MIX-DISTRIBUTION MODELING FOR OVERCOMPLETE DENOISING

Vladimir Katkovnik, Alessandro Foi, and Karen Egiazarian

Institute of Signal Processing,
Tampere University of Technology, Finland

Abstract: Localized or windowed data denoising based on linear transforms equipped with some thresholding operator is a usual approach in modern signal and image processing. With overlapping windows, techniques of this kind can be interpreted as overcomplete (redundant) data transforms (representations). In the simplest formulation, the final estimates for points belonging to multiple overlapping windows are calculated as the mean of the estimates independently obtained for each of the windows. In this paper we propose a general approach leading to a mix-distribution modeling of the overcomplete data and to least-squares optimal final estimates in the form of weighted average of the least-square estimates for the windowed data. Experiments demonstrate the advanced performance of this class of the algorithms, in particular in comparison with the standard ones using the sample averaging of the windowed estimates. *Copyright ©2007 IFAC*

Keywords: Adaptive denoising, overcomplete transform, estimate aggregation

1. INTRODUCTION

Overcomplete representations arise naturally when data are processed by overlapping subsets, i.e. windows or blocks, and multiple estimates are obtained for each individual point (e.g., (Elad, 2006), (Guleryuz, 2006) and references therein). Estimation is composed from three successive steps: first data windowing (blocking); second, window-wise processing; and, third, calculation of the final estimate by fusing the multiple window-wise estimates. It is found, that this sort of redundant estimates essentially improves results versus the standard non-redundant transform-thresholding procedures. For example, it is well known that translation-invariant undecimated wavelets (Coifman and Donoho, 1995) are a dramatically more effective domain than the basic fully decimated orthogonal (or biorthogonal) wavelets for denoised a signal by shrinkage; the realization of

such particular approach, so-called "cycle-spinning", is very simple: it applies the fully decimated transform on all translated replicas of the noisy signal; thresholds this overcomplete set of coefficients and inverts the transform, obtaining in this way an estimate of each of the translated replicas of the original signal; translates back these estimates to the original position, where they are averaged together, thus producing the final estimate of the signal. Because of its relative simplicity, this procedure is extremely popular and symbolizes the importance of overcompleteness and redundancy in signal filtering. In imaging, in particular, the overcomplete methods attenuate the ringing effects and other artifacts typical for the standard orthogonal transforms (see, e.g., (Hua and Orchard, 2003), (Öktem *et al.*, 1999), (Yaroslavsky *et al.*, 2001), and (Öktem *et al.*, 2001)).

A proper design of all the three main stages of these methods is of importance. In particular, it is well known that the sample average for fusing of the window-wise estimate, routinely used in many methods, can be improved by replacing it by a weighted

average. A number of advanced denoising algorithms demonstrating a very good performance have been recently developed explicitly based on the overcomplete filtering in localized transform domain using a data-adaptive weighted average to combine the redundant local estimates. The discrete cosine transform (DCT) filtering with the adaptive-size square window is proposed in (Foi *et al.,* 2006). Further development of this idea results in the shape-adaptive DCT transform filter (Foi *et al.,* 2007a), where starshaped adaptive neighborhoods are found for each estimation point. In both these methods for each point we obtain the multiple estimates fused together with the adaptive weights into the final estimate. Variants of these algorithms for non-Gaussian noise models are also proposed (Foi *et al.,* 2006), (Foi *et al.,* 2007b). A different and innovative idea is exploited in (Dabov *et al.,* 2007) where the matching of the image blocks is used for 3D grouping of 2D image data, then processed using the so-called collaborative filtering. In this case we again arrive to the multiple blockwise estimates fused for calculation of the final estimate of each individual point.

However, these advanced algorithms have been developed based mainly on the intuitive heuristic ideas. One of the motivation of this paper is to develop a theoretical base for this class of algorithms, with the aim to develop a general framework for design and optimization of the procedures based on the overcomplete representations.

In this paper we concentrate on the two last stages of the overcomplete signal processing, window-wise filtering and estimate fusing, assuming that the windowing (i.e., the first stage) is given and fixed. In particular, we propose a mix distribution model from which the filtering in the window-wise processing and the following fusing can be jointly derived, providing least-squares optimal estimates. Thus, this special modeling suggest a sense in which the above algorithms can be interpreted as optimal ones.

2. OBSERVATION MODELING AND OVERCOMPLETE DATA REPRESENTATION

Let $y(x)$, $x \in X \subset Z^1$ be a signal defined on the integer grid $X = \{x = 1, \dots, L\}$. We consider a windowing $\mathcal{C} = \{X_r, r = 1, \dots, N_s\}$ of X with N_s subsets (uniform windows) $X_r \subset X$ of size n_r such that $\cup_{r=1}^{N_s} X_r = X$. Thus, each $x \in X$ belongs to at least one subset X_r. The subsets may be overlapping and therefore some of the elements may belong to more than one subset. The total number of the subsets is N_s. Such windowing is called a *covering* of X.

Let Y_r be an n-vector composed from the samples y_j of y which are defined on the r-th subset X_r. The link between the windowed localized signal Y_r and the original L-vector y can be written in the form

$$Y_r = B_r y, \qquad (1)$$

where B_r is a $n_r \times L$ matrix. The matrix B_r has the $n_r \times n_r$ identity matrix in the columns corresponding to location of the elements of the r-th window. All other elements of B_r are equal to zero.

Let Y be a $\sum_r n_r$-vector composed from all windowed signals Y_r. It is a structured vector of the form $Y = (Y_1^T, Y_2^T, \dots, Y_{N_s}^T)^T$. The link between this composed vector and the original signal-vector y can be given as

$$Y = By, \qquad (2)$$

where B is a structured $\sum_r n_r \times L$ matrix, $B^T = (B_1^T, B_2^T, \dots, B_{N_s}^T)^T$.

It can be realized that B is a full rank matrix, $rang(B) = L$, and in this way the transform (2) is invertible in the following way

$$y = B^\# Y, \qquad B^\# = (B^T B)^{-1} B^T, \qquad (3)$$

where $B^\#$ is the Moore-Penrose generalized inverse. The matrix B can be called *the windowing matrix*, which forms the windowed signals and overcomplete data vector Y from the original data vector y. Analogously, the matrix $B^\#$ can be called the *de-windowing* matrix, as it returns the original data vector y from the structured vector Y of the windowed ones. The formula (3) can be given in the equivalent form $y = (\sum_{r=1}^{N_s} B_r^T B_r)^{-1} \sum_{r=1}^{N_s} B_r^T y$.

In the de-windowing matrix $B^\#$, the matrix B^T moves the elements of Y to the correct positions in the vector y. However, because the vector Y is overcomplete, the product $B^T y$ returns the sums of the elements which are repeated at the same position. The diagonal matrix $B^T B$ counts the number of these repetitions and then the factor $(B^T B)^{-1}$ produces the corresponding normalization of the vector $B^T y$ to the values corresponding to the vector y. Thus, the matrices B^T and $(B^T B)^{-1}$ in the pseudoinverse $B^\#$ perform these different operations which result in the perfect reconstruction of y from Y.

Let us consider noisy observations

$$z(x) = y(x) + \sigma \varepsilon(x), \quad x \in X, \qquad (4)$$

where ε is i.i.d. zero mean Gaussian noise, $\varepsilon(\cdot) \sim \mathcal{N}(0, 1)$. The above data windowing defines the noisy data windows and overcomplete noisy data set Z as

$$Z_r = B_r z, \qquad Z = Bz, \qquad z = B^\# Z.$$

Applying the B_r matrix to both sides of the equation (4), we can see that $Z_r = B_r y + \sigma B_r \varepsilon = Y_r + \sigma \cdot \tilde{\varepsilon}_r$, where $\tilde{\varepsilon}_r = B_r \varepsilon$ is a random vector of the size n_r and the elements of this vector are correlated with correlation matrix $R_r = E\{\tilde{\varepsilon}_r \tilde{\varepsilon}_r^T\} = B_r B_r^T$.

In a similar way, for the vector Z we have $Z = By + \sigma B \varepsilon = Y + \sigma \cdot \tilde{\varepsilon}$, where $\tilde{\varepsilon} = B\varepsilon$ is a random vector of the size $\sum_{r=1}^{N_s} n_r$ and the elements of this vector are correlated with correlation matrix $\tilde{R} = E\{\tilde{\varepsilon}\tilde{\varepsilon}^T\} = BB^T$. This $\sum_{r=1}^{N_s} n_r \times \sum_{r=1}^{N_s} n_r$ correlation matrix is singular with $rank(\tilde{R}) = L$.

3. MATHEMATICAL FORMULATION FOR ESTIMATION

3.1 Transform-domain sparsity and denoising

As a main tool for denoising we use transforms (orthonormal series) in conjunction with the concept of the redundancy of natural signals. Mainly these are the Fourier transform (FFT), DCT, polynomial and wavelet transforms. The transform is calculated for each window independently, with the spectrum-vector defined as

$$\theta_r = D_r Y_r, \quad r = 1, \ldots, N_s, \tag{5}$$

where θ_r is a n_r-vector and D_r is a squared orthonormal $n_r \times n_r$ transform matrix.

The sparsity of the signal means that it can be well approximated by a small number of non-zero elements of the spectrum θ_r. The number of the non-zero elements of the spectrum is interpreted as the complexity of the model. Then, the complexity for the windowed data modeling is the count of the non-zero elements in the vector θ_r, denoted using the standard notation $||\theta_r||_0$.

We use the observations (4) in the window-wise form

$$Z = Y + \sigma\tilde{\varepsilon}, \quad Y = D^T\theta, \tag{6}$$

$\theta^T = (\theta_1^T, \ldots, \theta_{N_s}^T)^T$ and $D = diag\{D_1, \ldots, D_{N_s}\}$. The goal is to estimate the signal y through the spectrum θ assuming the minimal complexity of the approximation model which, in particular, can be calculated as $||\theta||_0 = \sum_r ||\theta_r||_0$.

3.2 Maximum likelihood

For the Gaussian noise, the penalized minus log likelihood maximization gives the estimates as

$$\hat{\theta} = \underset{\theta}{\text{argmin}}(Z - D^T\theta, R^{\#}(Z - D^T\theta))\frac{1}{\sigma^2} + \tag{7}$$

$$+ \sum_r pen(\theta_r),$$

$$\hat{Y} = D^T\hat{\theta}, \quad \hat{y} = B^{\#}\hat{Y}. \tag{8}$$

The penalty term $pen(\theta)$ used for characterization of the model complexity appears naturally in this modeling, provided that the spectrum θ_r is random with the prior density $p(\theta_r) \propto \exp(-pen(\theta_r))$. The estimator (7) can be presented in the following equivalent form

$$\hat{\theta} = \arg\min \frac{1}{\sigma^2}||z - B^{\#}D^T\theta||^2 + \sum_r pen(\theta_r) \tag{9}$$

$$\hat{Y} = D^T\hat{\theta}, \quad \hat{y} = B^{\#}\hat{Y}.$$

The models (7)-(9) explicitly take into consideration the noise correlation between windows. The simultaneous optimization of the all components of the spectrum θ is a complex problem and the advantage of simplicity of the spectrum estimation for the separate windows is lost here.

The modeling (7)-(9) makes clear the following point which of importance for the overcomplete approach overall. The solution for both θ and y essentially depends on the used de-windowing matrix. The Moore-Penrose inverse gives

$$\hat{y} = (\sum_r B_r^T B_r)^{-1} \sum_r B_r^T \hat{Y}_r,$$

i.e. the final estimate is the sample average of the overlapping window-wise estimates. This is the standard fusing of the window-wise estimates, which is used, e.g., in (Coifman and Donoho, 1995), (Hua and Orchard, 2003), (Öktem et al., 1999), (Öktem et al., 2001), and (Elad, 2006).

It is obvious that there is an infinite number of alternative de-windowing matrices A different from the Moore-Penrose inverse, to invert the windowing matrix B and returning y from Y in the form of the linear transform $y = AY$. Using different inverses, we obtain different solutions. In this way, a proper selection of A can be a powerful tool for modification and optimization of the overcomplete estimation. In what follows, we demonstrate that the some randomization of the matrix A leads to the algorithms optimal in least-squares sense with the fused estimate calculated as the weighted sum of the window-wise estimates. Under some assumptions, the derived algorithms coincide with the advanced algorithms discussed in Introduction. The high efficiency of these algorithms explicitly proves that the use of the de-windowing different from the Moore-Penrose inverse is a good idea. As a complimentary result, it demonstrates that provided conditions used in the derivation of the algorithms in the following sections, the discussed heuristic advanced algorithms become the optimal ones in least-squares sense.

4. RANDOMIZED DE-WINDOWING OF OVERCOMPLETE DATA

As alternative to the reconstruction of y from the windows Y given by the Moore-Penrose generalized inverse (3), we introduce the following randomized model. Let $x \in X$ and $X_r, r = 1, \ldots, N_s$, be a set of windows such that $x \in X_r$. Define a random window $\tilde{X}$ as

$$\tilde{X} = \bigcup_{r:\alpha_r=1} X_r, \tag{10}$$

where α_r are binary (zero-one) random variables equal to 1 with probability p_r, $\sum p_r = 1$. The variables α_r are not independent, but random indicators (selectors) such that only one of the variables α_r takes value 1 while all other are zero in any random realization. The corresponding randomly windowed true values of the signal is

$$\tilde{Y} = \sum_r Y_r \alpha_r.$$

Denote as $\tilde{Y}(x), Y_r(x)$ the elements of Y_r corresponding to $y(x)$, so that $\tilde{Y}(x) = Y_r(x) = y(x)$, and assume

that y is random with the probability density $f(y)$. Then, the probability density $f_0(\tilde{Y})$ for $\tilde{Y}$ has a form of the mix distribution

$$f_0(\tilde{Y}) = \sum_r f_r(Y_r) p_r, \qquad (11)$$

where $f_r(Y_r)$ is a density for the windowed data Y_r. The similar representation is valid for the conditional distribution

$$f_0(\tilde{Y}|z) = \sum_r f_r(Y_r|z) p_r. \qquad (12)$$

Assume further that $f_r(Y_r|z) = f_r(Y_r|Z_r)$. This happens if the components of the random vector y are independent, $f(y) = \Pi f_s(y_s)$. Note, that this assumption means that only the random component of y are independent while the conditional means (regressions) can be smooth. The filtering of these smooth functions is the root of the considered estimation algorithms. It follows from (12) that

$$f_0(\tilde{Y}|z) = \sum_r f_r(Y_r|Z_r) p_r. \qquad (13)$$

Let $\mu_r(x)$ be a window-wise least-squares estimate of $y(x)$ obtained from the observations Z_r given in r-th window, $\mu_r(x) = E\{y(x)|Z_r\}$, and $\mu(x) = E\{y(x)|z\}$ be the least-squares estimate obtained from all data z. Then,

$$\mu(x) = \sum_r \mu_r(x) \cdot p_r. \qquad (14)$$

Let us prove this statement. The conditional mean of $\tilde{Y}$ is the least-squares estimate of $\tilde{Y}$ provided a fixed z. Calculations of the conditional means using (12) give

$$\mu_{\tilde{Y}} = \sum_r \mu_{Y_r} \cdot p_r, \qquad (15)$$

where $\mu_{\tilde{Y}} = E\{\tilde{Y}|Z\} = \int \tilde{Y} f_0(\tilde{Y}|z) d\tilde{Y}$ and the partial window-wise $\mu_{\tilde{Y}_r} = E\{Y_r|Z_r\}$ are the estimates for the windowed data. The formula (14) for the pointwise estimates follows from (15), as $\mu(x) = \mu_{\tilde{Y}}(x)$ and $\mu_r(x) = \mu_{Y_r}(x)$.

The formula (14) defines a decomposition of the least-squares estimate of $y(x)$ into the weighted sum of the window-wise least-squares estimates. It is a fusing formula for the window-wise estimates into the final estimate.

The parameters p_r can be used for optimization of this estimate. Assume that the window-wise estimates $\mu_r(x)$ are unbiased. Then, because $\sum p_r = 1$, the estimate $\mu(x)$ is also unbiased. The variance of the estimate $\mu(x)$ is calculated as

$$\sigma_\mu^2(x) = p^T A(x) p, \qquad (16)$$
$$A_{kl}(x) = cov\{\mu_{Y_r}(x), \mu_{Y_l}(x)\},$$

where $cov\{\cdot\}$ means the covariance of the estimates. To minimize the mean squared error, we select the parameters p_r in (14) that minimize the variance of the estimate:

$$\hat{p}(x) = \arg\min\{p^T A(x) p | \ p^T \mathbf{1} = 1\},$$

where $\mathbf{1}$ is a vector with all elements equal to 1. It gives $\hat{p}(x) = A^{-1}(x)\mathbf{1}/(\mathbf{1}^T A^{-1}(x)\mathbf{1})$. Here $\hat{p}$ depends on x and requires to track the correlations between the elements of different windows. In order to simplify calculations, let us assume that the correlation between windows can be neglected. Then, $cov\{\mu_{Y_r}(x), \mu_{Y_l}(x)\} = \sigma_{Y_r}^2(x)\delta_{r,l}$ where $\sigma_{Y_r}^2(x)$ is the variance of the window-wise estimate. It follows that the matrix A becomes diagonal and we arrive to the simple formula for $\hat{p}$, $\hat{p}_r = \sigma_{Y_r}^{-2}/\sum_r \sigma_{Y_r}^{-2}$, with the estimate of the signal given as the weighted average with the weights equal to the inverse variance of the estimates in the window:

$$\mu(x) = \mu_{\tilde{Y}} = \frac{\sum_r \sigma_{Y_r}^{-2} \mu_{Y_r}(x)}{\sum_r \sigma_{Y_r}^{-2}}. \qquad (17)$$

4.1 Partial window-wise estimates

For calculation of the regression approximations for the conditional means μ_{Y_r}, the Bayesian formula gives $f_r(Y_r|Z_r) = f_r(Z_r|Y_r) f_r(Y_r)/f_r(Z_r)$, where $f_r(Y_r)$ is a prior density for Y_r and $f_r(Z_r|Y_r)$ is the conditional density of Z_r provided a fixed Y_r. Using the parametric model (5) for Y_r, $Y_r = D_r^T \theta_r$, and the exponential prior for the parameters θ_r, we arrive to the posterior density in the form

$$f_r(Y_r|Z_r) \propto \exp\left(-\frac{\|Z_r - Y_r\|^2}{2\sigma^2} - pen(\theta_r)\right).$$

Inserting these $f_r(Y_r|Z_r)$ into (13), the maximum likelihood gives the estimates calculated independently for each window:

$$\hat{\theta}_r = \underset{\theta_r}{\arg\min}\left(\frac{\|Z_r - D_r^T \theta_r\|^2}{2\sigma^2} + pen(\theta_r)\right), \qquad (18)$$
$$\hat{Y}_r = D_r^{-1}\hat{\theta}_r, \ y_r(x) = \hat{Y}_r(x).$$

It is usually assumed that the penalty $pen(\theta_r)$ is additive with respect to the components of the vector θ_r, $pen(\theta_r) = \sum_{i=1}^{n_r} pen(\theta_{r,i})$. Then, the minimization in (18) can be solved as the scalar optimization independently for each $\theta_{r,i}$:

$$\hat{\theta}_{r,i} = \underset{\theta_{r,i}}{\arg\min}\left(\frac{((D_r Z_r)_i - \theta_{r,i})^2}{2\sigma^2} + pen(\theta_{r,i})\right), \qquad (19)$$

where $(\cdot)_i$ is the i-th element of the vector. Depending on the penalty criterion $pen(\theta_{r,i})$, the problem (19) defines a different kind of filtering procedures going under the generic name *thresholding*. A number of threshold operators derived from the optimization problem (19) for different penalty functions can be found in (Elad, 2006).

In this form, the derived algorithm coincides with the procedure used in (Foi *et al.*, 2006), (Foi *et al.*, 2007a), (Foi *et al.*, 2007b). Thus, we show that, under the proposed randomized inverse windowing (10), these algorithms can be seen as optimal ones. It is clear from the above formulas that in implementation we do not need to imitate the random selection of the windows,

as the average over this randomness is calculated explicitly with the final estimation formula (17) where the weights of the partial estimates are given.

5. GROUPING BY WINDOW-MATCHING

Let us consider again the windowing (covering) $\mathcal{C} = \{X_r, r = 1, \ldots, N_s\}$ from Section 2, further assuming that all windows are equal upon translation. We also assume that there is a similarity between some of the vectors Y_r, e.g., their elements take nearly identical or close values, and that these vectors can be *grouped* according to their closeness. For simplicity, the grouping rule is defined by the ℓ^2 norm as

$$||Y_{r'} - Y_r||^2 \leq \Delta, \; r, r' = 1, \ldots, N_s, \;\; \Delta > 0. \quad (20)$$

The aim of the grouping is a joint processing of the windowed data in the group.

5.1 Reference-grouping (matching)

Similar to (Dabov *et al.*, 2007), in the grouping we consider a window to be as the "reference" window X_R, with reference data Y_R. The corresponding group G_R is a collection of the windowed data similar to this reference one:

$$G_R = \{Y_r : ||Y_R - Y_r||^2 \leq \Delta, r = 1, \ldots, N_s, \; \Delta > 0\}. \quad (21)$$

Let R take values from 1 to N_s, i.e. each of the windows is treated as a reference one. The rule (21) defines the groups of the similar windowed data, for each of the windows covering the grid X.

A link between this composed grouped windows and the original signal-vector y can be given in the form (2) with the generalized inverse similar to (3). Using for $B^{\#}$ the Moore-Penrose generalized inverse, we obtain the reconstruction of the true $y(x)$ as the sample average of the windowed data estimates whose corresponding windows contain x, which estimates are produced by filtering different groups. We remark that it is not necessary that $x \in X_R$ for G_R to contribute to the estimation of $y(x)$, since there can be some other $r \neq R$ such that $Y_r \in G_R$ and $x \in X_r$.

In what follows, we use the randomized modeling resulting in the final estimate that is the weighted average of the multiple estimates obtained in different groups.

5.2 Randomized de-grouping

The randomized de-grouping models the reconstruction of the original signal $y(x)$ as a random procedure where the window is randomly selected from all windows containing x among the all groups in which such a window can be found. Let us use the notation $X_{r,k}$ for the window X_r located in k-th group. In our

procedure we randomly select one of these groups and randomly extract the corresponding window containing the point x.

Define a random window $\tilde{X}$ as

$$\tilde{X} = \bigcup_{r,k:\alpha_{r,k}=1} X_{r,k} = \bigcup_{r:\alpha_{r,k}=1} X_r,$$

where $\alpha_{r,k}$ are binary (zero-one) random variables equal to 1 with probability $p_{r,k}$, with $\sum p_{r,k} = 1$. As in Section 4, the variables $\alpha_{r,k}$ are random indicators such that only one of the variables $\alpha_{r,k}$ takes value 1. The corresponding randomly windowed true values and observations are

$$\tilde{Y} = \sum_{r,k} Y_{r,k}\alpha_{r,k} = \sum_{r,k} Y_r \alpha_{r,k},$$

$$\tilde{Z} = \sum_{r,k} Z_{r,k}\alpha_{r,k} = \sum_{r,k} Z_r \alpha_{r,k}.$$

It is clear that $\tilde{Y}(x) = y(x)$, i.e. this procedure gives the perfect reconstruction of the true signal from its grouped true values. Assuming that y is a random variable, the probability density for the randomly windowed data $\tilde{Y}$ has the form of the mix distribution

$$f(\tilde{Y}) = \sum_{r,k} f_{r,k}(Y_{r,k})p_{r,k}, \quad (22)$$

where $f_{r,k}(Y_{r,k})$ are the densities for the windowed data $Y_{r,k}$. A similar representation is valid for the conditional distribution

$$f(\tilde{Y}|z) = \sum_{r,k} f_{r,k}(Y_{r,k}|z)p_{r,k}. \quad (23)$$

Assuming, as it has been used in the proof of formula (14), that $f_k(Y_k|z) = f_k(Y_k|Z_k)$, we find that $f(\tilde{Y}|z) = \sum_{r,k} f_{r,k}(Y_{r,k}|Z_{r,k})p_{r,k}$ and

$$\mu(x) = \sum_{r,k} \mu_{r,k}(x) \cdot p_{r,k}, \quad (24)$$

where $\mu(x) = \mu_{\tilde{Y}}(x)$, $\mu_{\tilde{Y}} = E\{\tilde{Y}|z\} = \int \tilde{Y}f(\tilde{Y}|z)d\tilde{Y}$, $\mu_{r,k}(x) = \mu_{\tilde{Y}_{r,k}}(x)$, and $\mu_{\tilde{Y}_{r,k}} = E\{Y_{r,k}|Z_{r,k}\}$. The formula (24) shows that the least-squares estimate is calculated as a weighted sum of the least-squares estimates obtained for the point x in different windows belonging to different groups.

5.3 Group-wise estimation

Using the parametric model (5) for Y_r, $Y_r = D_r^T \theta_r$, and the exponential prior for the parameters θ_r, similar to those introduced for the above window-wise estimates, we arrive to the minus log posterior density

$$-\log f(Y_r|Z_r) \propto \frac{||D_r Z_r - \theta_r||^2}{2\sigma^2}.$$

Let $\Theta_k = \{\theta_r\}_{r \in G_k}$ be a set of spectrums $\theta_r = D_r Z_r$ of the windows in the group k. Consider the elements of Θ_k as a function of the index r and apply some

transform Υ with respect to this index. This new group-wise spectrum is $\Omega_k = \Upsilon(\Theta_k)$. Let the penalty be defined as restrictions on the elements of the group array Ω_k. Then, the thresholding of Ω_k and estimation of $\tilde{\Omega}_k$ are defined as

$$\hat{\Omega}_k = \underset{\Omega}{\arg\min}\left(\frac{||\tilde{\Omega}_k - \Omega||_F^2}{2\sigma^2} + pen(\Omega)\right), \qquad (25)$$

$$\hat{\Theta}_k = \Upsilon^{-1}\left(\hat{\Omega}_k\right), \; \hat{Y}_{r,k} = D_k^{-1}\hat{\theta}_{r,k}, \; y_{r,k}(x) = \hat{Y}_{r,k}(x).$$

Here, $|| \cdot ||_F$ stays for the Frobenius matrix norm. If the penalty $pen(\theta_k)$ is additive with respect to the components of the matrix Ω, the minimization in (25) is scalar and independent for each element of Ω. The inverse transform returns the estimates of $\{\theta_r\}_{r\in G_k}$. Because they are different in different groups, we use the double indexes for the estimates $\hat{\theta}_{r,k}$, $\hat{Y}_{r,k}$, $y_{r,k}(x)$.

5.4 Estimate fusing

Using (24) we obtain for the signal's estimate

$$\hat{y}(x) = \sum_{r,k} \hat{y}_{r,k}(x) \cdot p_{r,k}, \qquad (26)$$

$$\hat{y}_{r,k}(x) = Y_{r,k}(x).$$

This estimate can be optimized by selection of the appropriate values for $p_{r,k}$ as it is in (17). The simplified version is to take these parameters to be equal for all windows in the group, i.e. $p_{r,k}$ does not depend on r and assumes that the group estimates are independent. Then, $\hat{p}_k = \sigma_{\Theta_k}^{-2} / \sum_{r,k} \sigma_{\Theta_k}^{-2}$, where $\sigma_{\Theta_k}^2$ are the estimate variances of the groups.

The derived algorithm consists of three successive steps: grouping; spectrum filtering with calculation of the window-wise estimates $\hat{y}_{r,k}(x)$; and fusing of these estimates into the final one, as it is done in (26). It coincides with the block-matching collaborative filtering algorithm proposed in (Dabov *et al.,* 2007). In this way, we prove that this algorithm yields a close approximation of the least-square optimal estimate, provided that the image model can be given in the form of the mix probability density (23).

6. CONCLUSION

In this paper we focus on de-windowing and de-grouping of the overcomplete data as an important element of formulation of the denoising problem, because they influence both the estimates of the spectrums and the fusing of the window-wise estimates into the final one. We propose a randomized de-windowing that leads to a mix distribution image modeling and the weighted-averaging of the window-wise estimates. Simulation experiments presented in the papers (Foi *et al.*, 2006), (Foi *et al.*, 2007a), (Foi *et al.*, 2007b), and (Dabov *et al.*, 2007)

prove that with this sort estimation composed with the proper de-windowing and de-grouping is highly efficient. In these cited papers one can find full details concerning these algorithms. The MATLAB codes of the algorithms from (Foi *et al.*, 2007a) and (Dabov *et al.*, 2007) are available from the websites `http://www.cs.tut.fi/~foi/SA-DCT` and `http://www.cs.tut.fi/~foi/GCF-BM3D` .

REFERENCES

Coifman, R. R., and D. Donoho (1995). Translation-invariant de-noising. In: *Wavelets and Statistics* (A. Antoniadis and G. Oppenheim (Eds.)), Lecture Notes in Statistics, Springer-Verlag, 125-150.

Dabov K., Foi, A., V. Katkovnik, and K. Egiazarian (2007). Image denoising by sparse 3D transform-domain collaborative filtering. *IEEE Trans. Image Process.*, **16**, no. 8, (in press).

Elad M. (2006). Why shrinkage is still relevant for redundant representations? *IEEE Trans. Inf. Theory*, **52**, no. 12, 5559-5569.

Foi A., R. Bilcu, V. Katkovnik, and K. Egiazarian (2006). Adaptive-size block transforms for signal-dependent noise removal. *Proc. 7th Nordic Signal Processing Symposium, NORSIG 2006*, Reykjavik, Iceland.

Foi A., V. Katkovnik, and K. Egiazarian (2007a). Pointwise shape-adaptive DCT for high-quality denoising and deblocking of grayscale and color images. *IEEE Trans. Image Process.* **16**, no. 5, 1395-1411.

Foi, A., V. Katkovnik, and K. Egiazarian (2007b). Signal-dependent noise removal in pointwise shape-adaptive DCT domain with locally adaptive variance, to appear in *Proc. 15th European Signal Process. Conf., EUSIPCO 2007*, Poznan.

Guleryuz O. G. (2006). Weighted overcomplete denoising. In review, preprint available online at `http://eeweb.poly.edu/~onur` .

Hua, G. and M. T. Orchard (2003). A new interpretation of translation invariant image denoising. *Proc. Asilomar Conf. Signals Syst. Comput.*, Pacific Grove, CA, 332- 336.

Öktem, R., L. Yaroslavsky, K. Egiazarian and J. Astola (1999) Transform based denoising algorithms: comparative study. Tampere University of Technology, 1999.

Öktem, H., V. Katkovnik, K. Egiazarian, and J. Astola (2001). Local adaptive transform based image de-noising with varying window size. *Proc. IEEE Int. Conf. Image Process., ICIP 2001,* Thessaloniki, Greece, 273-276.

Yaroslavsky, L., K. Egiazarian, and J. Astola (2001). Transform domain image restoration methods: review, comparison and interpretation. *Proc. SPIE* **4304 -** *Nonlinear Image Process. Pattern Anal. XII*, San Jose, CA, 155-169.