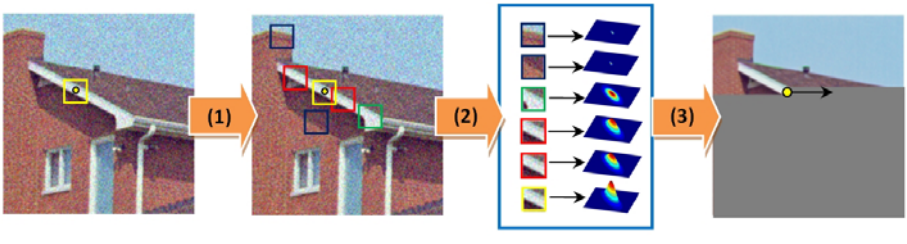

Fig. 1. Non-Local Kernel Regression. (1) Similar patch searching: different colors indicate the similarity (red the highest, green the medium and blue the least); (2) Structural kernel estimation and reweighting: estimate a regression kernel adapted to the structure at each position where the similar patches reside and re-weight them according to similarity; (3) Non-local kernel regression: estimate the value for the query point with both local structural and non-local similar information in raster-scan order.

2 Related Works

In this work, we are interested in image and video restorations where we desire to estimate the pixel value of a given location in the image plane (e.g., image super-resolution, inpainting and denoising). This section presents a brief technical review of local structural regression or filtering method as well as the non-local similarity-based approach.

2.1 Local Structural Regression

Typical image filtering methods usually perform in a local manner, i.e., the value of the estimated image at a query location is influenced only by the pixels within a small neighborhood of that position. They usually take the form of:

$$\hat{z}(\boldsymbol{x}_i) = \arg\min_{z} \sum_{j \in \mathcal{N}(\boldsymbol{x}_i)} (y_j - z)^2 K_{\boldsymbol{x}_i}(\boldsymbol{x}_j - \boldsymbol{x}_i) \tag{1}$$

where $\mathcal{N}(\boldsymbol{x}_i)$ denotes the neighbors of $\boldsymbol{x}_i$, and $K_{\boldsymbol{x}_i}(\boldsymbol{x}_j - \boldsymbol{x}_i)$ is the spatial kernel at location $\boldsymbol{x}_i$ that assigns larger weights to nearby similar pixels while smaller weights to farther non-similar pixels. Since the local image structure is not isotropic, local structure aware kernels are developed, with representative examples as Orientated Gaussian kernel [1] and Bilateral kernel [7]. To approximate the local structure better, higher order estimation can be used:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|R_{\boldsymbol{x}_i} Y - \boldsymbol{\Phi}\mathbf{a}\|^2_{W_{K_{\boldsymbol{x}_i}}} \tag{2}$$

Here $\boldsymbol{\Phi}$ is the polynomial bases given in Eq. 3 developed from Taylor expansion[1] with $\mathbf{a}$ the corresponding regression coefficients, and $\mathbf{tril}(\cdot)$ extracts the lower triangular part of a matrix and stack it to a column vector. $W_{K_{\boldsymbol{x}_i}} = \mathrm{diag}[K_{\boldsymbol{x}_i}(\boldsymbol{x}_1 - \boldsymbol{x}_i), K_{\boldsymbol{x}_i}(\boldsymbol{x}_2 - \boldsymbol{x}_i), \cdots, K_{\boldsymbol{x}_i}(\boldsymbol{x}_m - \boldsymbol{x}_i)]$ $(m = |\mathcal{N}(\boldsymbol{x}_i)|)$ is the weight matrix defined by the kernel. $R_{\boldsymbol{x}_i}$ takes a patch centered at $\boldsymbol{x}_i$ from Y and represents it as a vector.

$$\boldsymbol{\Phi} = \begin{bmatrix} 1 & (\boldsymbol{x}_1 - \boldsymbol{x}_i)^T & \mathbf{tril}\{(\boldsymbol{x}_1 - \boldsymbol{x}_i)(\boldsymbol{x}_1 - \boldsymbol{x}_i)^T\}^T & \cdots \\ 1 & (\boldsymbol{x}_2 - \boldsymbol{x}_i)^T & \mathbf{tril}\{(\boldsymbol{x}_2 - \boldsymbol{x}_i)(\boldsymbol{x}_2 - \boldsymbol{x}_i)^T\}^T & \cdots \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (\boldsymbol{x}_m - \boldsymbol{x}_i)^T & \mathbf{tril}\{(\boldsymbol{x}_m - \boldsymbol{x}_i)(\boldsymbol{x}_m - \boldsymbol{x}_i)^T\}^T & \cdots \end{bmatrix} \tag{3}$$

Therefore, the first element of the regression coefficient is the desired pixel value estimation at $\boldsymbol{x}_i$.

$$\hat{z}(\boldsymbol{x}_i) = \mathbf{e}_1^T \left[\boldsymbol{\Phi}^T W_{K_{\boldsymbol{x}_i}} \boldsymbol{\Phi}\right]^{-1} \boldsymbol{\Phi}^T W_{K_{\boldsymbol{x}_i}} R_{\boldsymbol{x}_i} Y \tag{4}$$

where $\mathbf{e}_1$ is a vector with the first element equal to one, and the rest zero.

2.2 Non-Local Similarity-Based Estimation

Local image structures tend to repeat themselves across the image and also the image sequence in videos. This property has been explored in many applications

[1] The regression bases do not have to be polynomial, and other choices are open. For more details about deriving the polynomial bases, one can refer to [6], which gives a nice tutorial on kernel regression.

such as texture synthesis [12], image inpainting [13], denoising [8] and super-resolution [9] [14]. This self-similarity property provides the redundancy that is sometimes critical for many ill-posed image processing problems, as similar structures can be regarded as multiple observations from the same underlying ground truth. For instance, the NL-Means algorithm recently introduced by Buades *et al.* in [8] for image denoising has become very popular, due to its effectiveness despite of its simplicity. The algorithm breaks the locality constraints of previous conventional filtering methods, making use of similar patterns found in different locations of the image to denoise the image. Specifically, NL-Means algorithm is a weighted average:

$$z(\boldsymbol{x}_i) = \frac{\sum_{j\in\mathcal{P}(\boldsymbol{x}_i)} w_{ij} y_j}{\sum_{j\in\mathcal{P}(\boldsymbol{x}_i)} w_{ij}} \tag{5}$$

where $\mathcal{P}(\boldsymbol{x}_i)$ denote the index set for similar pixel observations for $\boldsymbol{x}_i$ (includes $\boldsymbol{x}_i$ itself). The weight w_{ij} reflects the similarity between the observations at $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ [8]. Eq. 5, can be reformulated into an optimization problem:

$$\begin{aligned}\hat{z}(\boldsymbol{x}_i) &= \arg\min_{z} \sum_{j\in\mathcal{P}(\boldsymbol{x}_i)} \left[y_j - z(\boldsymbol{x}_i)\right]^2 w_{ij} \\ &= \arg\min_{z} \|\boldsymbol{y} - \mathbf{1}z(\boldsymbol{x}_i)\|^2_{W_{\boldsymbol{x}_i}}\end{aligned} \tag{6}$$

where $\boldsymbol{y}$ denotes the vector consisting of the pixels at the locations in the similar set $\mathcal{P}(\boldsymbol{x}_i)$, $\mathbf{1}$ denotes a vector of all ones, and $W_{\boldsymbol{x}_i} = \mathrm{diag}\,[w_{i1}, w_{i2}, ..., w_{im}]$ ($m = |\mathcal{P}(\boldsymbol{x}_i)|$). Compared with Eq. 2, the NL-Means estimation Eq. 6 can be regarded as a zero-order estimation, and the weight matrix is constructed from the similarity measures instead of the spatial kernel as before.

3 Non-Local Kernel Regression Model

3.1 Mathematical Formulation

We derive the Non-Local Kernel Regression (NL-KR) algorithm in this section. The approach makes full use of both cues from *local* structural regularity and *non-local* similarity for image and video restoration. We argue that the proposed approach is more reliable and robust for ill-posed inverse problems: local structural regression regularize the noisy candidates found by non-local similarity search; and non-local similarity provides the redundancy preventing possible overfitting of the local structural regression. Instead of using a *point prediction* model in non-local methods, we use the more reliable *local structure*-based prediction. On the other hand, rather than predicting the value with only one *local patch*, we can try to make use of all the *non-local similar* patches in natural images. Mathematically, the proposed high-order Non-Local Kernel Regression model is formulated as:

$$\begin{aligned}\hat{\mathbf{a}} &= \arg\min_{\mathbf{a}} \frac{1}{2}\overbrace{w_{ii}\|R_{\boldsymbol{x}_i}Y - \boldsymbol{\Phi}\mathbf{a}\|^2_{W_{K_{\boldsymbol{x}_i}}}}^{local} + \frac{1}{2}\overbrace{\sum_{j\in\mathcal{P}(\boldsymbol{x}_i)\setminus\{i\}} w_{ij}\|R_{\boldsymbol{x}_j}Y - \boldsymbol{\Phi}\mathbf{a}\|^2_{W_{K_{\boldsymbol{x}_j}}}}^{non-local} \\ &= \arg\min_{\mathbf{a}} \frac{1}{2}\sum_{j\in\mathcal{P}(\boldsymbol{x}_i)} w_{ij}\|R_{\boldsymbol{x}_j}Y - \boldsymbol{\Phi}\mathbf{a}\|^2_{W_{K_{\boldsymbol{x}_j}}} \\ &= \arg\min_{\mathbf{a}} \frac{1}{2}\sum_{j\in\mathcal{P}(\boldsymbol{x}_i)} \|R_{\boldsymbol{x}_j}Y - \boldsymbol{\Phi}\mathbf{a}\|^2_{\tilde{W}_{\boldsymbol{x}_j}}\end{aligned} \tag{7}$$

where $W_{K_{\boldsymbol{x}_j}}$ is the weight matrix constructed from kernel $K_{\boldsymbol{x}_j}$, and $\mathcal{P}(\boldsymbol{x}_i)$ again is the similar index set for $\boldsymbol{x}_i$. w_{ij} is calculated between the location $\boldsymbol{x}_i$ of interest and any position $\boldsymbol{x}_j$ ($j \in \mathcal{P}(\boldsymbol{x}_i)$) by measuring the similarity of their neighborhoods weighted by a Gaussian kernel:

$$w_{ij} = \exp\left(-\frac{\|R_{\boldsymbol{x}_i}Y - R_{\boldsymbol{x}_j}Y\|^2_{W_G}}{2\sigma^2}\right) \tag{8}$$

where W_G is the weight matrix constructed from a Gaussian kernel, which puts larger weights on the centering pixels of the patch. The proposed NL-KR regression model consists of two parts:

1. **Local regression term:** the traditional local regression or filtering term, with w_{ii} set to be 1. This term also contributes as a fidelity loss, as the estimation should be close to the observation.
2. **Non-local regression term:** instead of zero-order point estimation as in Non-Local means, higher-order kernel regression is also used to make full use the structural redundancy in the similar patches.

The effects of these two parts will be more clear with experimental results in Section 5. To get the regression coefficients, differentiate the right hand side of Eq. 7 with respect to $\mathbf{a}$ and set it to zero, we have

$$\hat{\mathbf{a}} = \left[\boldsymbol{\Phi}^T\left(\sum_{j\in\mathcal{P}(\boldsymbol{x}_i)} w_{ij}W_{K_{\boldsymbol{x}_j}}\right)\boldsymbol{\Phi}\right]^{-1}\boldsymbol{\Phi}^T\sum_{j\in\mathcal{P}(\boldsymbol{x}_i)} w_{ij}W_{K_{\boldsymbol{x}_j}}R_{\boldsymbol{x}_j}Y \tag{9}$$

Then $\hat{z}(\boldsymbol{x}_i) = \mathbf{e}_1^T\hat{\mathbf{a}}$. Examination on Eq. 9, we have the following two comments:

- The structural kernel is estimated from contaminated observations, and thus is not robust. Compared with Eq. 4, with non-local redundancy, our estimation is more stable because of the weighted average of kernel weight matrices inside the inverse, and the weighted average of the structural pixel values.
- Compared with Eq. 6, our model can regularize the estimation from the non-local patches by structural higher-order regression, and thus is more robust to outliers.

Therefore, the proposed model make full use of both important cues from local structure and non-local similarity, leading to more reliable and robust estimation, which will be verified by experimental results in Sec. 5.

3.2 Structural Kernel Estimation

It is desirable to use a structure adaptive kernel in estimation. Given the observation Y and a query location $\boldsymbol{x}_i$, we can construct a structure adaptive kernel as:

$$K_{\boldsymbol{x}_i}(\boldsymbol{x}-\boldsymbol{x}_i)=\frac{1}{\sqrt{\det(T)}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_i)^T T^{-1}(\boldsymbol{x}-\boldsymbol{x}_i)\right\} \tag{10}$$

where T is the diffusion tensor at $\boldsymbol{x}_i$ controlling the spatial structure of the kernel. Given two unit vectors $\mathbf{u}$ and $\mathbf{v}$ defined by the gradient and tangent direction respectively, we can construct $T=f\mathbf{u}\mathbf{u}^T+g\mathbf{v}\mathbf{v}^T$ and adjust f and g according to the underlying structure, so that the induced kernel is isotropic ($f\approx g$) at almost constant regions and aligned along the image contour ($g>f$) otherwise. One possible choice [2] is $f(\alpha,\beta)=\frac{\beta+\gamma}{\alpha+\gamma}$ and $g(\alpha,\beta)=\frac{\alpha+\gamma}{\beta+\gamma}$ ($\gamma\geq 0$), where α and β are the eigen values of the structure tensor [1], reflecting the strength of the gradient along each eigenvector directions. α, β, $\mathbf{u}$ and $\mathbf{v}$ can be calculated from the following relation using singular value decomposition (SVD):

$$\nabla Y_{\boldsymbol{x}_i}\nabla Y_{\boldsymbol{x}_i}^T=\alpha\mathbf{u}\mathbf{u}^T+\beta\mathbf{v}\mathbf{v}^T \tag{11}$$

where $\nabla Y_{\boldsymbol{x}_i}$ is a 2×1 vector, denoting the gradient of Y at $\boldsymbol{x}_i$.

3.3 Relations to Other Works

Tons of works have emerged recently based on non-local redundancy and local regressions for image and video processing. It is worthwhile to talk about the relations of the proposed model to those previously proposed algorithms. The non-local models in [4], [8] and [9] use the redundancy from non-local self-similarity, but do not include the spatial structure explicitly as a regularization. The high order generalization of non-local means in [15] uses the computation of non-local similarity to find the local kernel for regression, which actually violates the philosophy of the non-local model. Local structural regression [1][6][7] explicitly employ the spatial kernel for regularization, but neglect the redundancy of similar local patterns useful for robust estimation. The 3D kernel regression method [16] exploits the local spatial-temporal structure by extending their 2D spatial kernel regression, also discards the non-local self-similarity. Sparse representation for denoising [2] and super-resolution [3] do local regression using bases learned from a training database. They perform estimation on each individual local patch and discard the patch redundancy. The sparse representation model is later generalized in [17] for image denoising by doing simultaneous sparse coding over similar patches found in different locations of the image. However, the non-local redundancy is used in a hard assignment clustering way instead of a soft way. [10] fully explores the self-similarity property for single image super-resolution, but no spatial structural regularization is applied. To summarize, our model is the first work to explicitly unify the self-similarity and local structural regularization into a single model, allowing more robust estimation.

[2] One can refer to [1] for other choices of diffusion tensor.

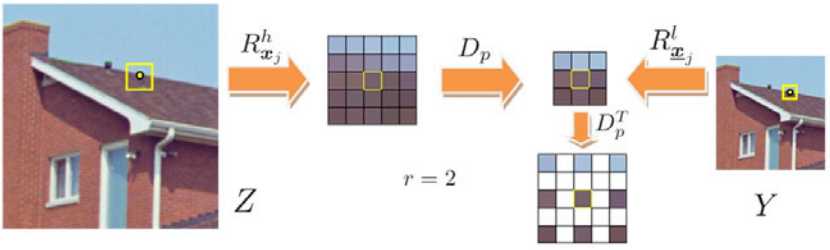

Fig. 2. Operator Illustration. $R^l_{\underline{\boldsymbol{x}}_j} Y$, the patch of the LR image Y at $\underline{\boldsymbol{x}}_j$, is formed by downsampling HR patch $R^h_{\boldsymbol{x}_j} Z$ by factor $r = 2$ as $D_p R^h_{\boldsymbol{x}_j} Z$, keeping the center pixel still in the center. Operator D_p^T up samples a patch with zero padding.

4 Non-Local Kernel Regression for Super-Resolution

The NL-KR model proposed above is a general model that can be applied to many image and video restoration tasks. In this work, we specifically apply the formulation Eq. 7 to image and video super-resolution.

Image super-resolution (SR) aims to estimate a high-resolution image (HR) from a single or a set of low-resolution (LR) observations. Conventional multi-frame SR follows the steps of (1) global motion estimation, (2) image wrapping and (3) data fusion. These methods are limited in the assumed global motion model, and can not be applied to realistic videos that almost certainly contain arbitrary motion patterns. Recently, several multi-frame SR algorithms based on fuzzy motion estimation of local image patches are proposed to process real videos [9][16]. We will show that similarly our model can also be applied to realistic videos, while achieving better results both qualitatively and quantitatively. Besides, due to the self-similarity property of the image, we can also perform single frame SR without additional training, arguing that the motion may not be that critical as in the conventional SR cases for image resolution enhancement. The LR image frames are usually modeled as blurring and downsampling the desired HR image, i.e.:

$$Y_k = D_k H X + \epsilon_k = DZ + \epsilon_k, k = 1, 2, ... \tag{12}$$

where D_k is the downsampling operator, H is the blurring operator and ϵ_k is a noise term, and k is the LR frame index. Therefore, the SR recovery problem can be divided into two steps: LR image fusion and deblurring. In our NL-KR model, we also target recovering Z followed by deblurring. As now we have two different spatial scales, i.e., high- and low-resolution image grids, the following notations are introduced for ease of presentation. We let r denote the zoom factor, $\boldsymbol{x}$ and $\underline{\boldsymbol{x}}$ denote the coordinates on HR and LR grids respectively. R^h and R^l denote the patch extraction and vectorization operator on HR and LR images, where the extracted vectors are of dimension $u^2 \times 1$ and $v^2 \times 1$ respectively, and $u = (v - 1) \times r + 1$ relates the two spatial scales. D_p is a patch downsampling

operator which keeps the center pixel of the patch on the LR grid, while D_p^T is a patch upsampling operator with zero-padding (refer to Fig. 2). For a given query position $\boldsymbol{x}_i$ on the HR grid, $\mathcal{P}(\boldsymbol{x}_i)$ can be constructed from the initial HR estimation of the current image or consecutive frames, while keeping only those corresponding to integer positions on the LR grid, i.e., $\boldsymbol{x}_j = (\underline{\boldsymbol{x}}_j - 1) \times r + 1$ ($j \in \mathcal{P}(\boldsymbol{x}_i)$). Then using Eq. 7 and Eq. 12, the NL-KR model tailored for SR tasks is formulated as:

$$\begin{aligned} \hat{\mathbf{a}} &= \arg\min_{\mathbf{a}} \frac{1}{2} \sum_{j \in \mathcal{P}(\boldsymbol{x}_i)} \| D_p^T (R^l_{\underline{\boldsymbol{x}}_j} Y - D_p \boldsymbol{\Phi} \mathbf{a}) \|^2_{\tilde{W}_{\boldsymbol{x}_j}} \\ &= \arg\min_{\mathbf{a}} \frac{1}{2} \sum_{j \in \mathcal{P}(\boldsymbol{x}_i)} \| R^l_{\underline{\boldsymbol{x}}_j} Y - D_p \boldsymbol{\Phi} \mathbf{a} \|^2_{\tilde{W}^D_{\boldsymbol{x}_j}} \end{aligned} \tag{13}$$

where we denote $\tilde{W}_{\boldsymbol{x}_j} = w_{ij} W_{K_{\boldsymbol{x}_j}}$ to keep the notation uncluttered, $\boldsymbol{\Phi}\mathbf{a}$ is a high order regression for the patch $R^h_{\boldsymbol{x}_j} Z$ centered at query location $\boldsymbol{x}_j$ for the blurred HR image Z, and $\tilde{W}^D_{\boldsymbol{x}_j} = D_p \tilde{W}_{\boldsymbol{x}_j} D_p^T$. Solution of Eq. 13 is straightforward:

$$\hat{\mathbf{a}} = \left[\boldsymbol{\Phi}^T \left(\sum_{j \in \mathcal{P}(\boldsymbol{x}_i)} D_p^T \tilde{W}^D_{\boldsymbol{x}_j} D_p \right) \boldsymbol{\Phi} \right]^{-1} \boldsymbol{\Phi}^T \sum_{j \in \mathcal{P}(\boldsymbol{x}_i)} D_p^T \tilde{W}^D_{\boldsymbol{x}_j} R^l_{\underline{\boldsymbol{x}}_j} Y \tag{14}$$

The estimated pixel value at query point $\boldsymbol{x}_i$ is $\mathbf{e}_1^T \hat{\mathbf{a}}$.

As we can see, the missing pixels in the high resolution grid are filled up by multiple low resolution observations found in a non-local way on the current frame or current sequence. These low resolution observations are further fused with regularization from the local structure. The estimated image is then deblurred with a Total Variation based algorithm [18]. Algorithm 1 describes the practical implementation for the proposed model.

5 Experimental Validation

The proposed NL-KR model can handle both single image and multiple frame SR naturally. In this section, we validate the performance of the proposed method with experiments on single images, synthetic and real video sequences. Performance comparisons are performed with related *state-of-the-art* algorithms. We use both Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM) index [19] to evaluate different algorithms objectively.

In all the experiments, we focus on zooming the LR frame(s) by factor of 3. These LR frames are modeled by first blurring the HR frames with a 3×3 uniform PSF and downsampling with decimation factor of 3. Gaussian noise of standard deviation 2 is added to the LR frames to model the real imaging system. In our algorithm, the LR patch size is fixed as 5×5, and the corresponding HR patch size is thus 13×13. The support of the non-local similar patch searching is fixed to be the 10-nearest neighbors. We set $\sigma = 169c$ with $c = 0.06$ for similarity weight calculation and $\gamma = 1$ for diffusion tensor computation. For image deblurring, we use a Total Variation based deblurring algorithm [18].

Algorithm 1. (Non-local kernel regression for image super-resolution).

1: **Input:** a low resolution video sequence $\underline{Y} = [Y_1, Y_2, ..., Y_M]$ and zoom factor r.
2: **Initialize** an enlarged sequence $\underline{\tilde{Y}} = [\tilde{Y}_1, \tilde{Y}_2, ..., \tilde{Y}_M]$ with bicubic interpolation.
3: **For** each pixel location $\boldsymbol{x}_i$ on the high resolution image grid for frame Y_m, do

- Construct the similar patch index set $\mathcal{P}(\boldsymbol{x}_i)$ with sequence $\underline{\tilde{Y}}$;
- Estimate the image gradient $\nabla Y_{\boldsymbol{x}_i}$;
- Calculate the diffusion tensor $K_{\boldsymbol{x}_i}$ use Eq. 11 and Eq. 10.

4: **End**
5: **For** each pixel location $\boldsymbol{x}_i$ on the high resolution image grid for frame Y_m, do

- Construct the spatial weight matrix $\tilde{W}^D_{\boldsymbol{x}_j}$ using estimated $K_{\boldsymbol{x}_j}$ for all $j \in \mathcal{P}(\boldsymbol{x}_i)$;
- Calculate the regression coefficients with Eq. 14 and update the current estimation of Z_m at $\boldsymbol{x}_i$ with $Z_m(\boldsymbol{x}_i) = \mathbf{e}_1^T \hat{\mathbf{a}}$.

6: **End**
7: **Perform** deblurring for Z_m: $X_m = \text{TVdeblur}(Z_m)$.
8: **Output:** a high resolution video frame X_m.

5.1 Single Frame Based Super-Resolution

We will first evaluate the proposed method on single image SR. In the first set of experiments, we specifically compare the proposed model with 2D case of Generalized NL-Means (GNL-Means) [9] and Kernel Regression (KR)[6] in order to show that our model is more reliable and robust for estimation. We take one frame from each of the popular test sequences: Foreman, Miss America and Suzie used in [9], degrade it and perform the SR estimation. The PSNR and SSIM results for the three frames are summarized in Table 1, which shows that the proposed method is constantly better than 2D GNL-Means and 2D Kernel Regression. The results of Nearest Neighbor (NN), Bicubic Interpolation (BI) and Sparse Coding (SC) method [3] are also provided as references. Fig. 3 shows the visual quality comparisons on Foreman. As shown, the 2D GNL-Means method is prone to block artifacts due to poor patch matching within a single image and the 2D Kernel Regression method generates ghost effects due to insufficient observation for regularizing the local regression. Our result, however, is free of either of these artifacts. We further make more comparisons with *state-of-the-art* methods on real images, where the input LR image is zoomed by a factor of 4, as shown in Fig. 4 and Fig. 5. Note that these methods are designed specifically to work on single images. In Fig. 4, it can be seen that the proposed method can preserve more details than Fattal's method [20] and is comparable with Kim's method [21] and the more recent work [10]. In Fig. 5, however, our algorithm outperforms both [20] and [10], where our result is free of the *jaggy* artifacts on the edges and the characters generated by our method is more realistic. The improvement comparison could be more impressive if one notices that in [10], multiple scales are used for similar pattern matching while our method only uses one scale.

Fig. 3. Single-frame super resolution (×3, PSNR, SSIM in brackets). Left to right: NN(28.6917, 0.8185), BI(30.9511, 0.8708), GNL-Means(31.9961, 0.8747)[9], KR(32.4479, 0.8862)[6], NL-KR(**32.7558, 0.8918**). GNL-Means generates block effect while KR generates ghost artifacts. Our method does not suffer from these problems.

Fig. 4. Single-frame super resolution for real color images (×4). From left to right: NN, BI, Kim's method [21], Fattal's method [20], Glasner's method [10], NL-KR. Note that our result preserves more details than Fattal's method and is comparable to results from Kim's learning based method and recently proposed method by Glasner.

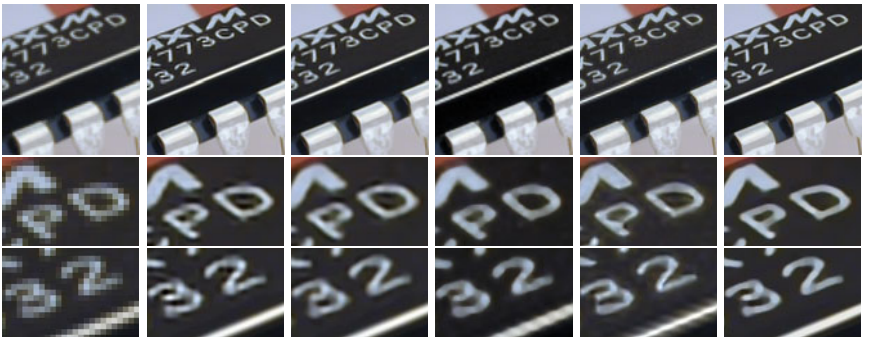

Fig. 5. More results on single-frame SR for real color images (×4). From left to right: NN, GNL-Means [9], KR [6], Fattal's method [20], Glasner's method [10], NL-KR. Note that Fattal's method and Glasner's method generate *jaggy* effects on the edge. Our method is free from the *jaggy* artifacts and preserves better structure.

Table 1. PSNR (Top) and SSIM (Bottom) results for single image super-resolution

Images	NN	BI	GNL-Means [9]	KR [6]	SC [3]	Proposed
Foreman	28.6917	30.9511	31.9961	32.4479	32.5997	**32.7558**
Miss America	31.5765	34.0748	34.4700	34.4419	34.9111	**35.4033**
Suzie	30.0295	31.4660	31.6547	31.8203	31.5208	**32.1033**
Foreman	0.8185	0.8708	0.8747	0.8862	0.8768	**0.8924**
Miss America	0.8403	0.8941	0.9008	0.8990	0.8843	**0.9117**
Suzie	0.7892	0.8286	0.8355	0.8285	0.8334	**0.8449**

Table 2. PSNR (Top) and SSIM (Bottom) results for synthetic test frames

Sequence	NN	BI	GNL-Means [9]	BM3D [22]	Proposed
Foreman	28.8977	30.9493	34.6766	34.9	**35.2041**
Miss America	31.6029	34.0684	36.2508	37.5	**37.8228**
Suzie	30.0307	31.4702	32.9189	33.6	**33.9949**
Foreman	0.8413	0.8709	0.9044	NA	**0.9234**
Miss America	0.8404	0.8928	0.8193	NA	**0.9346**
Suzie	0.7904	0.8290	0.8428	NA	**0.8864**

5.2 Synthetic Experiment for Multi-frame SR

Our second experiment is conducted on synthetic image frames. We generate 9 LR images from one HR image by blurring the HR image with a 3×3 uniform PSF and then decimating the blurred HR image every 3rd row or column with shifts of $\{0, 1, 2\}$ pixels. Gaussian noise with standard deviation of 2 is also added. The PSNR and SSIM results are summarized in Table 2, showing that the proposed method is again constantly better. Note that the results from BM3D is cited directly from [22], which are obtained from noise-free observations.

5.3 Evaluation on Real Video Sequences

Finally, we evaluate the performance of our model on three real image sequences with general motions: Foreman, Miss America and Suzie. Comparisons are made with the GNL-Means [9], BM3D [22], and 3D-KR [16]. The average PSNR and SSIM results on these three test sequences are given in Table 3. As shown, the proposed method achieves better reconstruction accuracy than GNL-Means and BM3D.[3] In Fig. 7, we further show the PSNR results on Foreman and Miss American frame by frame, compared with Bicubic and GNL-Means. The proposed method outperforms GNL-Means method by a notable margin in all frames. The SR results on Miss America and Foreman sequences are given in Fig. 8 and Fig. 6 respectively for visual comparison. Note that GNL-Means sometimes

[3] The PSNR results of 3D-KR are not listed, because they are not numerically available in their original papers (plotted in a figure). However, compared with their figure, our method improves over GNL-Means by a larger margin than the 3D-KR method.

Fig. 6. Video super-resolution for Foreman sequence (frame 8, zoom factor 3, PSNR and SSIM in brackets). Left to right: Ground-truth, BI(30.1758, 0.8739), GNL-Means(33.2233, 0.9041), BM3D(33.45, NA), 3D-KR(33.30, NA), NL-KR(**33.7589, 0.9137**). GNL-Means performs well at regular-structured area but generates severe block artifacts where few similar patches can be found; BM3D suffers from *jagged* effects at edges; 3D-KR can not preserve the straight structure well due to the non-robustness of its spatial-temporal kernel and can generate *ghost image*; our method preserves both the larger structure and fine details well and is free of these artifacts.

Table 3. Average PSNR (Top) and SSIM (Bottom) for the three video sequences

Sequence	NN	BI	GNL-Means[9]	BM3D[22]	Proposed
Foreman	28.8444	31.0539	32.8165	33.5	**34.0141**
Miss America	31.6555	34.2424	35.3453	36.3	**36.4377**
Suzie	30.0846	31.4363	32.9725	33.0	**33.0915**
Foreman	0.8207	0.8720	0.9025	NA	**0.9120**
Miss America	0.8426	0.8938	0.9136	NA	**0.9164**
Suzie	0.7857	0.8233	**0.8797**	NA	0.8671

generates severe block artifacts (see the *Mouth* part in Fig. 6 and *Eye* part in Fig. 8). The 3D-KR method, on the other hand, will generate some *ghost* effects, due to overfitting of the regression and inaccurate estimation of the 3D kernel (see the *Mouth* part in Fig. 6). Furthermore, the 3D-KR method has to employ a motion pre-compensation in order for good 3D kernel estimation, while our model does not require this step.

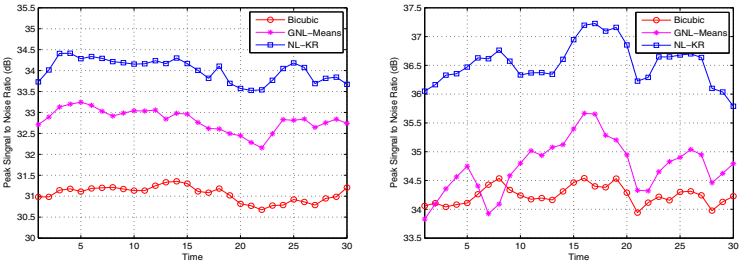

Fig. 7. PSNR plots for Video SR. Left: Foreman and right: Miss America data. The proposed method outperforms other two methods in terms of PSNR in all frames.

Fig. 8. Video super-resolution for Miss America sequence (frame 8, zoom factor 3, PSNR and SSIM in brackets). Left to right: NN(32.4259, 0.8580), BI(34.5272, 0.8958), GNL-Means(34.7635,0.9132), 3D-KR(35.53, NA), NL-KR(**36.6509, 0.9171**). The GNL-Means method suffers from block effects, while our method is free from artifacts and is comparable to 3D-KR in this case.

6 Conclusions and Future Work

This paper proposes a Non-Local Kernel Regression (NL-KR) model for image and video restoration tasks, which combines the local structural regularity as well as non-local similarity explicitly to ensure a more reliable and robust estimation. The proposed method is a general model that includes many related models as special cases. In this work, we focus on the image and video super-resolution task, and experiments on both single frame and video sequence demonstrate the effectiveness and robustness of our model. Further more, the NL-KR on single image super-resolution may suggest that the image itself contains enough information that SR without training and motion is possible. Also, incorporating more self-similarity information by extending the image into multi-scale space is straightforward under our model. In the current algorithm, the patch matching and spatial kernel calculation are most computationally heavy, which can be speeded up by KD-tree searching and parallel computing respectively. The proposed model can also be applied to other image and video restoration tasks, e.g. inpainting and denoising, and we leave those to be our future work.

Acknowledgement. The authors would like to thank all the reviewers for their valuable comments. Haichao Zhang would like to thank the Chinese Government for supporting his PhD Study. This work is supported in part by NSF of China (No.60872145) and Cultivation Fund from Ministry of Education of China (No.708085). The work is also supported in part by the U.S. Army Research Laboratory and U.S. Army Research Office under grand number W911NF-09-1-0383.

References

1. Tschumperle, D.: PDE's Based Regularization of Multivalued Images and Applications. PhD thesis (2002)
2. Elad, M., Aharon, M.: Image denoising via learned dictionaries and sparse representation. In: CVPR, pp. 17–22 (2006)

3. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution as sparse representation of raw image patches. In: CVPR (2008)
4. Chang, H., Yeung, D.Y., Xiong, Y.: Super-resolution through neighbor embedding. In: CVPR (2004)
5. Li, X.: Video processing via implicit and mixture motion models. IEEE Trans. on Circuits and Systems for Video Technology 17, 953–963 (2007)
6. Takeda, H., Farsiu, S., Milanfar, P.: Kernel regression for image processing and reconstruction. IEEE TIP 16, 349–366 (2007)
7. Tomasi, C.: Bilateral filtering for gray and color images. In: ICCV, pp. 839–846 (1998)
8. Buades, A., Coll, B.: A non-local algorithm for image denoising. In: CVPR (2005)
9. Protter, M., Elad, M., Takeda, H., Milanfar, P.: Generalizing the non-local-means to super-resolution reconstruction. IEEE TIP, 36–51 (2009)
10. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: ICCV (2009)
11. Peyre, G., Bougleux, S., Cohen, L.: Non-local regularization of inverse problems. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 57–68. Springer, Heidelberg (2008)
12. Efros, A., Leung, T.: Texture synthesis by non-parametric sampling. In: ICCV, pp. 1033–1038 (1999)
13. Wong, A., Orchard, J.: A nonlocal-means approach to exemplar-based inpainting. In: ICIP (2002)
14. Protter, M., Elad, M.: Super resolution with probabilistic motion estimation. IEEE TIP, 1899–1904 (2009)
15. Chatterjee, P., Milanfar, P.: A generalization of non-local means via kernel regression. In: SPIE Conf. on Computational Imaging (2008)
16. Takeda, H., Milanfar, P., Protter, M., Elad, M.: Super-resolution without explicit subpixel motion estimation. IEEE TIP (2009)
17. Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: ICCV (2009)
18. Getreuer, P.: (2009), `http://www.math.ucla.edu/~getreuer/tvreg.html`
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Member, S., Simoncelli, E.P., Member, S.: Image quality assessment: From error visibility to structural similarity. IEEE TIP 13, 600–612 (2004)
20. Fattal, R.: Image upsampling via imposed edge statistics. In: SIGGRAPH (2007)
21. Kim, K.I., Kwon, Y.: Example-based learning for single-image super-resolution and jpeg artifact removal. Technical report (2008)
22. Danielyan, A., Foi, A., Katkovnik, V., Egiazarian, K.: Image and video super-resolution via spatially adaptive block-matching filtering. In: Int. Workshop on Local and Non-Local Approx. in Image Process (2008)