MULTIPLE COPY IMAGE DENOISING VIA WAVELET THRESHOLDING

S. Grace Chang[1] *Bin Yu*[2] *Martin Vetterli*[1,3]

[1]Department of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA 94720, USA
[2]Department of Statistics
University of California, Berkeley, CA 94720, USA
[3]Département d'Electricité
Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland
grchang@eecs.berkeley.edu, binyu@stat.berkeley.edu, vetterli@de.epfl.ch

ABSTRACT

This work addresses the recovery of an image from noisy observations when multiple noisy copies of the image are available. The standard method is to compute the average of these copies. Since the wavelet thresholding technique has been shown to effectively denoise a single noisy copy, it is natural to consider combining these two operations of averaging and thresholding. The first important task is to find the optimal ordering. The second issue is the threshold selection for each method. By modeling the signal wavelet coefficients as Laplacian distributed and the noise as Gaussian, our investigation finds the optimal ordering to depend on the number of available copies and on the signal-to-noise ratio. We propose thresholds that are nearly optimal under the assumed model for each ordering. With the optimal and near-optimal thresholds, the two methods yield very similar performance, and both show considerable improvement over merely averaging.

1. INTRODUCTION

Since the seminal work of denoising via wavelet thresholding proposed by Donoho and Johnstone [3], there have been many variations in both theory and practice. Most of these works are for applications in which there is only one set of observations (e.g. one time series sequence or one still image). However, in some applications there are multiple copies of the same or similar images, thus it is necessary to investigate denoising techniques which remove noise from multiple corrupted copies of the same signal. For a corrupted video sequence, suppose we choose a few consecutive frames in which the motion is not significant and that we have already taken care of the registration problem, one can view the frames as multiple noisy copies of the same image. Another example is when one scans a picture, but with unsatisfactory result, thus one does multiple scans, and then combines these copies to obtain the most noise-free copy possible. Since wavelet thresholding has worked well for one copy, it is natural to consider its extension to multiple copies.

The standard method for combining the multiple copies is to simply compute their weighted sum. One can only do better by incorporating a thresholding step. The question is, which ordering is better, thresholding first or averaging first, and what is the threshold value for each method? These are the issues to be addressed in this work. With the coefficients of each subband modeled as samples of a Laplacian random variable and the noise as samples of a Gaussian variable, we will show that the optimal ordering (in the mean squared error sense) depends on the number of available copies and the ratio between the noise power and the signal power. Moreover, we propose near-optimal subband adaptive thresholds for both orderings. Results show that with the optimal or the proposed near-optimal thresholds, the two methods yield very similar performance, and both outperforms weighted averaging substantially.

2. COMBINING WEIGHTED AVERAGING AND WAVELET THRESHOLDING

Let f denote the $M \times M$ matrix of the original image to be recovered. The signal f has been transmitted over a additive Gaussian noise channel N times, and at the receiver we have N copies of noisy observations, $g^{(n)} = f + \varepsilon^{(n)}, n = 1, \ldots, N$. For the n-th copy, the pixels $\varepsilon_{ij}^{(n)}$ are *iid* Gaussian $N(0, \sigma_n^2)$, where σ_n^2 is the variance of the n-th copy of noise. The noise between different copies are also assumed independent. The goal is to find an estimate $\hat{f}$ which minimizes the mean squared error (MSE), $\frac{1}{M^2}\sum_{i,j=1}^{M}(\hat{f}_{ij} - f_{ij})^2$.

The recovery of the image is done in the orthogonal wavelet transform domain (the readers are referred to standard wavelet literature such as [4, 5] for details of the 2D dyadic wavelet transform). The wavelet coefficients can be grouped into *subbands* of different scale and orientation, with one lowest frequency subband, and the rest called *detail subbands*. It has been found that for a large class of images, the coefficients in each detail subband can be well described by a Generalized Gaussian distribution [6], which is often simplified to the special case of Laplacian distribution. In this work, we also use the Laplacian distribution. Under probability distributions, the MSE is well approximated by the *expected* squared error. Thus for each detail subband, we wish to find the estimator of the coefficients which minimizes the expected squared error.

2.1. Wavelet Thresholding and Averaging

To denoise one copy, the wavelet thresholding operation by Donoho and Johnstone [3] has three steps. First, take the wavelet transform of the noisy observation $g = f + \varepsilon$, denoted by $Y = X + V$. Then each coefficient is thresholded with a chosen threshold. Finally, the thresholded coefficients are transformed back to yield the recovered signal. The thresholding operation is performed only on the coef-

0-8186-8821-1/98 $10.00

ficients of the detail subbands.

There are two popular thresholding functions: the soft-threshold function, $\eta_\lambda(t) = \text{sgn}(t) \cdot \max(0, |t| - \lambda)$, which shrinks the input towards zero, and the hard-threshold function, $\psi_\lambda(t) = t \cdot 1_{|t|>\lambda}$, which keeps the input only if it is above the threshold λ. Although the soft-thresholding operation tends to smooth the image slightly more than the hard-threshold function, it yields images with better visual quality especially when the noise power is significant. Furthermore, with the chosen probability distributions, soft-thresholding yields a lower MSE than hard-thresholding, as was shown in [1]. Thus, soft-thresholding will be the preferred operation in this work. The next issue is the selection of the threshold value.

As in [1], we model each detail subband of the wavelet coefficients of the original uncorrupted image f as samples from a centered Laplacian random variable with an unknown parameter. That is, let $X \sim p(x) = \text{LAP}(\beta) \triangleq \frac{\beta}{2} e^{-\beta|x|}$, $Y|X \sim p(y|x) = N(x, \sigma^2)$, and the estimator be $\hat{X} = \eta_\lambda(Y)$, then the optimal threshold is

$$\begin{aligned} \lambda^* &= \arg\min_\lambda E_{Y|X} E_X (\hat{X} - X)^2 \\ &= \arg\min_\lambda \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\eta_\lambda(y) - x)^2 \, p(y|x) \, p(x) \, dy \, dx. \end{aligned}$$

A good approximation of λ^* was found in [1] to be $\tilde{\lambda}(\beta) = \frac{\sigma^2}{\sigma_x} = \frac{\sigma^2 \beta}{\sqrt{2}}$, where σ_x is the standard deviation of X. This threshold $\tilde{\lambda}(\beta)$ is simple and effective and has an intuitive explanation. When the noise power is much smaller than the signal power, $\sigma/\sigma_x \ll 1$, the normalized threshold λ/σ is small to preserve most of the signal features; on the other hand, when $\sigma/\sigma_x \gg 1$, λ/σ is chosen to be large to remove the noise which has overwhelmed the signal. By modeling each detail subband as samples from a Laplacian random variable with the unknown parameter β to be estimated, this method also allows the threshold to be adaptive to each subband.

When there are multiple copies available, the standard method is to use the (pixel-wise) weighted average as the estimate. Let $V^{(n)} \sim N(0, \sigma_n^2), n = 1, \ldots, N$, be the random variables representing the n-th copy noise, define Z to be the weighted sum of the N random variables $Y^{(n)} = X + V^{(n)}$,

$$Z = \sum_{n=1}^{N} \alpha_n Y^{(n)} = X + \sum_{n=1}^{N} \alpha_n V^{(n)},$$

where $\sum \alpha_n = 1$. It is well-known that the optimal α_n are $\alpha_n^* = \frac{1}{\sigma_n^2} / \sum_{i=1}^{N} \frac{1}{\sigma_i^2}$, and the resulting MSE is $\sigma^2_{\text{total}} = \text{Var}(Z - X) = \text{Var}\left(\sum_{n=1}^{N} \alpha_n^* V^{(n)}\right) = 1/(\sum_{n=1}^{N} \frac{1}{\sigma_n^2})$.

Now let us incorporate thresholding into averaging. The weighted sum Z is essentially a new random variable and $Z|X \sim N(x, \sigma^2_{\text{total}})$. Since this is exactly the setting for one copy thresholding, the next straightforward step is to simply find the best threshold and apply it on Z. However, can we do better than that? More specifically, since we have two operations here — averaging and thresholding — it is natural to ask which ordering is best in the mean squared sense.

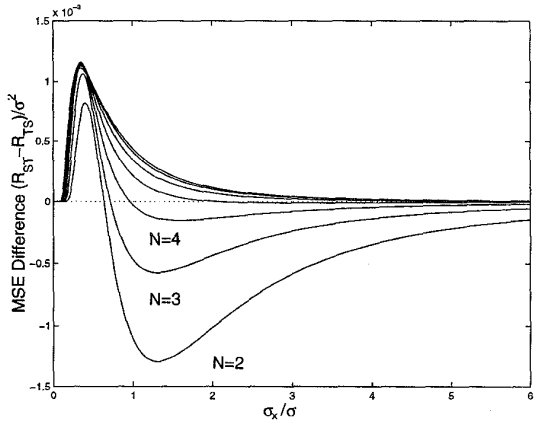

Figure 1. Scaled MSE difference $(R_{ST}(\lambda^*_{ST}) - R_{TS}(\lambda^*_{TS}))/\sigma^2$ as a function of N and σ_x/σ.

2.2. Combining Thresholding With Averaging

Consider the special case when $\sigma_1 = \sigma_2 = \cdots = \sigma_N \triangleq \sigma$. Thus, $\alpha_1 = \cdots = \alpha_N = \frac{1}{N}$. To make references more convenient, let $S(\cdot)$ denote the weighted sum operation and $T(\cdot)$ the thresholding operation, and we give the following notation to the two orderings:

$$\begin{aligned} S(T(Y^{(1)}, \ldots, Y^{(N)})): &\quad \hat{X}_{ST}(\lambda) = \tfrac{1}{N} \textstyle\sum_{n=1}^{N} \eta_\lambda(Y^{(n)}) \\ T(S(Y^{(1)}, \ldots, Y^{(N)})): &\quad \hat{X}_{TS}(\lambda) = \eta_\lambda\left(\tfrac{1}{N} \textstyle\sum_{n=1}^{N} Y^{(n)}\right). \end{aligned}$$

The MSE or *risk* of the $S(T(\cdot))$ method is

$$\begin{aligned} R_{ST}(\lambda) &= E_X E_{Y^{(1)},\ldots,Y^{(N)}|X} (\hat{X}_{ST}(\lambda) - X)^2 \\ &= \frac{1}{N} E_X E_{Y|X} (\eta_\lambda(Y) - X)^2 \\ &\quad + \frac{N-1}{N} E_X \left[E_{Y|X}(\eta_\lambda(Y) - X)\right]^2, \quad (1) \end{aligned}$$

where $Y|X \sim N(x, \sigma^2)$ and (1) follows from the fact that $\{Y^{(1)}, \ldots, Y^{(N)}\}$ conditioned on X are independent. The risk of $T(S(\cdot))$ is

$$\begin{aligned} R_{TS}(\lambda) &= E_X E_{Y^{(1)},\ldots,Y^{(N)}|X} (\hat{X}_{TS}(\lambda) - X)^2 \\ &= E_X E_{Z|X} (\eta_\lambda(Z) - X)^2, \end{aligned}$$

where $Z|X \sim N(x, \frac{\sigma^2}{N})$. The optimal threshold is the argument which minimizes the risk, that is, $\lambda^*_{ST} = \arg\min_\lambda R_{ST}(\lambda)$, and $\lambda^*_{TS} = \arg\min_\lambda R_{TS}(\lambda)$.

To compare the risks of these two methods, we look at the scaled difference $(R_{ST}(\lambda^*_{ST}) - R_{TS}(\lambda^*_{TS}))/\sigma^2$ as a function of N (the number of copies available) and of the ratio σ_x/σ, as illustrated in Figure 1. For each $N \leq 5$, there is a cutoff point C^*_N below which $R_{ST}(\lambda^*_{ST}) > R_{TS}(\lambda^*_{TS})$, and above

which $R_{ST}(\lambda^*_{ST}) < R_{TS}(\lambda^*_{TS})$. For $N > 5$, however, the $T(S(\cdot))$ method is better for any value of σ_x/σ. This finding indicates that the best method depends on the relative power between the noise and signal, and also on the value of N. With the optimal thresholds, the improvement of one method over the other is small, on the order of $10^{-3}\sigma^2$. The $T(S(\cdot))$ method requires much less computation than the $S(T(\cdot))$ method (since the former can be implemented by computing the wavelet transform once, whereas the latter computes it N times), thus if computation is an issue, the $T(S(\cdot))$ method is preferred.

We do not have closed form solutions for λ^*_{TS} and λ^*_{ST}, thus their values would need to be calculated each time or be tabulated. However, we have found that they can be well approximated by simple closed form expressions. For the $T(S(\cdot))$ estimator, the threshold is simply a modification of $\tilde{\lambda}$ for one copy denoising, but with a change in the noise variance,

$$\tilde{\lambda}_{TS} = \frac{\sigma^2/N}{\sigma_x}. \tag{2}$$

For the $S(T(\cdot))$ method, we use the approximation

$$\tilde{\lambda}_{ST} = \frac{\sigma^2/N^{(3/4)}}{\sigma_x}. \tag{3}$$

The threshold for $S(T(\cdot))$ needs to decrease as N increases, even though at the thresholding stage, each copy is thresholded independently of the other copies. To explain this, note that the inner expectation of $R_{ST}(\lambda)$ can be written as

$$E_{Y^{(1)},\ldots,Y^{(N)}|X}(\hat{X}_{ST}(\lambda) - X)^2 =$$
$$\tfrac{1}{N}E_{Y|X}\left(\eta_\lambda(Y) - E_{Y|X}\eta_\lambda(Y)\right)^2 + \left(E_{Y|X}\eta_\lambda(Y) - X\right)^2.$$

The first term is the variance due to thresholding, while the second term is the square of the bias. The optimal threshold is obtained from the tradeoff between the variance term (which decreases with increasing λ) and the bias term (which increases with increasing λ). As N becomes larger, the variance term decreases due to the $1/N$ factor while the bias term stays the same. Thus, λ needs to be decreased as well to obtain the minimum total.

Figure 2 compares the optimal and approximate thresholds for both methods as a function of N, for $\sigma = 1$ and $\sigma_x = 1$. Using the approximate thresholds $\tilde{\lambda}_{TS}$ and $\tilde{\lambda}_{ST}$ result in less than .2% loss of MSE optimality for any value of σ_x and σ. Figure 3 compares the optimal threshold λ^*_{ST} and the approximation $\tilde{\lambda}_{ST}$ as a function of σ_x/σ for $N = 2,\ldots,6$. It shows that the approximation is good for large σ_x/σ but not as well for very small σ_x/σ, especially for large N. The loss of MSE optimality is less than 3.5% for $\sigma_x/\sigma < 1$ and less than .1% for $\sigma_x/\sigma > 1$. However, since typically the signal power is much larger than the noise power, inaccurate approximations for small σ_x/σ are acceptable. The thresholds $\tilde{\lambda}_{TS}$ and $\tilde{\lambda}_{ST}$ also yield a different set of cutoff values $\tilde{C}_N$, but the scaled MSE difference $(R_{ST}(\tilde{\lambda}_{ST}) - R_{TS}(\tilde{\lambda}_{TS}))/\sigma^2$ is similar to the curves shown in Figure 2 for optimal thresholds and is of the same order of magnitude. Thus, the use of $\tilde{\lambda}_{TS}$ and $\tilde{\lambda}_{ST}$ do not perturb the previous results much.

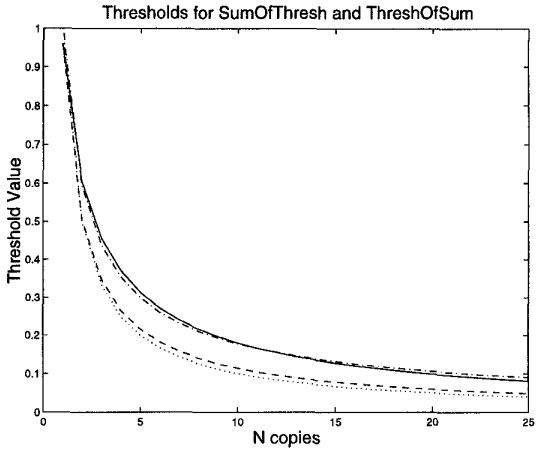

Figure 2. Comparing λ^*_{TS} (− − −) versus $\tilde{\lambda}_{TS}$ (···), and λ^*_{ST} (—) versus $\tilde{\lambda}_{ST}$ (− · −·), when $\sigma = 1$ and $\sigma_x = 1$.

Up to now we have assumed that we have knowledge of the noise variance σ^2 and the standard deviation σ_x. In practice, these two values are not known and have to be estimated from the noisy observations. For both methods, these two parameters are estimated the same way for a fair comparison. First the noise variance σ_n^2 is estimated by the robust median estimator in the highest subband (also used in [3]), $\hat{\sigma}_n = \text{Median}(|Y_{ij}|)/.6745$, with all Y_{ij} in the HH_1 subband of the n-th copy, then $\hat{\sigma}^2$ is taken to be the average of these N estimates. Since the noise is independent from the signal, $\text{Var}(Z) = \text{Var}(X) + \sigma^2/N = \sigma_x^2 + \sigma^2/N$. Thus, for each subband of $Z = \frac{1}{N}\sum_{n=1}^{N} Y^{(n)}$, the sample variance estimate of $\text{Var}(Z)$ is calculated, and the estimate of the standard deviation σ_x of the Laplacian distribution is $\sqrt{(\text{Var}(Z) - \hat{\sigma}^2/N)}$.

Now consider the case when the noise variances σ_n^2 are different. This extension is straightforward in the $T(S(\cdot))$ case. The multiple copies are averaged with coefficients α_n, and the threshold is $\tilde{\lambda}_{TS}$ in (2) but with σ^2/N replaced by σ^2_{total}.

For the $S(T(\cdot))$ method, one needs to find the optimal threshold for each copy and the optimal weights in the summation. By minimizing the risk $E_X E_{Y^{(1)},\ldots,Y^{(N)}|X}\left(\sum_{n=1}^{N}\alpha_n\eta_{\lambda_n}(Y^{(n)}) - X\right)^2$ with respect to $\alpha_1,\ldots,\alpha_N$ subject to $\sum\alpha_n = 1$, and also with respect to $\lambda_1,\ldots,\lambda_N$, one can find the optimal values α_n^* and λ_n^*. The optimal α_n^* are found to be very close to $\frac{1}{\sigma_n^2}/\sum_i\frac{1}{\sigma_i^2}$, and the optimal thresholds can be approximated by $\tilde{\lambda}^{(n)}_{ST} = \frac{\sigma_n^{1/2}}{\sigma_x}\left(1/(\sum_{i=1}^{N}\frac{1}{\sigma_i^2})\right)^{3/4}$, $n = 1,\ldots,N$, which yields $\tilde{\lambda}_{ST}$ in (3) when $\sigma_1 = \sigma_2 = \cdots = \sigma_N$. For a given set of σ_n's, this approximation is good for the threshold corresponding to the smallest σ_n, and it worsens for thresholds corresponding to larger σ_n. This inaccuracy is

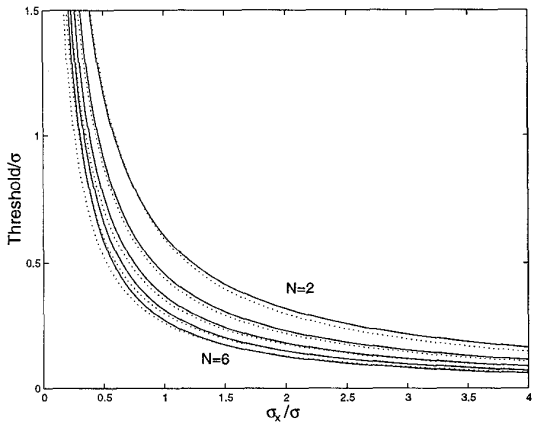

Figure 3. Comparing λ^*_{ST} (—) and $\tilde{\lambda}_{ST}$ ($\cdots$) for $\sigma_1 = \cdots, \sigma_N \stackrel{\Delta}{=} \sigma$ as a function of σ_x/σ and $N = 2, \ldots, 6$.

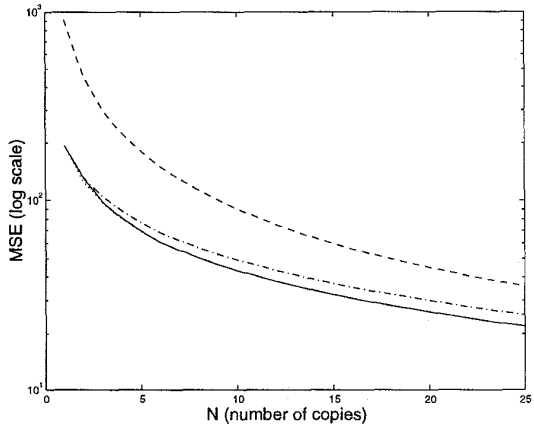

Figure 4. Comparing for each N the MSE of averaging ($- - -$), $S(T(\cdot))$ ($- \cdot -$), $T(S(\cdot))$ ($\cdots$), and switching (—), for $\sigma = 30$.

mitigated by the fact that the weights α^*_n's for copies with large σ_n's are small, thus the overall MSE is still close to the optimal MSE.

3. EXPERIMENTAL RESULTS

To validate the theory, we take as the test image a 256×256 block from the image *barbara*, with $\sigma_1 = \ldots = \sigma_N = \sigma = 30$, using Daubechies' least unsymmetric wavelet with 8 vanishing moments and 4 scales of wavelet transform [2]. The parameters σ_x and σ are estimated as in prior discussion. We compare the MSEs of four methods for a range of N: averaging, $S(T(\cdot))$, $T(S(\cdot))$, and switching between the two thresholding methods (only for $N \leq 5$) with cutoff values $\tilde{C}_N$ (thus the switching method becomes $S(T(\cdot))$ for $N > 5$) . The resulting MSEs are shown in Figure 4. The three thresholding methods show significant improvement over merely averaging, ranging from 70% to 30% reduction in MSE for N varying from 2 to 30. The removal of noise due to thresholding is also significant visually (see Figure 5), especially for small N. Among the thresholding methods, the $T(S(\cdot))$ method is the best in terms of MSE, even better than switching, suggesting that perhaps the $S(T(\cdot))$ method is more sensitive to model errors and threshold estimation errors. For $1 < N \leq 5$, the switching method yields MSEs that are between those of $S(T(\cdot))$ and $T(S(\cdot))$. Visually, one does not discern any difference between the results from these three thresholding methods. The $T(S(\cdot))$ method also requires the least amount of computation since it can be implemented with only one wavelet transform. Thus, in practice, this method suffices to combine multiple noisy copies.

It is curious to investigate if an additional stage of thresholding can have a significant improvement. It cannot do worse, since we can always choose the second stage threshold to be zero. To test this idea, we take the output of $S(T(\cdot))$ and optimally threshold it assuming that we have the original. The resulting MSE is only slightly better than the $T(S(\cdot))$, suggesting that thresholding of the weighted sum yields a sufficiently denoised image already. Furthermore, finding the optimal thresholds of a two-stage thresholding operation is difficult.

4. CONCLUSION

In this paper we addressed the issue of image recovery from multiple copies of noisy observations, and explored the idea of combining the wavelet thresholding technique with the more traditional averaging operation. The investigation showed that the optimal ordering of these two operations is not so straightforward and is in fact a function of the number of available copies and of the relative energy between noise and signal. We also proposed near-optimal thresholds for each ordering. With these thresholds, the performances are similar, and for computational reasons, averaging followed by thresholding is recommended. Furthermore, all of these thresholding methods show substantial improvement over mere averaging, both visually and in the MSE sense.

REFERENCES

[1] S.G. Chang, B. Yu, and M. Vetterli, "Image Denoising via Lossy Compression and Wavelet Thresholding," *Proc. IEEE Int. Conf. Image Processing*, Vol.1, pp. 604-607, Nov. 1997.

[2] I. Daubechies, *Ten Lectures on Wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics*, SIAM, Philadelphia, 1992.

[3] D.L. Donoho and I.M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol 81, pp. 425-455, 1994.

[4] S. Mallat, "A theory for multiresolution signal decomposition: The wavelet representation," *IEEE Pat. Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674-693, July 1989.

[5] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice Hall, Englewood Cliffs, NJ, 1995.

[6] P.H. Westerink, J. Biemond, and D.E. Boekee, "An optimal bit allocation algorithm for sub-band coding",

Figure 5. Denoised images, for $N = 5$. From left to right, top to bottom: original, noisy image with $\sigma = 30$, averaging, switching, $S(T(\cdot))$, and $T(S(\cdot))$. This image can also be found on the Web, at `http://www-wavelet.eecs.berkeley.edu/~grchang/icip98MultThresh.pgm`.

Proc. Int. Conf. on Acous., Speech and Signal Process., Dallas, Texas, pp. 1378-1381, April 1987.