Patch-wise Ideal Stopping Time for Anisotropic Diffusion

Hossein Talebi, and Peyman Milanfar*

Department of Electrical Engineering
University of California, Santa Cruz

ABSTRACT

Data-dependent filtering methods are powerful techniques for image denoising. Beginning with any base procedure (nonlinear filter), repeated applications of the same process can be interpreted as a discrete version of anisotropic diffusion. As such, a natural question is "What is the best stopping time in iterative data-dependent filtering?" This is the general question we address in this paper. To develop our new method, we estimate the mean-squared-error (MSE) in each image patch. This estimate is used to characterize the effectiveness of the iterative filtering process, and its minimization yields the ideal stopping time for the diffusion process.

Keywords: Image denoising, Anisotropic diffusion, Nonlinear filter, Stopping time criterion

1. INTRODUCTION

In the past few years, non-parametric restoration methods have become extremely popular. These new algorithms are mostly patch-wise, and also employ local and non-local similarities in the signals.[1–3] Perhaps the most well known algorithm in this class is the bilateral filter,[2] which smooths images by means of a nonlinear combination of nearby image values. The method combines pixel values based on both their geometric closeness and their photometric similarity. The Non-Local Means (NLM)[3] is another very popular data-dependent filter which closely resembles the bilateral filter except that the photometric similarity is captured in a patch-wise way. More recently, the LARK kernel[1] exploits the geodesic distance based on estimated gradients. In general, all of these restoration algorithms work based on the same framework in which some data-adaptive weights are assigned to each pixel contributing to the filtering. The measurement model for the denoising problem is defined as:

$$\mathbf{y} = \mathbf{z} + \mathbf{e} \tag{1}$$

where $\mathbf{z}$, $\mathbf{e}$ and $\mathbf{y}$ are column vectors and respectively denote the underlying signal, noise and noisy signal. The noise $\mathbf{e}$ is zero mean, white, and uncorrelated with the signal $\mathbf{z}$. Eq. 2 shows the matrix-vector multiplication form of the denoising filter where $\widehat{\mathbf{z}}$ and $\mathbf{W}$ denote the filtered signal and the matrix of filter weights respectively. The square matrix $\mathbf{W}$ is in general a function of the given data vector[4] $\mathbf{y}$. This filter is given by the weights defined by any one of the data-dependent methods discussed above

$$\widehat{\mathbf{z}} = \mathbf{W}\mathbf{y}. \tag{2}$$

In general, $\mathbf{W}$ is a positive definite, row-stochastic matrix (each row sums to one.) Although this matrix is not symmetric, it has been shown that it can be very closely approximated with a symmetric matrix.[5] The spectrum of $\mathbf{W}$ specifies the effect of the filter on the noisy signal. Considering the symmetric $\mathbf{W}$ matrix, its eigen-decomposition is:

$$\mathbf{W} = \mathbf{V}\mathbf{S}\mathbf{V}^T \tag{3}$$

where $\mathbf{S} = \mathrm{diag}[\lambda_1, ..., \lambda_n]$ contains the eigenvalues in decreasing order $0 < \lambda_n \leq ... < \lambda_1 = 1$, and V is an orthogonal matrix containing the eigenvectors of W in its columns. It is possible to improve the performance of

Further author information:
htalebi@soe.ucsc.edu, milanfar@ucsc.edu
*This work was supported by AFOSR Grant FA9550-07-1-0365 and NSF Grant CCF-1016018.

these filters by applying them multiple time. That is, $\widehat{\mathbf{z}}_k = \mathbf{W}^k\mathbf{y}$. In this framework, each application of $\mathbf{W}$ can be thought of as one step of anisotropic diffusion with the kernel[4,6] defined by $\mathbf{W}$. Diffusion filtering gradually removes noise in each iteration, but also takes away latent details from the underlying signal. Choosing a small iteration number k preserves the underlying structure, but also does little denoising. On the other hand, a large k tends to over-smooth and remove noise and high frequency details at the same time. The question we address here is when to stop the diffusion process so as to get the ideal denoising result. In other words, we look for the optimal iteration number which keeps the balance between noise removal and detail preservation.

Two existing stopping time criteria are detailed in Refs. 5,8. Sporring and Weickert[7] showed that the behavior of generalized entropies over time can provide a clue as to the optimal stopping. Monotonicity and smoothness of the entropy indicates the earliest time when the iteration should be stopped because the output signal entropy has reached its steady state. Mrázek and Navara[8] developed a time-selection strategy for diffusion. In their proposed method, the stopping time is chosen so that the correlation of the filtered signal and noise is minimized. While these stopping criteria are applicable to anisotropic diffusion, there is a need to compute the diffusion scale for patch-wise methods. In the next section we show that the second method does not work in patch-based denoising, and propose an effective alternative.

2. EXISTING STOPPING TIME STRATEGIES

Mrázek and Navara[8] developed a time-selection strategy for where the stopping time is chosen so that the correlation of signal and noise in the filtered image is minimized. More specifically, they first define the residual after k iterations as:

$$\mathbf{r}_k = \mathbf{y} - \widehat{\mathbf{z}}_k = (\mathbf{I} - \mathbf{W}^k)\mathbf{y}. \tag{4}$$

To find the optimal stopping time they propose to minimize the normalized correlation between $\mathbf{r}_k$ and $\widehat{\mathbf{z}}_k$ in each iteration:

$$\rho(\mathbf{r}_k, \widehat{\mathbf{z}}_k) = \frac{\operatorname{tr}\left(\operatorname{cov}(\mathbf{r}_k, \widehat{\mathbf{z}}_k)\right)}{\sqrt{\operatorname{tr}\left(\operatorname{cov}(\mathbf{r}_k)\right)\operatorname{tr}\left(\operatorname{cov}(\widehat{\mathbf{z}}_k)\right)}} \tag{5}$$

where the covariance of two vectors $\mathbf{r}_k$ and $\widehat{\mathbf{z}}_k$ is given by

$$\operatorname{cov}(\mathbf{r}_k, \widehat{\mathbf{z}}_k) = \mathrm{E}[(\mathbf{r}_k - \bar{\mathbf{r}}_k)(\widehat{\mathbf{z}}_k - \bar{\mathbf{z}}_k)^T] \tag{6}$$

in which $\bar{\mathbf{r}}_k$ and $\bar{\mathbf{z}}_k$ are expected values of $\mathbf{r}_k$ and $\widehat{\mathbf{z}}_k$.[9] Without additional assumptions on the noise, the signal and the filter, $\rho(\mathbf{r}_k, \widehat{\mathbf{z}}_k)$ is not guaranteed to be unimodal and possess a unique single minimum.[8] Let us go back to (5) and rewrite the criterion. For a deterministic signal $\mathbf{z}$ we have:

$$\bar{\mathbf{r}}_k = \mathrm{E}[(\mathbf{I} - \mathbf{W}^k)\mathbf{y}] \approx (\mathbf{I} - \mathbf{W}^k)\mathbf{z} \tag{7}$$

$$\bar{\mathbf{z}}_k = \mathrm{E}[\mathbf{W}^k\mathbf{y}] \approx \mathbf{W}^k\mathbf{z} \tag{8}$$

We also can rewrite $\mathrm{E}[\mathbf{r}_k\widehat{\mathbf{z}}_k^T]$ as signal and noise components

$$\mathrm{E}[\mathbf{r}_k\widehat{\mathbf{z}}_k^T] = (\mathbf{I} - \mathbf{W}^k)(\mathrm{E}[\mathbf{z}\mathbf{z}^T] + \sigma^2\mathbf{I})\mathbf{W}^k \tag{9}$$

where σ^2 is the noise variance. Overall, considering (7),(8) and (9), and after some simplifications the covariance is given:

$$\operatorname{cov}(\mathbf{r}_k, \widehat{\mathbf{z}}_k) = (\mathbf{I} - \mathbf{W}^k)\mathbf{W}^k\sigma^2\mathbf{I} \tag{10}$$

Doing the same for the other terms, we can express them in terms of $\mathbf{W}$, $\mathbf{z}$ and σ.

$$\operatorname{tr}\left(\operatorname{cov}(\mathbf{r}_k)\right) = \operatorname{tr}\left((\mathbf{I} - \mathbf{W}^k)^2\sigma^2\mathbf{I}\right) \tag{11}$$

$$\mathrm{tr}\left(\mathrm{cov}(\widehat{\mathbf{z}}_k)\right) = \mathrm{tr}(\mathbf{W}^{2k}\sigma^2\mathbf{I}) \tag{12}$$

Overall, the correlation criterion can be written as:

$$\rho(\mathbf{r}_k, \widehat{\mathbf{z}}_k) = \frac{\mathrm{tr}((\mathbf{I}-\mathbf{W}^k)\mathbf{W}^k)}{\sqrt{\mathrm{tr}((\mathbf{I}-\mathbf{W}^k)^2)\mathrm{tr}(\mathbf{W}^{2k})}} = \frac{\sum_{i=1}^{n}(1-\lambda_i^k)\lambda_i^k}{\sqrt{\sum_{i=1}^{n}(1-\lambda_i^k)^2\sum_{i=1}^{n}\lambda_i^{2k}}} \tag{13}$$

where $\{\lambda_i\}$ denote the eigenvalues of the filter. Mrázek and Navara[8] claim that minimizing (13) should give an optimal k for minimizing MSE in the diffusion process. This function generally does not have a well-defined minimum in a patch-wise denoising approach, and we present an effective alternative. We can see that (13) does not depend on the noise, so the criterion would not seem to work generally. In fact as k grows, the correlation criterion always converges to zero. In section 3 , we give an alternative for the optimal stopping time.

3. PROPOSED METHOD

If we clairvoyantly had the mean-square error MSE function in each iteration, finding the optimal stopping time would be trivial, as this would correspond to the minimum of the MSE. Otherwise, to find the minimum of the MSE, an estimate in each iteration is needed. A close approximation of the MSE for the diffusion process has been proposed.[5] Considering the eigen-decomposition of $\mathbf{W} = \mathbf{V}\mathbf{S}\mathbf{V}^T$, the image $\mathbf{z}$ can be written in the column space of $\mathbf{V}$ as $\mathbf{z} = \mathbf{V}\mathbf{b}_0$, where $\mathbf{b}_0 = [b_{01}, b_{02}, ..., b_{0n}]^T$, where $\{b_{0i}^2\}$ are the signal energy distribution over all the modes. We start our discussion with a description of the predicted MSE for the diffusion process after k iterations:

$$\mathrm{MSE}_k = \sum_{i=1}^{n}(1-\lambda_i^k)^2 b_{0i}^2 + \sigma^2\lambda_i^{2k} \tag{14}$$

Our objective is to find $\widehat{k}$ which is the optimal stopping time that minimizes MSE_k for the diffusion process. Since in practical denoising problems there is no access to the clean image, the MSE needs to be estimated. In evaluating (14), we need to have λ_i and b_{0i}. In general, we have the eigenvalues λ_i in advance from the given filter $\mathbf{W}$. What is missing is an estimate of the coefficients b_{0i}, which we denote by $\widehat{b}_{0i}$. With this notation, the estimate of the MSE can be written as:

$$\widehat{\theta}_k = \widehat{\mathrm{MSE}}_k = \sum_{i=1}^{n}(1-\lambda_i^k)^2\,\widehat{b}_{0i}^2 + \sigma^2\lambda_i^{2k} \tag{15}$$

Next, we propose a way to estimate the coefficients $\mathbf{b}_0 = [b_{01}, b_{02}, ..., b_{0i}]^T$. We apply the given filter to the data as a "pre-processor" as follows:

$$\widehat{\mathbf{b}}_0 = \mathbf{V}^T\widehat{\mathbf{z}} = \mathbf{V}^T\mathbf{W}^m\mathbf{y} \tag{16}$$

in which $\mathbf{W}^m = \mathbf{V}\mathbf{S}^m\mathbf{V}^T$, $\mathbf{S}^m = \mathrm{diag}[\lambda_1^m, ..., \lambda_n^m]$ where m can be any positive real number. This means that in our search for m we are not just limited to integer values. With this estimate of the vector $\mathbf{b}_0$, we can then study the behavior of the estimated MSE, $\widehat{\theta}_{k,m}$ as a function of m as follows. The bias for this estimate can be expressed as:

$$\begin{aligned}\text{bias}(\widehat{\theta}_{k,m}) &= \text{E}(\widehat{\theta}_{k,m}) - \theta_k &(17)\\ &= \sum_{i=1}^{n}(1-\lambda_i^k)^2(\lambda_i^{2m}b_{0i}^2 + \lambda_i^{2m}\sigma^2 - b_{0i}^2) &(18)\\ &= \sum_{i=1}^{n}(1-\lambda_i^k)^2\text{bias}(\widehat{b}_{0i}^2) &(19)\end{aligned}$$

It is evident that $\text{bias}(\widehat{\theta}_{k,m})$ is a linear combination of the bias in each channel i. As k grows, the bias tends to a constant value of $\sum_{i=2}^{n}\text{bias}(\widehat{b}_{0i}^2)$ in each patch*. Similarly, the variance of the estimator also can be written as a linear combination of the variance in each channel i:

$$\begin{aligned}\text{var}(\widehat{\theta}_{k,m}) &= 2\sigma^2\sum_{i=1}^{n}\lambda_i^{4m}(1-\lambda_i^k)^4(2b_{0i}^2+\sigma^2) &(20)\\ &= \sum_{i=1}^{n}(1-\lambda_i^k)^4\text{var}(\widehat{b}_{0i}^2) &(21)\end{aligned}$$

As k grows , the variance tends to a constant value of $\sum_{i=2}^{n}\text{var}(\widehat{b}_{0i}^2)$ in each patch. The overall error in estimating θ_k across all iterations can be written as the mean *integrated* squared error defined as follows:

$$f(m) = \sum_{k=1}^{T}\text{MSE}(\widehat{\theta}_{k,m}) = \sum_{k=1}^{T}\|\text{bias}(\widehat{\theta}_{k,m})\|^2 + \text{var}(\widehat{\theta}_{k,m}) \quad (22)$$

where T is the total number of iterations which should be set as a sufficiently large integer (well beyond the optimal stopping time). We employ a gradient descent method to find a local minimum to this function. Once we have the estimate for m, the optimal "pre-processor" (16) gives us the estimate for the vector $\mathbf{b}_0$. Now we are ready to minimize $\widehat{\text{MSE}}_k$ with respect to k and estimate the ideal stopping time $\widehat{k}$. In the next section we compare our estimate with the true stopping time, but first some notational matters: $\widehat{k}$ minimizes the actual MSE_k, $\widehat{k}_c$ is the minimum of the clairvoyant approximation MSE_k (given in (14)) and finally $\widehat{k}_e$ is our estimation for the ideal stopping time which minimizes $\widehat{\theta}_k$. Hence, we can formulate these minimizations as:

$$\widehat{k} = \arg\min_k \|\mathbf{z} - \mathbf{W}^k\mathbf{y}\|^2 \qquad (\text{True} \quad \text{MSE}) \quad (23)$$

$$\widehat{k}_c = \arg\min_k \sum_{i=1}^{n}(1-\lambda_i^k)^2 b_{0i}^2 + \sigma^2\lambda_i^{2k} \qquad (\text{Clairvoyant} \quad \text{MSE}) \quad (24)$$

$$\widehat{k}_e = \arg\min_k \sum_{i=1}^{n}(1-\lambda_i^k)^2\widehat{b}_{0i}^2 + \sigma^2\lambda_i^{2k} \qquad (\text{Estimated} \quad \text{MSE}) \quad (25)$$

in which the clairvoyant MSE uses the clean $\mathbf{b}_0$. Starting from $k = 1$ we check if the MSE in (14) and (15) are decreasing. The iteration number k is increased over integer values as long as the MSE is descending.

*Since $\lambda_1 = 1$, the index i in the sum starts from 2.

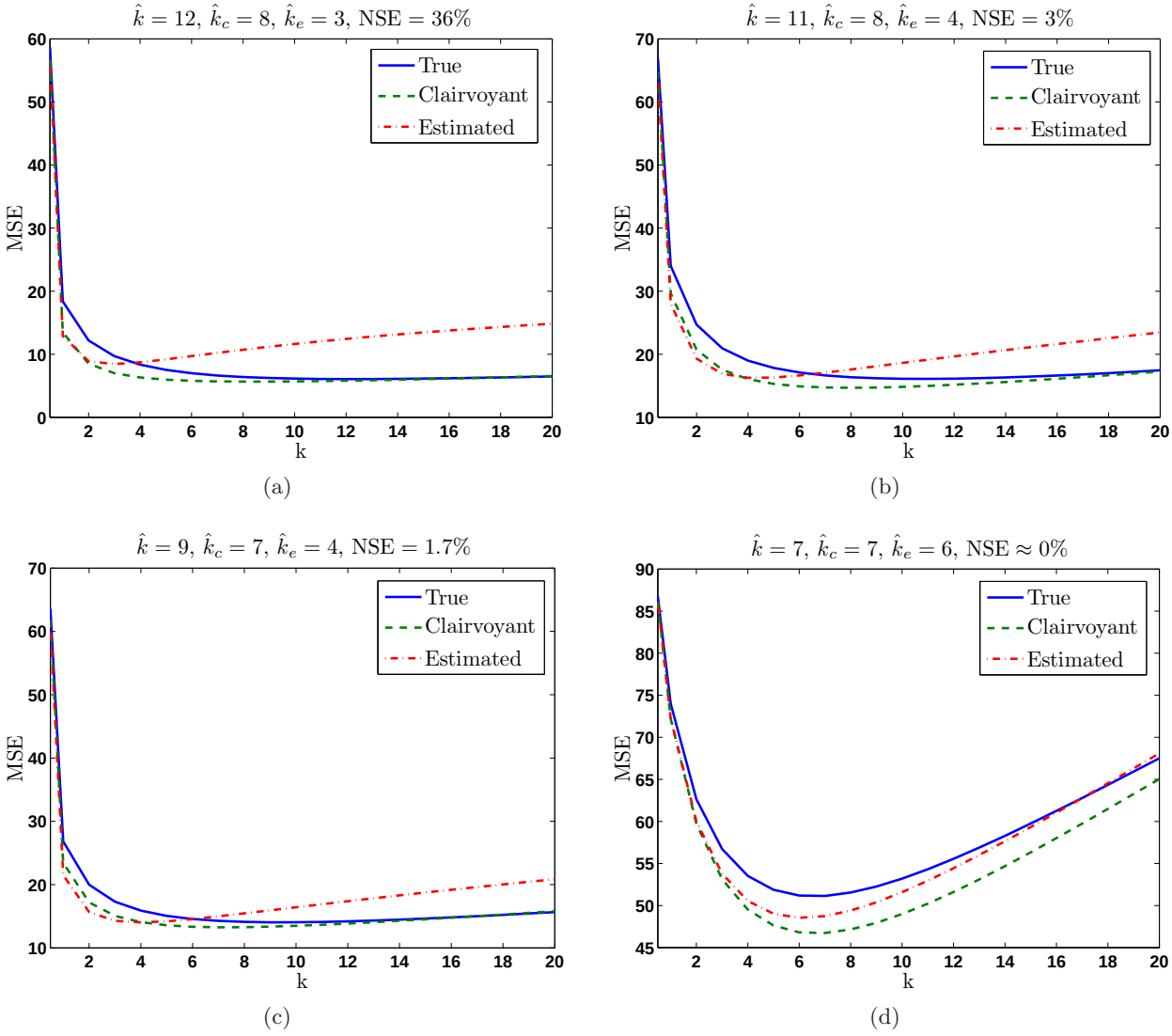

Figure 1. Monte-Carlo estimation of ideal stopping time for LARK filter in diffusion process. Different types of patches:(a) flat,(b) edge, (c) corner,(d) texture.($\sigma^2 = 100$).

4. SIMULATION RESULTS

Before discussing our simulation results, we should note that the non-parametric filters $\mathbf{W}$ are approximately deterministic when the noise variance σ^2 is small relative to the clean data $\mathbf{z}$. In the other words, when signal-to-noise ratio is high, the matrix $\widetilde{\mathbf{W}} = \mathbf{W}(\mathbf{y})$ computed from noisy data is close to the matrix $\mathbf{W}(\mathbf{z})$.[5] In order to get a closer approximation of $\mathbf{W}(\mathbf{z})$, in our simulations the filter $\mathbf{W}(\widehat{\mathbf{z}})$ is used which is computed from the pre-filtered image $\widehat{\mathbf{z}} = \widetilde{\mathbf{W}}\mathbf{y}$. To measure the effectiveness of our stopping criterion, we introduce a measure to compare the relative size of the correct mean squared error at the true minimum versus its value at the estimated minimum given by $\widehat{k}_e$. Normalized-Squared Error (NSE) is defined as follows:

$$\mathrm{NSE} = 100 \times \frac{(\mathrm{MSE}_{\widehat{k}} - \mathrm{MSE}_{\widehat{k}_e})^2}{\mathrm{MSE}^2_{\widehat{k}}}\% \tag{26}$$

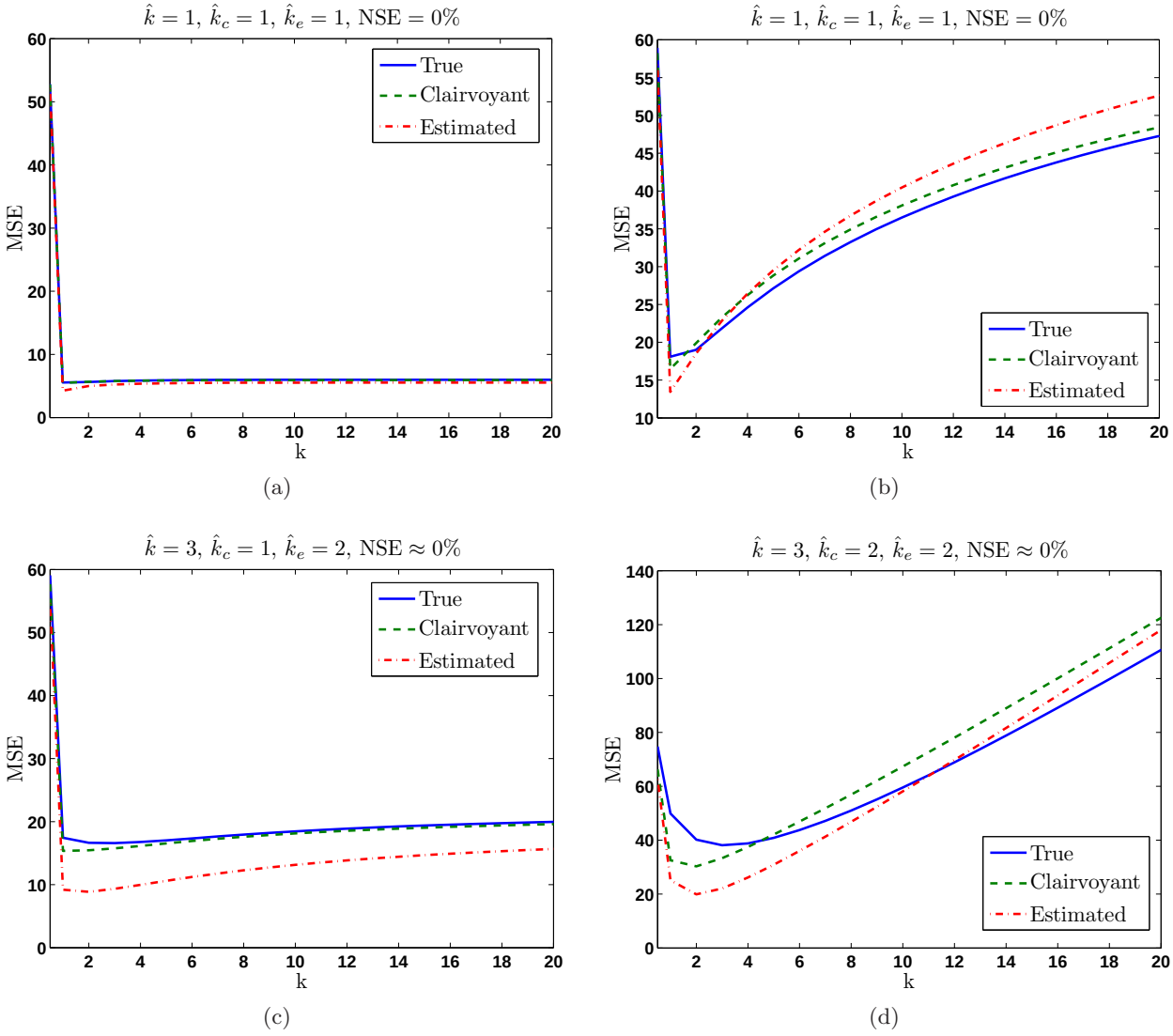

Figure 2. Monte-Carlo estimation of ideal stopping time for NLM filter in diffusion process. Different types of patches:(a) flat,(b) edge, (c) corner,(d) texture.($\sigma^2 = 100$).

In our simulations the patch size was 21×21 and additive white Gaussian noise with $\sigma^2 = 100$ was added to the image. The total number of iterations, T was set to 20. Fig. 1 illustrates actual, ideally predicted and estimated MSE for different types of patches shown in Fig. 3(a)-(d). The Monte-Carlo estimation is done for 50 noise realizations and each time the filter $\mathbf{W}$ is computed from the pre-filtered image for the LARK kernel.[1] It can be observed that the estimated MSEs are quite close to the true and predicted ones. The experiment is repeated, this time using the NLM kernel[3] for the construction of $\mathbf{W}$ in Fig. 2. As can be seen, in all the tested patches for NLM filter, the normalized error in locating the optimal MSE is close to zero. Experiments for denoising different patches (see Fig. 3) are carried out, where the same LARK and NLM filters were used for the diffusion process. As can be seen the proposed stopping time strategy shows a good performance in all the patches for both LARK and NLM filters. The largest differences between the estimated and true MSE occur in the case of the flat patch, as shown in figures 1(a) and 3(i). Even in this case, where the NSE is not quite negligible, we observe that the visual quality of the results is essentially the same.

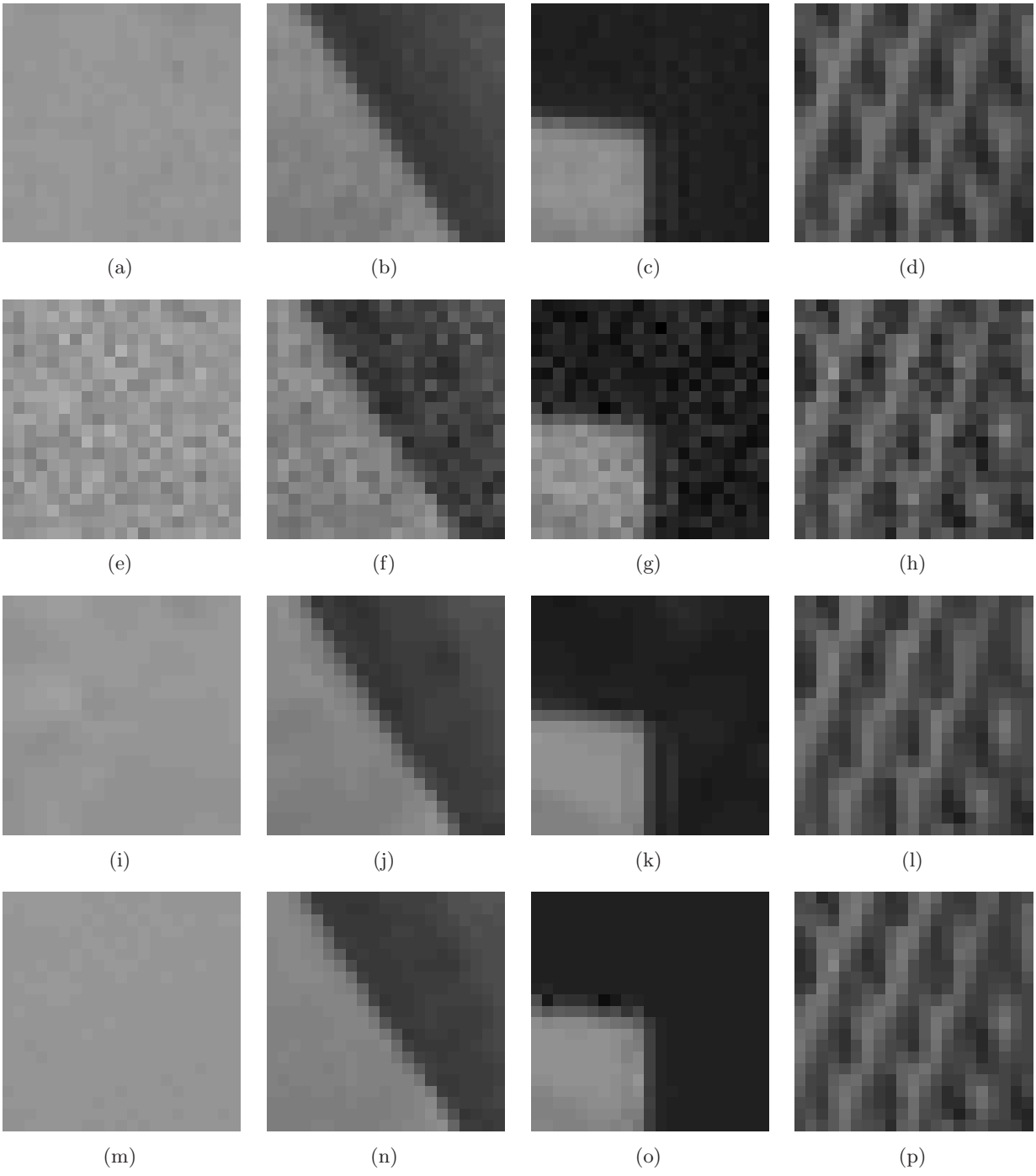

Figure 3. Denoising example for different patches: (a) flat, (b) edge, (c) corner, (d) texture. (e)-(h) noisy patches $\sigma^2 = 100$. (i)-(l) denoised patches by LARK filter (MSEs are 9.71, 18.99, 15.88 and 51.19 respectively). (m)-(p) denoised patches by NLM filter(MSEs are 5.48, 18.09, 17.44 and 40.19 respectively).

5. CONCLUSION AND FUTURE WORKS

To improve the effectiveness of data-dependent filtering, we introduced a new stopping criterion based on minimization of MSE for each patch. Our estimation of MSE was biased and an essential study might be comparison of the presented algorithm with other unbiased estimation methods such as SURE.[10] The proposed method also can easily be extended to the whole image with just an aggregation step for overlapping patches.

REFERENCES

[1] Takeda, H., Farsiu, S., and Milanfar, P., "Kernel regression for image processing and reconstruction," *IEEE Transactions on Image Processing* **16**(2), 349–366 (February 2007).

[2] Tomasi, C. and Manduchi, R., "Bilateral filtering for gray and color images," *International Conference of Compute Vision* , 836–846 (January 1998).

[3] Buades, A., Coll, B., and Morel, J. M., "A review of image denoising algorithms, with a new one," *Multiscale Modeling and Simulation (SIAM interdisciplinary journal)* **4**(2), 490–530 (2005).

[4] Singer, A., Shkolinsky, Y., and Nadler, B., "Diffusion interpretation of nonlocal neighborhood filters for signal denoising," *SIAM Journal on Imaging Sciences* **2**(1), 118–139 (2009).

[5] Milanfar, P., "A tour of modern image filtering," *Invited article to appear in IEEE Signal Processing Magazine* (2011).

[6] Perona, P. and Malik, J., "Scale-space and edge detection using anistropic diffusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **12**(9), 629–639 (July 1990).

[7] Sporring, J. and Weickert, J., "Information measures in scale-spaces," *IEEE Transactions on Information Theory* **45**(3), 1051–1058 (1999).

[8] Mrázek, P. and Navara, M., "Selection of optimal stopping time for nonlinear diffusion filtering," *International Journal of Computer Vision* **52**(2-3), 189–203 (2003).

[9] Papoulis, A., [*Probability and statistics*], Prentice-Hall, Englewood Cliffs, NJ (1990).

[10] Stein, C. M., "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics* **9**(6), 1135–1151 (November 1981).