A Generalization of Non-Local Means via Kernel Regression

Priyam Chatterjee and Peyman Milanfar

Department of Electrical Engineering,
University of California, Santa Cruz, CA 95064, USA.

ABSTRACT

The Non-Local Means (NLM) method of denoising has received considerable attention in the image processing community due to its performance, despite its simplicity. In this paper, we show that NLM is a zero-th order kernel regression method, with a very specific choice of kernel. As such, it can be generalized. The original method of NLM, we show, implicitly assumes local constancy of the underlying image data. Once put in the context of kernel regression, we extend the existing Non-Local Means algorithm to higher orders of regression which allows us to approximate the image data locally by a polynomial or other localized basis of a given order. These extra degrees of freedom allow us to perform better denoising in texture regions. Overall the higher order method displays consistently better denoising capabilities compared to the zero-th order method. The power of the higher order method is amply illustrated with the help of various denoising experiments.

1. INTRODUCTION

Image denoising is perhaps the simplest of image processing problems that researchers have studied over a long time. In the recent years various new techniques have been proposed which perform very good denoising even in the presence of large amounts of noise. Yet this linear problem remains a wide open field of research as we see progress in image capturing sensor technology. The increase in the number of pixels in sensors typically translates to smaller pixels which consequently gives rise to an increase in the perceived noise in the captured image due to the availability of fewer photons. This makes image denoising still a very much relevant problem necessitating continuing research.

For image denoising we assume that the original scene, when captured by a camera, is degraded only by the presence of noise introduced by the image acquisition process. The forward model for each pixel in the image thus can be mathematically written as

$$y_i = z(\mathbf{x}_i) + \eta_i \tag{1}$$

where $z(.)$ denotes a local function describing the true image which is corrupted by additive noise η_i resulting in the observed intensity y_i at location $\mathbf{x}_i = [x_1\ x_2]_i^T$. The problem of denoising then translates to obtaining an estimate $\hat{z}(.)$ of the intensity function at each pixel position. The recently proposed Non-Local Means[1] algorithm assumes that the noise is zero mean and uncorrelated across locations. Such an assumption has also been made by other spatial domain filtering methods for denoising.[2–5] Many filtering methods perform anisotropic filtering by a weighted averaging process in a small neighborhood. The *bilateral filter*[3] is one such neighborhood filter which performs well under reasonably small amounts of noise. However, the reliance on intensity matching of pixels makes the method much less useful in cases where noise is strong. Recently Takeda *et al.*[5] proposed a method of image restoration which performs denoising well even when the image is corrupted by large amounts of noise. The tolerance to noise of their method results from an iterative process where the weights are calculated in a robust manner. The authors model the problem of denoising in a kernel regression framework and also extend the standard *bilateral filter* to good effect under this framework in a related work.[4] In this paper we study the Non-Local Means method and show how it can be generalized for better denoising under the kernel regression model.

Further author information: (Send correspondence to Priyam Chatterjee)
P. Chatterjee : E-mail: priyam@soe.ucsc.edu, Web: http://www.soe.ucsc.edu/~priyam
P. Milanfar : E-mail: milanfar@soe.ucsc.edu, Web: http://www.soe.ucsc.edu/~milanfar

Figure 1. An image showing similar and dissimilar patches. The patch under consideration is the shaded red patch. Other red patches are similar and hence contribute more to denoising of the pixel at the center of the shaded patch. The blue patches are totally dissimilar and hence are ignored. The orange patches bear some resemblance but are not totally similar and hence their contributing weights are lesser than the red patches.

2. NON-LOCAL MEANS

Non-Local Means (NLM) is a method of denoising that was introduced recently by Buades *et al.*[1] and has become quite popular. The method was further enhanced for speed in subsequent works by Mahmoudi *et al.*[6] and Bilcu *et al.*[7] The difference of this method from previous adaptive spatial domain filtering methods is that the theory behind this method does not require a locality constraint. The authors make use of similar patterns occurring in different parts of any image and use these similarities to denoise an image assuming that the corrupting noise is uncorrelated. Figure (1) explains the concept of patch similarities that exist throughout natural images.

The NLM algorithm is a weighted averaging process where similarity of regions determines the amount of contribution of a particular region to the denoising mechanism for a region of interest. Mathematically this process can be written down as

$$\hat{z}(\mathbf{x}_i) = \frac{1}{C_i} \sum_{j \in I} w_{ij} y_i \quad \text{where} \quad C_i = \sum_j w_{ij}. \tag{2}$$

Here C_i acts as a normalizing constant. The index j runs over the entire image I which essentially denotes that for every pixel the weights are calculated with respect to all other pixels in the image. The weights are calculated by matching neighborhoods (patches) and searching for pixels which have similar neighborhoods to the one under consideration. This is done by calculating weights for each pair of pixels by the formula

$$w_{ij} = e^{-\frac{\|\mathcal{N}(\mathbf{x}_i) - \mathcal{N}(\mathbf{x}_j)\|_a^2}{h^2}} \tag{3}$$

where $\mathcal{N}(\mathbf{x}_i)$ is the vectorized version of a patch centered at location $\mathbf{x}_i$ and h is a global smoothing parameter which controls the amount of blurring introduced in the denoising process. For higher values of noise present in an image this constant is set to be larger. The parameter a is the variance of a normal distribution which is used to smooth out (i.e. give decaying weight to pixels away from the center of) the neighborhood while calculating the weights. This has the effect of reducing the influence of difference in pixel intensities as they go spatially further away from the center of the patch being considered.

The NLM process thus gives rise to a method of weighted average filtering which calculates the contributing weights making use of Equation (3). This weight calculation mechanism effectively sorts out dissimilar patches and does not consider them (or gives them significantly lower weight) in the denoising process. The final restoration process consists of using these weights to determine the ratio in which the central pixels of the patches contribute to estimate the denoised intensity of the central pixel of the patch under consideration, as expressed by Equation (2). It can be shown that the noise free estimate for each pixel is found such that it solves an optimization problem

$$\begin{aligned}
\hat{z}(\mathbf{x}_i) &= \arg\min_{z(\mathbf{x}_i)} \sum_{j=1}^{N} w_{ij}(y_j - z(\mathbf{x}_i))^2 \\
&= \arg\min_{z(\mathbf{x}_i)} \ (\mathbf{y} - \mathbf{1}z(\mathbf{x}_i))^T \mathbf{W}_i (\mathbf{y} - \mathbf{1}z(\mathbf{x}_i)) \quad \text{where} \quad \mathbf{W}_i = \text{diag}[w_{i1} \ldots w_{ij} \ldots w_{iN}] \\
&= \arg\min_{z(\mathbf{x}_i)} \ \|\mathbf{y} - \mathbf{1}z(\mathbf{x}_i)\|^2_{\mathbf{W}_i} \qquad (4)
\end{aligned}$$

where $\mathbf{y}$ denotes the lexicographically ordered vectorized version of the image and $\mathbf{1}$ denotes a vector of all ones. The method, though simple, takes a lot of time to calculate the weights since the weights need to be calculated for each and every pair of pixels. To speed the process up, the authors introduce a much smaller search window within which the weights are calculated. Taking such a speedup into account, we can rewrite Equation (4) as

$$\begin{aligned}
\hat{z}(\mathbf{x}_i) &= \arg\min_{z(\mathbf{x}_i)} \sum_{j=1}^{P} w_{ij}(y_j - z(\mathbf{x}_i))^2 \\
&= \arg\min_{z(\mathbf{x}_i)} \|\mathbf{y}_i - \mathbf{1}z(\mathbf{x}_i)\|^2_{\mathbf{W}_i} \qquad (5)
\end{aligned}$$

where now $\mathbf{y}_i$ stands for the lexicographically ordered vectorized version of the P pixels within the search window centered at location $\mathbf{x}_i$. The weights are likewise calculated within this search window only making $\mathbf{W}_i = \text{diag}[w_{i1} \ldots w_{ij} \ldots w_{iP}]$. While this strictly does not make the method non-local in nature, with a large enough search window the method works nearly as well. The ability to write the problem of denoising as approached by the NLM method in this way brings out an interesting aspect of the method. From the cost function in Equation (5) it can be seen that the method tends to find a single constant at each pixel location which minimizes the error in reconstruction of the patch under consideration. The weights detect similar patches and ensure that the cost function penalizes only the error with respect to pixels with similar neighborhoods. The fact that the single constant $z(\mathbf{x}_i)$ is multiplied by a constant vector of all ones and compared to the search window in calculating the cost in Equation (5) results in an implicit assumption of local constancy of the image region. In a global sense the method of Non-Local Means assumes that an image can be modelled to be locally constant. Such an assumption of local constancy particularly does not hold in regions of finer detail and results in a denoised image lacking much of the original texture.

3. HIGHER ORDER NON-LOCAL MEANS

In this section we extend the Non-Local Means method of denoising to overcome the inherent assumption of local constancy. Assuming that the underlying image data is locally sufficiently smooth, we can write a Taylor expansion of the form

$$\begin{aligned}
z(\mathbf{x}_j) &= z(\mathbf{x}_i) + \{\nabla z(\mathbf{x}_i)\}^T (\mathbf{x}_j - \mathbf{x}_i) + \frac{1}{2}(\mathbf{x}_j - \mathbf{x}_i)^T \{\mathcal{H}z(\mathbf{x}_i)\} (\mathbf{x}_j - \mathbf{x}_i) + \ldots \\
&= \beta_{\{0\}i} + \beta^T_{\{1\}i}(\mathbf{x}_j - \mathbf{x}_i) + \beta^T_{\{2\}i}\text{vech}\{(\mathbf{x}_j - \mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)^T\} + \ldots \qquad (6)
\end{aligned}$$

where $\nabla(.)$ denotes the two dimensional gradient vector, $\mathcal{H}$ denotes the Hessian and vech(.) denotes the lexicographically ordered upper triangular part of a symmetric matrix, written out as a vector. Such an expansion can be written with respect to all the points in the search windows centered at any given pixel, though naturally, as one moves away from the center of the window the fidelity of this expansion becomes diminished. This latter issue is addressed later in the choice of the kernel weights. Restricting the expansion to some order, say M, we thus may naturally want to estimate the denoised image intensity as that which best fits the local expansion. This can be done by minimizing the cost function with respect to the unknown β_i parameters

$$\begin{aligned}\hat{\beta}_i &= \arg\min_{\beta_i} \sum_{j=1}^{P} (y_j - \beta_{\{0\}i} - \beta_{\{1\}i}^T(\mathbf{x}_j - \mathbf{x}_i) - \beta_{\{2\}i}^T \text{vech}\{(\mathbf{x}_j - \mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)^T\} + \dots)^2 \\ &= \arg\min_{\beta_i} \|\mathbf{y}_i - \mathbf{\Phi}\beta_i\|^2 \end{aligned} \tag{7}$$

$$\text{where } \mathbf{\Phi} = \begin{bmatrix} 1 & (\mathbf{x}_1 - \mathbf{x}_i)^T & \text{vech}^T\{(\mathbf{x}_1 - \mathbf{x}_i)(\mathbf{x}_1 - \mathbf{x}_i)^T\} & \cdots \\ 1 & (\mathbf{x}_2 - \mathbf{x}_i)^T & \text{vech}^T\{(\mathbf{x}_2 - \mathbf{x}_i)(\mathbf{x}_2 - \mathbf{x}_i)^T\} & \cdots \\ \vdots & \vdots & \vdots & \\ 1 & (\mathbf{x}_P - \mathbf{x}_i)^T & \text{vech}^T\{(\mathbf{x}_P - \mathbf{x}_i)(\mathbf{x}_P - \mathbf{x}_i)^T\} & \cdots \end{bmatrix}, \ \beta_i = \begin{bmatrix} \beta_{\{0\}i} \\ \beta_{\{1\}i} \\ \beta_{\{2\}i} \\ \vdots \end{bmatrix}.$$

β_i is a vector and the matrix $\mathbf{\Phi}$ is formed from polynomial basis vectors. Once we find the β_i parameters, we simply retain the $\hat{\beta}_{\{0\}i}$ entry as the estimated intensity at location $\mathbf{x}_i$. The justification of doing so comes readily from the fact that once we estimate the unknown $\beta_{\{n\}i}$ parameters, we can reconstruct the search window centered at $\mathbf{x}_i$ under the local polynomial data model assumption. The intensity of the pixel under consideration is the intensity of the central pixel of the search window which can be mathematically shown as

$$\hat{z}(\mathbf{x}_i) = \mathbf{c}^T \mathbf{\Phi} \hat{\beta}_i = \hat{\beta}_{\{0\}i} \tag{8}$$

where $\mathbf{c} = [0 \ \dots \ 1 \ \dots \ 0]^T$ is a vector of all zeros except for a one at the center. However, in the above optimization, we wish to give lower weights to dissimilar neighborhoods. This can be achieved by making use of the weights calculated using Equation (3) to determine the confidence in the local expansion with respect to each pixel in the search window. Thus, the problem of optimization can be framed as

$$\begin{aligned}\hat{\beta}_i &= \arg\min_{\beta_i} \sum_{j=1}^{P} w_{ij}(y_j - \beta_{\{0\}i} - \beta_{\{1\}i}^T(\mathbf{x}_j - \mathbf{x}_i) - \beta_{\{2\}i}^T \text{vech}\{(\mathbf{x}_j - \mathbf{x}_i)(\mathbf{x}_j - \mathbf{x}_i)^T\} + \dots)^2 \\ &= \arg\min_{\beta_i} \|\mathbf{y}_i - \mathbf{\Phi}\beta_i\|^2_{\mathbf{W}_i}. \end{aligned} \tag{9}$$

This optimization problem is simply a weighted least squares estimation problem and can be shown to have a closed form solution as

$$\hat{\beta}_i = \left(\mathbf{\Phi}^T \mathbf{W}_i \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{W}_i \mathbf{y}_i = \mathbf{K}_i \mathbf{y}_i. \tag{10}$$

We define the matrix $\mathbf{K}_i = \left(\mathbf{\Phi}^T \mathbf{W}_i \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^T \mathbf{W}_i$ to be the *equivalent kernel* with which we perform the weighted averaging process. The estimated denoised intensity at position $\mathbf{x}_i$ is chosen to be $\hat{z}(\mathbf{x}_i) = \hat{\beta}_{\{0\}i}$ which is simply the first element of $\hat{\beta}_i$ calculated from Equation (10).

It can be easily seen that the NLM method of Buades *et. al*[1] can be derived as a special case of the above denoising formulation. In the zero-th order case of the regression problem of Equation (9), the matrix $\mathbf{\Phi}$ simplifies to a column vector of all ones and we obtain the formulation of Equation (4). When this $\mathbf{\Phi}$ is plugged into Equation (10) we obtain the NLM method.

$$\begin{aligned}\hat{\beta}_i &= \left(\mathbf{1}^T \mathbf{W}_i \mathbf{1}\right)^{-1} \mathbf{1}^T \mathbf{W}_i \mathbf{y}_i \\ &= \frac{\mathbf{1}^T \mathbf{W}_i \mathbf{y}_i}{\mathbf{1}^T \mathbf{W}_i \mathbf{1}} \\ &= \frac{\sum_{j=1}^{P} w_{ij} y_j}{\sum_{j=1}^{P} w_{ij}}. \end{aligned} \tag{11}$$

The method of NLM can thus be generalized to any arbitrary order by determining the number of terms we decide to retain in the Taylor expansion. In this paper, we work with higher orders of 1 and 2. It may be mentioned here that Buades *et al.*[8] do mention a first order "regression correction" to obtain a better reconstruction but

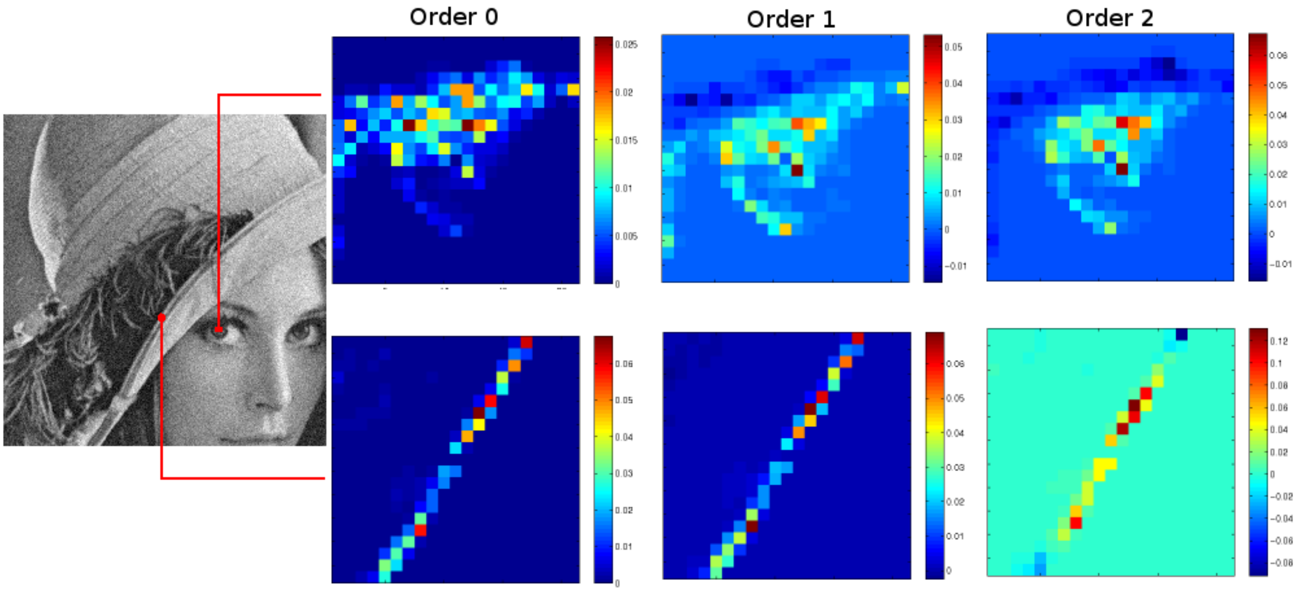


Figure 2. Lena image showing the equivalent kernels of orders 0, 1 and 2 at two different regions. Note how the equivalent kernel of 2^{nd} order adapts better to the underlying image data. We also see that the higher order equivalent weights can go negative at some points.

we provide a more general framework to extend the method to arbitrary orders, as dictated by the underlying image data, and illustrate that in many cases, regression of order 2 in fact results in superior performance than first order regression. Though it should be said that the choice of the order of regression is itself an interesting topic to be analyzed in future research, on balance, our approach allows for greater flexibility in modelling the data and results in better denoising as a whole. In Figure (2) we show the equivalent kernels resulting out of the NLM method of different orders when processing the Lena image.

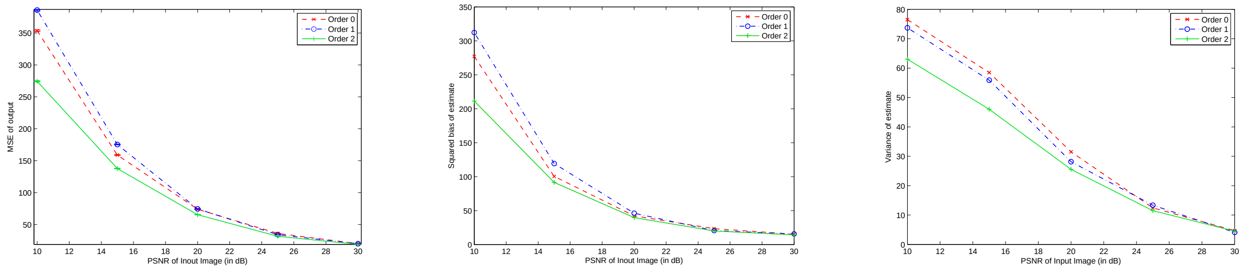
It is worth noting that the equivalent kernel weights in the zeroth order (standard NLM) case are by definition always positive, which illustrates a basic shortcoming. By contrast, higher order NLM as we have described above allows for the coefficients of the weighted sum in Equation (10) to have negative values as well. This advantage accounts for the superior performance of the higher order methods.

4. RESULTS

We performed our experiments using the higher order NLM method using images artificially corrupted by noise as well as those corrupted by noise introduced in the image capturing process. In the artificial noise case we perform denoising on the Lena image with differing amounts of zero mean white Gaussian noise. Since the original noiseless image is available to us, we measure the performance of the denoising results by comparing their mean squared error (MSE). The global smoothing parameter for each of the cases was tuned to obtain the best results, in terms of MSE, for the different orders. In Figure (3) we show the denoising results obtained on the Lena image where the input images had a peak signal to noise ratio (PSNR)* of 30, 20 and 10 dB. We see that in each of these cases the 2^{nd} order displays a better denoised output in terms of MSE. For each input PSNR, 20 different experiments were carried out where the image was degraded by different realizations of white Gaussian noise. Such experiments were carried out for NLM of orders 0, 1 and 2. The graph in Figure (4) shows the average MSE results for each method as a function of the noise level. The superiority of the higher order version of NLM is clearly shown in the graph.

*Peak Signal to Noise Ratio $= 10\log_{10}\left(\frac{255^2}{\text{Mean Squared Error}}\right)$ [dB]

Figure 3. Results of denoising using NLM method of orders 0, 1 & 2. The MSE values reported are the average MSE of the denoised results obtained over simulations with 20 different realizations of zero mean Gaussian noise in each PSNR case.



(a) MSE plot **(b) Squared bias plot** **(c) Variance plot**

Figure 4. Graph showing results from a simulation of denoising the Lena image with the input image having different levels of zero mean Gaussian noise. For each PSNR, the experiments were conducted 20 times with a different realization of the noise. (a) The mean MSE value is plotted for each PSNR along with the standard deviation of the results about the mean as error bars, (b) the squared bias plot for the estimates, and (c) the variance plot for the estimates.

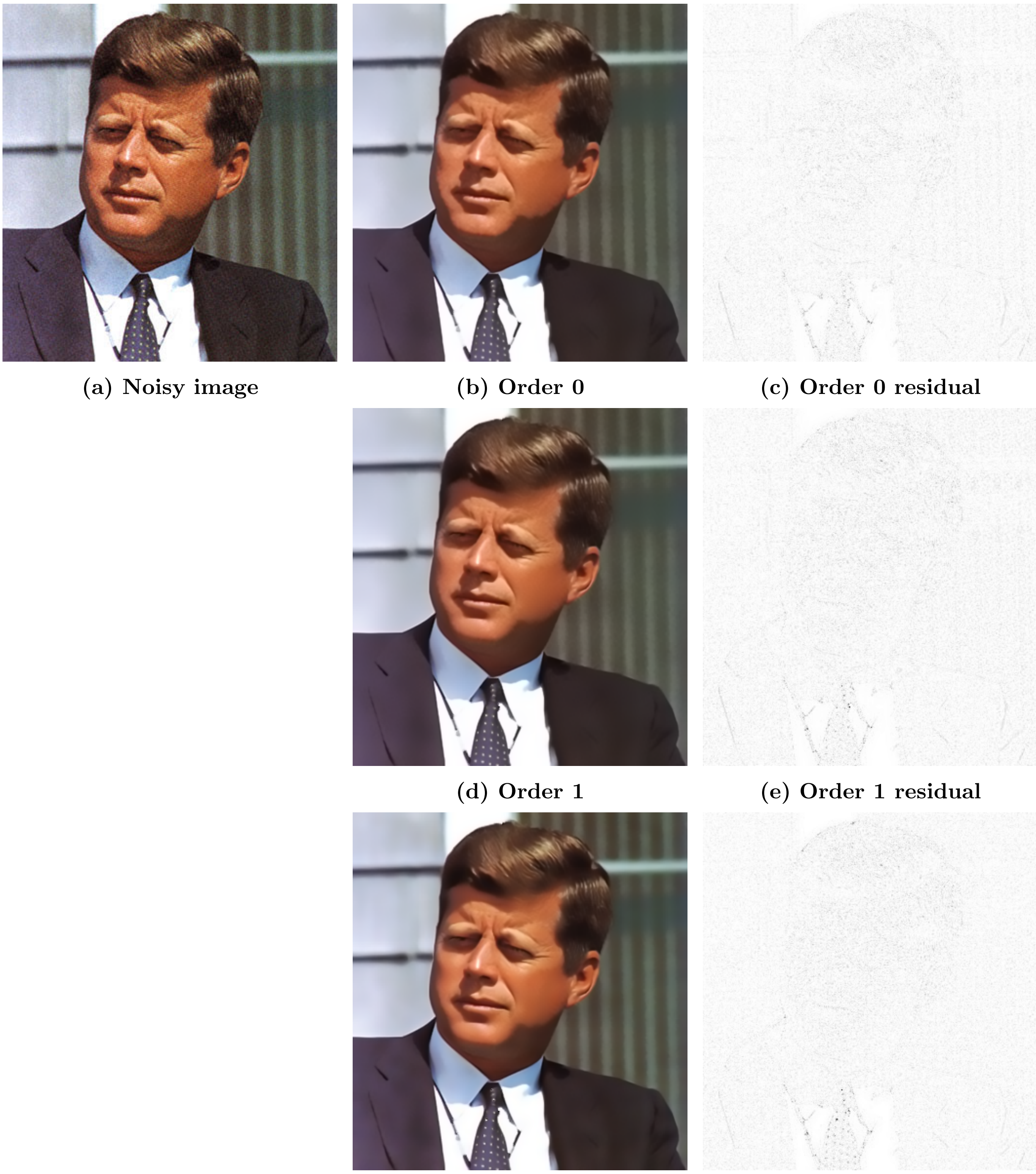
We also conducted experiments on real data corrupted by noise introduced in the image acquisition stage. Here the noise statistics are unknown to us. Even though the noise need not necessarily be zero mean or Gaussian in nature in these cases, we see that the NLM method which makes such an assumption is still able to perform considerable denoising. We fix the value of the global smoothing parameter h by making a guess on the variance of the noise present. Since most images captured by digital cameras are color images, we apply the method to these pictures by separately denoising the image in each of the YCbCr channels. While processing the images in other color coordinates and taking into account the inter-channel dependencies could possibly result in a better reconstruction, we consider it to be outside the scope of this paper. We perform denoising on different color images using the NLM algorithm of different orders and present the results in Figures (5) & (6). The global smoothing parameter h was chosen to achieve the visually best denoising results. Since the original denoised image is not available to us there is no method of performing a quantitative analysis on the performance of the methods. However, we show the residual image of the luminance channel after denoising which gives an indication of the noise removal capabilities of the method. A good denoising algorithm should ideally be able to distinguish between areas of texture and noise and perform denoising accordingly. In the results, it is apparent that the higher order methods allow for the retention of more image structure, specially in the finer textured areas.

5. CONCLUSION

In this paper we studied the Non-Local Means algorithm for denoising and identify an inherent limitation in the method. We bring out the implicit assumption of local constancy in the NLM method. Casting the problem as one of kernel regression, we are able to generalize the method to arbitrary orders, allowing us greater flexibility in modelling the underlying image data. We pose the generalized NLM algorithm as an optimization problem which produces a closed form weighted least squares solution. We see, through experiments, that the generalized higher order method can out-perform the standard zero-th order NLM algorithm. This improvement is better noticed in cases where the corrupting noise is strong. The NLM method also assumes that the corrupting noise is uncorrelated in nature. However, we show that the method performs well in natural color images corrupted by noise (e.g film grain). The generalized NLM algorithm displays improvements even in these cases which is seen by presence of lesser structure in the residual of the luminance channel. Overall we find the generalization to result in better denoising, both visually and in terms of MSE when such a quantitative analysis is possible.

REFERENCES

1. A. Buades, B. Coll, and J.-M. Morel, "A Non-Local Algorithm for Image Denoising," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, **2**, pp. 60–65, IEEE Computer Society, (Washington, DC, USA), October 2005.

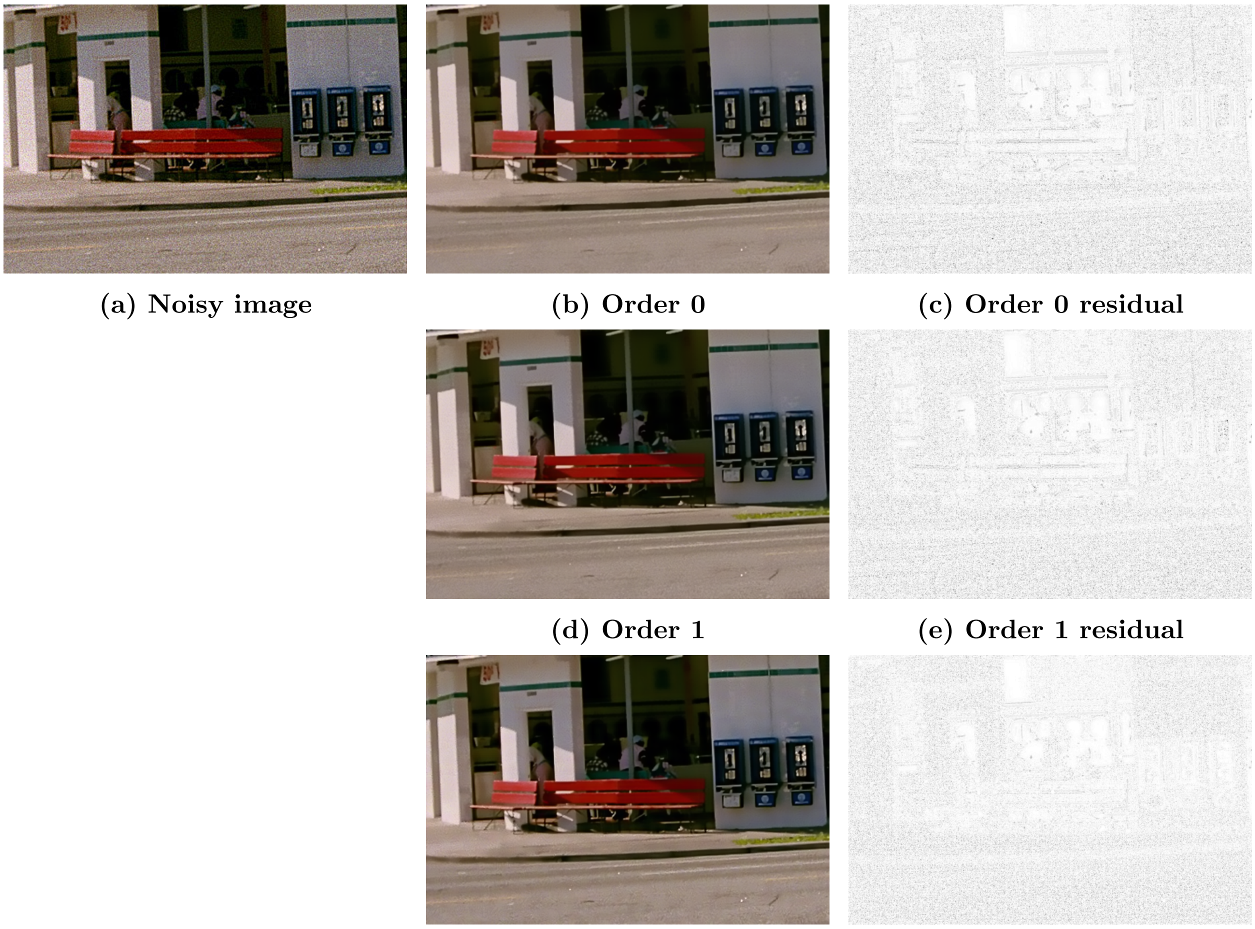


(a) Noisy image (b) Order 0 (c) Order 0 residual

(d) Order 1 (e) Order 1 residual

(f) Order 2 (g) Order 2 residual

Figure 5. Denoising a real noisy image where the noise statistics are unknown by NLM of different orders. The standard NLM (order 0) produces a more cartoonish result whereas the higher order restoration seems more natural. Also note how the higher order (order 2) residual displays less structure in it.



(a) Noisy image **(b) Order 0** **(c) Order 0 residual**

(d) Order 1 **(e) Order 1 residual**

(f) Order 2 **(g) Order 2 residual**

Figure 6. Denoising a real noisy image where the noise statistics are unknown by NLM of different orders. The higher order method produces a truer reconstruction which is noticeable in the residual which displays less image structure.

2. S. M. Smith and J. M. Brady, "SUSAN - A New Approach to Low Level Image Processing," *International Journal of Computer Vision* **23**(1), pp. 45–78, 1997.
3. C. Tomasi and R. Manduchi, "Bilateral Filtering for Gray and Color Images," *Proceedings of the Sixth International Conference on Computer Vision* , pp. 839–846, IEEE Computer Society, (Washington, DC, USA), January 1998.
4. H. Takeda, S. Farsiu, and P. Milanfar, "Higher Order Bilateral Filters and their properties," in *Proceedings of the SPIE Conference on Computational Imaging V*, **6498**, p. 64980S, February 2007.
5. H. Takeda, S. Farsiu, and P. Milanfar, "Kernel Regression for Image Processing and Reconstruction," *IEEE Transactions on Image Processing* **16**(2), pp. 349–66, 2007.
6. M. Mahmoudi and G. Sapiro, "Fast Image and Video Denoising via Nonlocal Means of Similar Neighborhoods," *IEEE Signal Processing Letters* **12**, pp. 839–842, December 2005.
7. R. C. Bilcu and M. Vehvilainen, "Fast Nonlocal Means for Image Denoising," in *Proceedings of the SPIE Conference on Digital Photography III*, **6502**, p. 65020R, March 2007.
8. A. Buades, B. Coll, and J.-M. Morel, "The Staircasing Effect in Neighborhood Filters and its Solution," *IEEE Transactions on Image Processing* **15**(6), pp. 1499–1505, 2006.