Nonlocal Collaborative l_0-Norm Prior for Image Denoising

Vladimir Katkovnik
Department of Signal Processing,
Tampere University of Technology, Finland
vladimir.katkovnik@tut.fi

Abstract

Spatially adaptive nonparametric regression estimation is one of the most promising recent directions in image processing. The Transforms and Spectral Techniques Research Group at the Department of Signal Processing, Tampere University of Technology, has been active in this novel field starting from about 2002. The results achieved with application to different image and video processing problems are very positive and completely support optimism following from general speculations concerning nonparametric modeling (e.g. [1]-[9]). Within this framework the Block Matching and 3-D Filtering (BM3D) algorithm has been developed which is currently one of the best performing denoising algorithms. In this paper a special prior is proposed allowing to reformulate mainly semi-heuristic nonlocal nonparametric techniques as global minimization of an energy criterion. It is shown that the basic hard-thresholding part of the BM3D algorithm can be derived as a minimizer of the proposed prior. The outstanding performance of BM3D is a strong argument in favor of this prior as an efficient multilayer redundant image model.

1 Introduction

Suppose we have independent random observation pairs $\{z_i, x_i\}$ given in the form

$$z_i = y_i + \varepsilon_i, \tag{1}$$

where $y_i = y(x_i)$ is a signal of interest, $x_i \in \mathbb{R}^2$ denotes a vector of "features" or explanatory variables which determines the signal observation y_i, and $\varepsilon_i = \varepsilon(x_i)$ is an additive noise, $\varepsilon_i \sim N(0, \sigma^2)$. The denoising problem is to reconstruct $y(x)$ from $\{z_i\}$.

A variety of denoising methods are derived by considering image processing as a variational problem where the restored image is computed by minimization of an energy functional. Typically, such functionals consist of the fidelity term calculated as the Eucledian norm of the difference between the true image and the observed noisy image and the regularization penalty:

$$\hat{y} = \arg\min_y ||y - z||_2^2/\sigma^2 + \lambda \cdot pen(y). \tag{2}$$

If $\lambda = 0$, the solution of (2) is trivial $\hat{y} = z$. This optimal estimate is equal to the noisy signal and denoising is not produced. It demonstrates that the filtering ability of the optimal solution is completely defined by the penalty pen(y) called also an image prior.

For imaging, the variational formulation (2) has been introduced in terms of *Markov Random Field* (*MRF*) modeling with Bayesian estimation and Gibbs distribution as a prior [10]. The corresponding prior probability density has an exponential form $p(y) = \exp(-U(y)/T)/Z$, where $U(y)$ is *"potential function"* and T is "*temperature*", the latter term is inherited from statistical mechanics, where this distribution has been introduced by J. Willard Gibbs in 1878. Then the maximum likelihood leads to

the penalty $pen(y) = U(y)$. Depending on the problem at hand $U(y)$ is defined in various ways and $U(y)$ calculated over *"cliques"*, sets of pixels close to each other in some sense.

In imaging , the variational approach is very popular with the penalty functions introduced using different arguments varying from strong statistical-mathematical to pure heuristical ones.

For illustration we mention some of the penalty functions in common use:

- *Quadratic penalties*
$$\text{pen}(y) = ||y||_2^2, \ \text{pen}(y) = ||Ly||_2^2 \tag{3}$$
have a form of the quadratic Eucledian norm of y or of a linear functional of y. These penalties are popular in the standard Tikhonov's regularizators [11];

- *Total variation (ROF)* [12], [13]
$$\text{pen}(y) = \int ||\Delta y||_2 dx, \tag{4}$$
where Δy is a vector-gradient of y. The success of this penalty stems from the fact that it allows discontinuous solutions and hence preserves edges while filtering out high-frequency oscillations due to noise;

- *Products of experts (PoE) and field of experts (FoE)* [14], [15]
$$\text{pen}(y_k) = \sum_{i=1}^{N} \alpha_i \log(1 + \frac{1}{2}(f_i^T y_k)^2), \tag{5}$$
$$\text{pen}(y) = \sum_{k} \sum_{i=1}^{N} \alpha_i \log(1 + \frac{1}{2}(f_i^T y_k)^2), \tag{6}$$
where y_k are vectorized image patches projected on f_i, where f_i are analysis filters, α_i and N are parameters;

- *Nonlocal means* [16], [17],
$$\text{pen}(y) = \int g\left(\frac{|y(x) - y(v)|^2}{h^2}\right) w(|x - v|) dx dv, \tag{7}$$
where $w > 0$ is a window, and g is a differentiable function;

- *Complexity penalty* is formulated usually for spectrum representations of the image as $\theta = \mathcal{T}\{y\}$, where $\mathcal{T}$ stands for orthonormal or overcomplete transforms (e.g. [6], [18]). This penalty is calculated as
$$pen(\theta) = ||\theta||_{l_0}, \tag{8}$$
where the l_0-norm gives a number of active spectrum elements different from zero. The l_1-norm is often used as a replacement for the l_0-norm as giving a close solution and computationally much more efficient.

1.1 Overcomplete transform domain modeling

Let the signals from (1) be given in the matrix form as Y and Z and defined on the regular 2-D grid X. Following [1] consider a windowing $\mathcal{C} = \{X_r, r = 1, \ldots, N_s\}$ of X with N_s blocks (uniform windows) $X_r \subset X$ of size $n_r \times n_r$ such that $\cup_{r=1}^{N_s} X_r = X$. Mathematically speaking, this windowing is a *covering*

of X. Thus, each $x \in X$ belongs to at least one subset X_r. The noise-free data Y and the noisy data Z windowed on X_r are arranged in $n_r \times n_r$ blocks denoted as Y_r and Z_r, respectively. Typically, the blocks are overlapping and therefore some of the elements may belong to more than one block.

We use transforms (orthonormal series) of pixels in the blocks in conjunction with the concept of the redundancy of natural signals. Mainly these are orthogonal polynomials, discrete Fourier, cosine and wavelet transforms. The transform, denoted as $\mathcal{T}_r^{2\mathrm{D}}$, is applied for each window X_r independently as

$$\theta_r = \mathcal{T}_r^{2\mathrm{D}}\,(Y_r)\,, \qquad \left[\ = D_r Y_r D_r^T\ \right] \qquad r = 1, \ldots, N_s, \tag{9}$$

where θ_r is the spectrum of Y_r. The equality enclosed in square brackets holds when the transform $\mathcal{T}_r^{2\mathrm{D}}$ is realized as a separable composition of 1-D transforms, each computed by matrix multiplication against an $n_r \times n_r$ orthogonal matrix D_r. The inverse $\mathcal{T}_r^{2\mathrm{D}\,-1}$ of $\mathcal{T}_r^{2\mathrm{D}}$ defines the signal from the spectrum as

$$Y_r = \mathcal{T}_r^{2\mathrm{D}\,-1}\,(\theta_r)\,, \qquad \left[\ = D_r^T \theta_r D_r\ \right] \qquad r = 1, \ldots, N_s.$$

The noisy spectrum of the noisy signal is defined as

$$\tilde{\theta}_r = \mathcal{T}_r^{2\mathrm{D}}\,(Z_r)\,, \qquad \left[\ = D_r Z_r D_r^T\ \right] \qquad r = 1, \ldots, N_s. \tag{10}$$

The signal y is *sparse* if it can be well approximated by a small number of non-zero elements of the spectrum θ_r. The number of non-zero elements of θ_r, denoted using the standard notation as $||\theta_r||_{l_0}$, is interpreted as the complexity of the model in the block.

If the blocks are overlapping the total number of the spectrum elements θ_r, $r = 1, \ldots, N_s$, is larger (much larger) than the image size and we arrive to the *overcomplete* or *redundant* data approximation. This redundancy is an important element of the efficiency of this modeling overall.

The blockwise estimates are simpler for calculation than the estimates produced for the whole image because the blocks are much smaller than the whole image. This is a computational motivation for the blocking. Another even more important point is that the blocking imposes a localization of the image on small pieces where simpler models may fit the observations.

The data windowing can be produced in many different ways. In deterministic non-adaptive design, fixed-size square windows cover the image entirely. One example of this sort of windowing is the sliding windowing where to each pixel in the image a window is assigned having this pixel as, say, its upper-left corner (e.g., [9], Ch. 5).

1.2 Estimation

For the white Gaussian noise in the observation model (9)-(10), the penalized minus log-likelihood maximization gives the estimates of θ_r as

$$\begin{aligned} \hat{\theta}_r &= \operatorname*{argmin}_{\vartheta} ||Z_r - \mathcal{T}_r^{2\mathrm{D}\,-1}(\vartheta)||_2^2/\sigma^2 + \lambda\,\mathrm{pen}(\vartheta), \\ \hat{Y}_r &= \mathcal{T}_r^{2\mathrm{D}\,-1}(\hat{\theta}_r), \end{aligned} \tag{11}$$

where ϑ is a matrix of the size of Z_r, $\mathrm{pen}(\vartheta) = ||\vartheta||_{l_0} = \sum_{k,l} 1(\vartheta(k,l) \neq 0)$, $1(\vartheta(k,l) \neq 0) = 1$ if $\vartheta(k,l) \neq 0$ and $1(\vartheta(k,l) = 0) = 0$, and $\lambda > 0$ is a parameter that controls the trade-off between the penalty and the fidelity.

The spectrum penalty is used for characterizing the model complexity and appears naturally in the Bayesian interpretation of this modeling, provided that the spectrum ϑ is random with a prior density $p(\vartheta) \propto e^{-\lambda\,\mathrm{pen}(\vartheta)}$. The estimator (11) can be presented in the following equivalent form

$$\hat{\theta}_r = \operatorname*{argmin}_{\vartheta} ||\tilde{\theta}_r - \vartheta||_2^2/\sigma^2 + \lambda\,\mathrm{pen}(\vartheta), \tag{12}$$

where the noisy spectrum is calculated as (10).

Because the penalty is additive for the items of ϑ the problem in (10) can be solved independently for each element of the $\hat{\theta}_r$ as a scalar optimization problem:

$$\hat{\theta}_r(k,l) = \underset{x \in R^1}{\operatorname{argmin}} \left(\tilde{\theta}_r(k,l) - x)\right)^2 / \sigma^2 + \lambda 1(|x| > 0). \tag{13}$$

This solution has a form of the hard-thresholding

$$\hat{\theta}_r(k,l) = \rho(\tilde{\theta}_r(k,l), \lambda\sigma) \triangleq \tilde{\theta}_r(k,l) \cdot 1\left(|\tilde{\theta}_r(k,l)| > \sigma\sqrt{\lambda}\right), \tag{14}$$

i.e. $\hat{\theta}_r(k,l) = 0$ if $|\tilde{\theta}_r(k,l)| \le \sigma\sqrt{\lambda}$ and $\hat{\theta}_r(k,l) = \tilde{\theta}_r(k,l)$ if $|\tilde{\theta}_r(k,l)| > \sigma\sqrt{\lambda}$.

The corresponding estimates for the pixels of the block are $\hat{Y}_r = \mathcal{T}_r^{2\mathrm{D}\,-1}(\hat{\theta}_r)$.

1.3 Aggregation of windowed estimates

At the points where the windows overlap, multiple estimates appear. Then, the final estimate for each x is calculated as the sample mean or the weighted mean of these multiple estimates [1]:

$$\hat{y}(x) = \frac{\sum_r \mu_r \hat{y}_r(x)}{\sum_r \mu_r \chi_{X_r}(x)},\ x \in X, \tag{15}$$

where $\hat{y}_r$ is obtained by returning the window-wise (multipoint) estimates $\hat{Y}_r$ to the respective place X_r (and extending it as zero outside X_r), μ_r are the weights used for these estimates, and χ_{X_r} is the indicator function (characteristic function) of the set X_r.

Although in many works equal weights $\mu_r = 1\ \forall r$ are traditionally used it is a well established fact that the efficiency of the aggregated estimates (15) sensibly depends on the choice of the weights.

In particular, using weights μ_r inversely proportional to the variances of the corresponding estimates $\hat{y}_r$ is found to be a very effective choice, leading to the dramatic improvement of the accuracy of estimation (e.g. [3], [8]). In [19] this sort of effects are studied for different weights for aggregating blockwise estimates from sliding window DCT and demonstrated essential improvements of the algorithms.

In [18], Elad and Aharon derive an optimal estimator for the windowed data (9)-(10) as a minimizer of the global energy criterion:

$$\mathcal{E} = \frac{1}{\sigma^2}||Z - Y||_2^2 + \sum_r \left(||Y_r - \mathcal{T}_r^{2\mathrm{D}\,-1}(\vartheta_r)||_2^2 + \lambda\,\mathrm{pen}(\vartheta_r)\right).$$

The algorithm proposed in [18] uses the alternative minimization on Y and ϑ_r and defines the spectrum estimates at the first step as

$$\tilde{\theta}_r = \arg\min_{\vartheta_r} ||Z_r - \mathcal{T}_r^{2\mathrm{D}\,-1}(\vartheta_r)||_2^2 + \lambda\,\mathrm{pen}(\vartheta_r).$$

Given $\tilde{\theta}_r$ the signal estimate of Y is calculated as

$$\hat{Y} = \arg\min_{Y} \frac{1}{\sigma^2}||Z - Y||_2^2 + \sum_r ||Y_r - \hat{Y}_r||_2^2,$$
$$\hat{Y}_r = \mathcal{T}_r^{2\mathrm{D}\,-1}\left(\tilde{\theta}_r\right).$$

Repeating this procedure we arrive to the recursive algorithm

$$\begin{aligned}
&\tilde{\theta}_r^{(t)} = \arg\min_{\vartheta_r} ||\hat{Y}_r^{(t)} - \mathcal{T}_r^{2\mathrm{D}\,-1}(\vartheta_r)||_2^2 + \lambda\,\mathrm{pen}(\vartheta_r), \\
&\hat{Y}_r^{(t)} = \mathcal{T}_r^{2\mathrm{D}\,-1}\left(\tilde{\theta}_r^{(t)}\right), \\
&\hat{Y}^{(t+1)} = \arg\min_{Y} \frac{1}{\sigma^2}||Z - Y^{(t)}||_2^2 + \sum_r ||Y_r - \hat{Y}_r^{(t)}||_2^2, \\
&\hat{Y}_r^{(1)} = Z_r,\ t = 1, \ldots.
\end{aligned} \tag{16}$$

The third formula in (16) defines the aggregation of the windowed signal estimates $\hat{y}_r^{(t)}(x)$ and can be rewritten as their sample mean:

$$\hat{y}^{(t+1)}(x) = \frac{z(x)/\sigma^2 + \sum_r \hat{y}_r^{(t)}(x)}{1/\sigma^2 + \sum_r \chi_{X_r}(x)}, \qquad x \in X. \tag{17}$$

The optimal estimator minimizing a global energy criterion can be achieved as a limit of this recursive procedure. However, as it is mentioned above, the sample mean (17) is not a good aggregation formula. It means that the recursive energy minimization used for the windowed estimates results in a procedure which can be essentially improved.

Indeed, the very good denoising results shown in [18] are obtained mainly due to combining the recursive procedure (16) with a "*dictionary update*" stage (K-SVD algorithm [20]).

As overcomplete estimation with multiple estimates for each pixel demonstrates high-efficiency, the aggregation of these estimates becomes a hot topic because of two different reasons. The first one is pragmatic, what is the best way to aggregate, and the second one is principal, why the aggregation can be so efficient.

2 Nonlocal transform domain modeling

2.1 Group-wise and global penalty

Following Section 1.1, we consider the signal Y_j and observation Z_j blocks corresponding to a given windowing. The transforms are defined and calculated for these blocks. Furthermore, it is assumed that there is a similarity between some of these blocks and the similar blocks are clustered in *"groups"*. As a measure of this similarity between the blocks r and j we use the Eucledian norm $||Y_j - Y_r||_2^2$ [2]:

$$w_h(r,j) = 1(||Y_r - Y_j||_2^2 \le h). \tag{18}$$

These binary weights $w_h(r,j)$ take value 1 if the Eucledian distance is smaller or equal to h, then the block j belongs to the group r. Otherwise if the Eucledian distance is larger than h then, $w_h(r,j) = 0$ and the block j is not included in the group r.

We introduce the penalty first *locally* for the group and further *globally* for the whole image. The penalty for the rth group is defined as

$$\begin{aligned}
&pen_r(\{\vartheta_{r,j}\}_j) = \\
&= \left(\sum_j w_h(r,j)||\theta_j - \vartheta_{r,j}||_2^2\right) + \lambda_r||\{\vartheta_{r,j}\}_j||_{l_0}.
\end{aligned} \tag{19}$$

Here $\{\vartheta_{r,j}\}_j$ is a set of the models for all jth blocks included in the rth group, and $\{\theta_j\}_j$ is a set of the spectrums of the true signal blocks in this group. This *group-wise penalty* model has been proposed and developed in [1].

In this paper we go further and introduce the *global* penalty as the weighted mean of the *group-wise local* penalties (19):

$$PEN(\{\vartheta_{r,j}\}_{r,j}) = \sum_r g_r pen_r(\{\vartheta_{r,j}\}_j) = \tag{20}$$

$$\sum_r g_r \left(\sum_j w_h(r,j) ||\theta_j - \vartheta_{r,j}||_2^2 + \lambda_r ||\{\vartheta_{r,j}\}_j||_{l_0} \right),$$

with the group-weights g_r calculated as

$$g_r = \frac{1/||\{\vartheta_{r,j}\}_j||_{l_0}}{\sum_r 1/||\{\vartheta_{r,j}\}_j||_{l_0}}. \tag{21}$$

The group-wise penalty characterizes the quality of the rth group, where the accuracy of the spectrum approximations as well as the complexity of these approximations are taken into consideration.

In the global penalty the *group-wise* ones are weighed with the weights inversely proportional to the complexity of the group-wise models. This rule perfectly corresponds to the idea of the sparse image modeling when a low complexity model is the main goal. According to this idea the low complexity groups are preferable and taken in (20) with larger weights.

For the white Gaussian noise the image denoising using the global penalty (20) is formalized as the following optimization problem:

$$\begin{aligned} \hat{Y} &= \arg \min_{Y, \{\vartheta_{r,j}\}_{r,j}} J, \\ J &= ||Z - Y||_2^2/\sigma^2 + \mu \cdot PEN(\{\vartheta_{r,j}\}_{r,j}), \end{aligned} \tag{22}$$

where μ defines a balance between the fidelity $||Z - Y||_2^2/\sigma^2$ and the penalty $PEN(\{\vartheta_{r,j}\}_{r,j})$.

2.2 Collaborative block-wise and global penalty

It is demonstrated in [2] that a much higher sparsity of the signal representation can be achieved using a 3D group-wise transform instead of 2D block-wise transforms (with spectrums $\vartheta_{r,j}$) as it is in (22). This sparsity in the 3D transform space improves the efficiency of filtering and implemented in [2] as *collaborative filtering*.

We are going to use these 3D collaborative transforms for the introduced global penalty.

Let $\Theta_r^Y = \{\theta_{r,j}\}_{j \in K_r^\Delta}$ be a collection of the $2D$ block-wise spectrums treated as 3-D array, where j is the index used for the third dimension. We will denote the elements of the $3D$ array Θ_r^Y as $\Theta_{r,j}^Y(k,l)$, where the indices (k,l) concern $2D$ array of the jth block in the rth group. Apply a $1D$ orthonormal transform $\mathcal{T}^{1D}$ with respect to j. In this way we arrive to a group-wise $3D$ spectrum of the signal Y in the rth group as

$$\Omega_r^Y = \mathcal{T}^{1D}(\Theta_r^Y). \tag{23}$$

Following [1], [2] we replace the $2D$ spectrum-estimates $\{\vartheta_{r,j}\}_{j \in K_r^h}$ with the corresponding $3D$ spectrum $\Omega = \mathcal{T}^{1D}(\{\vartheta_j\}_{j \in K_r^\Delta})$ obtained by applying the $1D$ transform $\mathcal{T}^{1D}$ on the collection of $2D$ spectra $\{\vartheta_j\}_{j \in K_r^\Delta}$. Then, the l_0-norm $||\{\vartheta_{r,j}\}_{j \in K_r^h}||_{l_0}$ in (20) is replaced with the equivalent norm in this $3D$ spectrum space defined as $||\Omega||_0 = \sum_{k,l,j \in K_r^\Delta} 1(\Omega_j(k,l) \neq 0)$.

This $3D$ spectrum representation is used as a joint *collaborative* model of the signal clustered in the rth group. For this group the group-wise penalty (19) takes the form

$$pen_r(\Omega) = ||\Omega_r^Y - \Omega||_2^2 + \lambda_r||\Omega||_{l_0}. \tag{24}$$

Recall again, that here Ω is the $3D$ array of the spectrum approximations (estimates) we are looking for, and Ω_r^Y (with index Y) is the spectrum of the blocks of the true signal values Y_j collected into the rth group according to the rule $\{Y_j\}_{j\in K_r^\Delta}$.

Then the global penalty (20) takes the form

$$\begin{aligned} &PEN(\{\Omega_r\}_r) = \sum_r g_r \cdot pen_r(\Omega) = \\ &\sum_r g_r \left(||\Omega_r^Y - \Omega_r||_2^2 + \lambda_r||\Omega_r||_{l_0}\right), \\ &g_r = \frac{1/||\Omega_r||_{l_0}}{\sum_r 1/||\Omega_r||_{l_0}}, \end{aligned} \tag{25}$$

where the spectrum Ω_r is an estimate for the spectrums Ω_r^Y in the rth group, and $\lambda_r||\Omega_r||_{l_0}$ is the l_0-norm penalty for this estimate.

2.3 Nonlocal energy minimization (NEM)

Using the spectrum representation for the signals and passage to the $3D$ spectrum as well as the l_0-norm global penalty (25) defined for this $3D$ space the estimation problem (22) is reformulated as follows

$$\begin{aligned} \hat{Y} &= \arg\min_{Y,\{\Omega_r\}_r} J, \\ J &= ||Z - Y||_2^2/\sigma^2 + \mu \cdot PEN(\{\Omega_r\}_r). \end{aligned} \tag{26}$$

This estimator is an essential development of the well known nonlocal means algorithms [21], [22], the algorithms with the block-matching [23] as well as the basic concepts imbedded in the nonlocal collaborative filtering [2], [1].

Let us consider the alternative minimization of J on $\{\Omega_r\}_r$ and Y assuming that the weights g_r are fixed.

If Y is given minimization on $\{\Omega_r\}_r$ concerns the penalty term $PEN(\{\Omega_r\}_r)$ only, and it is reduced to scalar calculations independent for each element of Ω_r, because $\min_{\{\Omega_r\}_r} J \Longrightarrow \min_\Omega pen_r(\Omega)$ and further

$$\hat{\Omega}_r(k,l) = \arg\min_{x\in R^1} \left(\Omega_r^Y(k,l) - x)^2 + \lambda_r \cdot 1(x \neq 0)\right).$$

According to (14) this solution is the hard-thresholding of $\Omega_r^Y(k,l)$ calculated as

$$\hat{\Omega}_r(k,l) = \rho(\Omega_r^Y(k,l), \sqrt{\lambda_r}). \tag{27}$$

When $\hat{\Omega}_r(k,l)$ are found the signal estimates are calculated as

$$\begin{aligned} \hat{\Theta}_r &= \{\hat{\theta}_{r,j}\}_{j\in K_r^\Delta} = \mathcal{T}^{1D\,-1}\left(\hat{\Omega}_r\right), \\ \hat{Y}_{r,j} &= \mathcal{T}^{2D\,-1}\left(\hat{\theta}_{r,j}\right). \end{aligned} \tag{28}$$

The consecutive $\mathcal{T}^{1D\,-1}$ and $\mathcal{T}^{2D\,-1}$ inverse transforms return first the estimates $\hat{\Theta}_r = \{\hat{\theta}_{r,j}\}_{j\in K_r^{\Delta}}$ of $\mathcal{T}^{2D}$-spectra of the blocks in the group, and hence the multipoint estimates $\hat{Y}_{r,j}$ of these blocks. Because these estimates can be different in different groups, we use the double indexes for the signal estimates $\hat{Y}_{r,j}$, where j stays for the index of the block and r for the group where these estimates are obtained.

Consider minimization in (26) on Y provided $\{\Omega_r\}_r$ are given as $\left\{\hat{\Omega}_r\right\}_r$. The spectrums Ω_r^Y depend on Y and this dependence should be taken into considerations in minimization on Y. It is convenient to give this solution for the penalty in the form (20) where the spectrums in the quadratic norms are replaced by the corresponding signals

$$PEN(\{\vartheta_{r,j}\}_{r,j}) = \tag{29}$$
$$= \sum_r g_r \left(\sum_j w_h(r,j) ||Y_j - \hat{Y}_{r,j}||_2^2 + \lambda_r ||\{\vartheta_{r,j}\}_j||_{l_0} \right).$$

Let us use for the signals given by the matrices Y, Z, Y_j, $\hat{Y}_{r,j}$ the lexicographical vector representations and use for this vectors the corresponding bold letters $\mathbf{Y}$, $\mathbf{Z}$, $\mathbf{Y}_j$, $\hat{\mathbf{Y}}_{r,j}$. The vectors $\mathbf{Y}_j$ are the parts (projections) of the vectors $\mathbf{Y}$ and they can be defined through projection matrices $\mathbf{P}_j$ of the corresponding sizes, $\mathbf{Y}_j = \mathbf{P}_j\mathbf{Y}$. Note that $\mathbf{P}_j$ is a binary matrix with items (0,1).

Using this vector-matrix notation and the penalty in the form (29) the criterion J in (22) can be represented as

$$J = ||\mathbf{Z} - \mathbf{Y}||_2^2/\sigma^2 + \mu \cdot \sum_r g_r \left(\sum_j w_h(r,j) ||\mathbf{P}_j\mathbf{Y} - \hat{\mathbf{Y}}_{r,j}||_2^2 + \lambda_r ||\{\vartheta_{r,j}\}_j||_{l_0} \right). \tag{30}$$

Differentiation on $\mathbf{Y}$ gives after some manipulations the estimate of $\mathbf{Y}$ in the form:

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{\Phi}^{-1} \left(\mathbf{Z}/\sigma^2 + \mu \cdot \sum_r g_r \sum_j w_h(r,j) \mathbf{P}_j^T \hat{\mathbf{Y}}_{r,j} \right), \\ \mathbf{\Phi} &= \mathbf{I}/\sigma^2 + \mu \cdot \sum_r g_r \sum_j w_h(r,j) \mathbf{P}_j^T \mathbf{P}_j. \end{aligned} \tag{31}$$

In conclusion of this section we wish to discuss the meaning of the global penalty in the form (25):

1. The penalty (25) is a dictionary (transform, basis) dependent one defined first of all by the approximation accuracy of the true image spectrums Ω_r^Y by the basis spectrums Ω_r. This accuracy is complemented by the cost of this approximation calculated through the l_0-norms of the used bases. The penalty (25) is different from (3), (4), (7), which are dictionary independent, and similar to (5), (6) and (8), which are dictionary dependent;

2. The global penalty (25) is unusual in a number of aspects. One of the most important is that this penalty function is multilayer. The blocks in the reference rth group are selected as the ones close (similar) to the reference rth block and taken from different part of the image. These blocks form a 3D multilayer group used for group-wise processing. Each block can be selected for various reference blocks and in this way it can serve as layers in many different blocks. The sets of the multilayer constructions can be tracked in the formula (25) explicitly. The multilayer constructions can be quite complex and signal/observation adaptive.

3. Another type of multilayer constructions appear when we go from the spectrum to signal estimates. The windows are ovelapping and for each image pixel there are multiple window-wise estimates. These constructions cannot be tracked in (25) because they depend on image location of the windows collected in the groups. This grouping is explicitly revealed by the projection matrices $\mathbf{P}_j$ in (30). The weights g_r defined by the complexity of the group-wise models gives the aggregation weights fusing the multiple estimates in the final estimates for each pixel as it clear from (31).

4. The proposed penalties, group-wise and global, are inspired by the similar constructions developed in the BM3D algorithm for the group-wise multi-model collaborative filtering [2], [1]. Using the global penalty and variation formulation of estimation with the criterion in the form (30) we obtain a novel recursive algorithm sharing with BM3D the distinctive features of this algorithm: grouping of windows, 3D collaborative filtering and fusing of the group-wise estimates into the final one.

2.4 Implementation on the NEM algorithm

The NEM minimization is implemented as a recursive algorithm composed from the following steps:
1: Initialization: $\hat{Y}^{(1)} = Z$ and $g_r^{(1)} = 1$;
2: For every $t = 1, 2, ...$

- Calculate the windowed signals $\hat{Y}_r^{(t)}$, the weights

$$w_h^{(t)}(r,j) = 1(||\hat{Y}_r^{(t)} - \hat{Y}_j^{(t)}||_2^2 \leq h) \tag{32}$$

  and the windowed spectrums $\tilde{\theta}_{r,j}^{(t)} = \mathcal{T}_r^{\text{2D}}\left(\hat{Y}_j^{(t)}\right)$, $j \in K_r^{\Delta}$, for all groups r;

- Calculate the group-wise *"noisy"* spectrums $\tilde{\Omega}_r^{\hat{Y}^{(t)}}$, the updated windowed spectrum estimates $\hat{\theta}_{r,j}^{(t)}$ and the corresponding updated windowed signal estimates $\hat{Y}_{r,j}^{(t)}$ using thresholding (27) and the inverse transforms (28);

- Calculate the complexity $||\{\hat{\theta}_{r,j}^{(t)}\}_j||_{l_0}$ of the group models and the weights

$$g_r^{(t)} = \frac{1/||\{\hat{\theta}_{r,j}^{(t)}\}_j||_{l_0}}{\sum_r 1/||\{\hat{\theta}_{r,j}^{(t)}\}_j||_{l_0}};$$

- Update the signal estimate $\hat{\mathbf{Y}}^{(t+1)}$ using (31)

$$\hat{\mathbf{Y}}^{(t+1)} = \mathbf{\Phi}^{-1}\left(\mathbf{Z}/\sigma^2 + \mu \cdot \sum_r g_r^{(t)} \sum_j w_h^{(t)}(r,j)\mathbf{P}_j^T \hat{\mathbf{Y}}_{r,j}^{(t)}\right), \tag{33}$$
$$\mathbf{\Phi} = \mathbf{I}/\sigma^2 + \mu \cdot \sum_r g_r^{(t)} \sum_j w_h^{(t)}(r,j)\mathbf{P}_j^T \mathbf{P}_j;$$

- Continue until convergence.

The recursive procedure (33) of the NEM algorithm looks similar to the recursive algorithm (17) which is also derived by a global minimization of the energy function. We wish to note some distinctive features making a principal difference between these algorithms:

- The estimate $\hat{\mathbf{Y}}^{(t+1)}$ in (33) is a weighted mean of the partial block-wise estimates $\hat{\mathbf{Y}}_{r,j}$ with the weights defined by the complexity of the group-wise models while in (17) the sample mean of the block-wise estimates is used with the weights defined by the number of the windows overlapping for the particular pixel;

- In the algorithm (33) we have varying adaptive number of the windowed estimates for each pixel. This effect follows from the multiple-window modeling in the collaborative filtering. In (17) a fixed number of the window-estimate defined by the geometry of the image windowing is used;

- The algorithm (33) is nonlocal with the adaptive similar windows selection while the estimator (17) is local.

2.5 Links between NEM and BM3D algorithms

The BM3D algorithm [2] is composed from two successive stages: the hard-thresholding (*basic*) and the Wiener (*post-processing*) filtering. These two stages being similar in structure are different by the used filtering: the hard-thresholding for the basic stage and the Fourier domain Wiener filtering for the second stage. The both filtering are the collaborative ones produced for 3D spectral variables.

The BM3D algorithm as well as its further developments (e.g. [24], [25]) currently demonstrates the state-of-the-art visual and numerical performance [26].

A few works has been done to formalize the image modeling implemented in the BM3D algorithm, in particular, in order to connect this type of the algorithms with the frameworks of the more traditional image processing approaches and techniques.

The concept of the *group-wise multiple-models* and *group-wise* penalty are proposed in [1], where the 3D transform domain estimates are found by minimization of the criteria formulated as the group-wise penalties (19) and (24), where the true signal spectrums θ_j and Ω_r^Y are replaced by the noisy ones $\tilde{\theta}_j$ and Ω_r^Z:

$$pen_r(\{\vartheta_{r,j}\}_j) = \left(\sum_j w_h(r,j)||\tilde{\theta}_j - \vartheta_{r,j}||_2^2\right) + \lambda_r||\{\vartheta_{r,j}\}_j||_{l_0}, \tag{34}$$

$$pen_r(\Omega) = ||\Omega_r^Z - \Omega||_2^2 + \lambda_r||\Omega||_{l_0}. \tag{35}$$

Minimization of this $pen_r(\Omega)$ gives two first steps of the *basic thresholding* BM3D's: *grouping* and *collaborative* filtering. However, the third step *aggregation* is appeared in [1] as a separate complementary procedure of BM3D.

The mix-distribution modeling in [27] is proposed with intention to obtain the all three steps of the *basic thresholding* BM3D. The thresholded group-wise estimates are derived as the l_0-norm penalized conditional means (regressions) while for the aggregation an auxiliary optimization problem should be formulated giving the estimate in the desired form of the weighted mean with the weights inverse to the variances of the group-wise estimates.

The introduced *global penalty* (in the forms (20) and (25)) is different from the models considered in [1] and [27]. Being used in the variational formulation it gives the all three steps of the basic thresholding BM3D algorithm in all-in-one package.

To be precise with this statement we show that the basic thresholding BM3D algorithm can be derived as a minimizer of the *global penalty*:

- *Grouping.* The windowed data Z_j are collected in the groups K_r^{Δ} according to the rule (18) used in (20). With $h = \sigma^2\chi$ ($\chi > 0$ is a parameter) it gives the grouping rule implemented in the basic BM3D algorithm;

- *3-D collaborative filtering.* Let the true spectrums Ω_r^Y in the global penalty (25) be replaced by the noisy Ω_r^Z, and this global penalty be minimized on all Ω_r provided that g_r are fixed:

$$\left\{\hat{\Omega}_r\right\}_r = \min_{\{\Omega_r\}_r} PEN(\{\Omega_r\}_r), \tag{36}$$

$$PEN(\{\Omega_r\}_r) = \sum_r g_r \left(||\Omega_r^Z - \Omega_r||_2^2 + \lambda_r ||\Omega_r||_{l_0}\right). \tag{37}$$

  Then, the estimates $\hat{\Omega}_r$ are the hard-thresholded Ω_r^Z calculated according to the formula (27).

  If λ_r are selected as the corresponding thresholds in [24] then $\hat{\Omega}_r$ from (36) are the group-wise spectrum estimates of the BM3D algorithm.

  The corresponding collaborative windowed spectrum $\hat{\theta}_{r,j}$ and signal estimates $\hat{Y}_{r,j}$ are calculated according to (28).

- *Aggregation.* Let us use the global penalty in the form (20). Minimization of this penalty on $\mathbf{Y}$ using the vectorized representation of the signal in the penalty (see (30)) gives the estimate of the signal in the form

$$\hat{\mathbf{Y}} = \mathbf{\Phi}^{-1} \sum_r g_r \sum_j w_h(r,j) \mathbf{P}_j^T \hat{\mathbf{Y}}_{r,j}, \tag{38}$$

$$\mathbf{\Phi} = \sum_r g_r \sum_j w_h(r,j) \mathbf{P}_j^T \mathbf{P}_j,\ g_r = \frac{1}{\sigma^2 ||\{\hat{\theta}_{r,j}\}_j||_{l_0}},$$

  which is identical to used in BM3D for aggregation of the estimates obtained by the hard-thresholding.

Thus, it is shown that the basic thresholding BM3D algorithm can be derived by the alternative minimization of the global penalty using two different forms of this criterion given in the spectrum (25) and signal (29) domains.

Table 1: $PSNR$ values (in dB) obtained by NEM and basic thresholding BM3D.

σ \ Image	Cameraman	Lena	Barbara
	256×256	512×512	512×512
5	**38. 20** (38.12)	**38.70** (38.57)	38.29 (**38.40**)
10	**34.00** (33.80)	**35.69** (35.56)	**34.93** (34.88)
25	**29.30** (29.01)	**31.89** (31.37)	**30.31** (30.17)
35	**27.70** (27.32)	**30.34** (29.57)	**28.58** (28.25)

3 Simulation experiments

We have produced a number of experiments in order to test the performance of the recursive NEM algorithm.

Figure 1: Fragment of the cameraman test-image, $\sigma = 25$: true image (left); novel algorithm, $PSNR = 29.30$ dB (middle); basic thresholding BM3D algorithm $PSNR = 29.01$ dB (right).

It is shown that provided a proper selection of the main parameters of the NEM algorithm such as μ in (33) and λ_r in (30) this algorithm is able to give results which are better than obtained by BM3D. The comparison versus the BM3D is done for the basic thresholding part of this algorithm omitting the Wiener filtering used as the second stage in BM3D. In this case we enable a more or less fair comparison where both the BM3D and NEM algorithms use identical filtering instruments including the collaborative filtering and aggregation.

The numerical results are shown in Table 1. $PSNR$ values are given for 10 iterations of the NEM algorithm. The $PSNR$ values obtained by the basic BM3D algorithm are shown in brackets. The NEM algorithm mainly demonstrates the accuracy which is not worse and even better in particular for higher level of the noise. We found that these preliminary simulation results are promising and show that further improvement can be achieved by modifying the NEM algorithm and tuning its parameters.

A visual performance of the algorithms is demonstrated in Fig.1 and Fig.2. The fragments of the cameraman test-image in Fig.1 show the advantage of the NEM algorithm giving a better filtered background. Visually the estimates by BM3D and NEM for the lena test-image are more less equivalent with a better value of $PSNR$ for the later algorithm. This visual similarity is also important as it shows that the NEW recursive procedure does not produce any artifacts sometimes typical for recursive algorithms.

Figure 2: The lena test-image, $\sigma = 25$: true image (left); novel algorithm, $PSNR = 31.89$ dB (middle); basic thresholding BM3D algorithm $PSNR = 31.37$ dB (right).

4 Conclusion and further work

- The main results of this paper are the proposed global nonlocal penalty in the energy criterion (25) and the recursive algorithm (33). The great performance of BM3D is a strong argument in favor of the proposed global penalty. While the basic thresholding part of BM3D can be derived as a minimizer of the global penalty, the performance of the recursive algorithm (33) and the improvement which could be achieved by this algorithm are the problems which require a further study.

  The presented simulation results are initial steps in this direction. A serious work should be done, in particular concerning the parameter's tuning in order to obtain the algorithm competitive with the complete BM3D including both the thresholding and Wiener stages. It deserves to be mentioned that the BM3D algorithm and all its development are equipped with the very well optimized parameters.

- The proposed global penalty is quite universal and can be incorporated in various data processing problems. One of the interesting applications concerns the inverse problems.

  Let us make clear this proposal. Assume that the observation model has a form

$$\mathbf{Z} = \mathbf{A}\mathbf{Y} + \boldsymbol{\varepsilon},$$

  where $\mathbf{Z}$ and $\mathbf{Y}$ are the vectorized observed noisy and true images as it is in (30) and $\mathbf{A}$ is a matrix blur operator.

  Then, from minimization of

$$J = ||\mathbf{Z} - \mathbf{A}\mathbf{Y}||_2^2/\sigma^2 + \mu \cdot \sum_r g_r \left(\sum_j w_h(r,j) ||\mathbf{P}_j\mathbf{Y} - \hat{\mathbf{Y}}_{r,j}||_2^2 + \lambda_r ||\{\vartheta_{r,j}\}_j||_{l_0} \right) \tag{39}$$

  we arrive instead of (33) to the recursive algorithm

$$\hat{\mathbf{Y}}^{(t+1)} = \mathbf{\Phi}^{-1} \left(\mathbf{A}^T\mathbf{Z}/\sigma^2 + \mu \cdot \sum_r g_r^{(t)} \sum_j w_h^{(t)}(r,j) \mathbf{P}_j^T \hat{\mathbf{Y}}_{r,j}^{(t)} \right), \tag{40}$$

$$\mathbf{\Phi} = \mathbf{A}^T\mathbf{A}/\sigma^2 + \mu \cdot \sum_r g_r^{(t)} \sum_j w_h^{(t)}(r,j) \mathbf{P}_j^T\mathbf{P}_j.$$

A remarkable and promising feature of this algorithm is that the regularization of the matrix $\mathbf{A}^T\mathbf{A}$ is produced by the data-depending matrix $\sum_r g_r^{(t)} \sum_j w_h^{(t)}(r,j)\mathbf{P}_j^T\mathbf{P}_j$.

5 Acknowledgement

This work was supported by the Academy of Finland, project No. 213462 (Finnish Centre of Excellence program 2006 – 2011). We highly appreciate critical comments by A. Foi used in order to improve the paper.

References

[1] V. Katkovnik, A. Foi, K. Egiazarian, and J. Astola, "Nonparametric regression in imaging: from local kernel to multiple-model nonlocal collaborative filtering," in *Proc. 2008 Int. Workshop on Local and Non-Local Approximation in Image Processing, LNLA 2008*, Lausanne, Switzerland, August 2008.

[2] Dabov, K., A. Foi, V. Katkovnik, and Egiazarian, K., "Image denoising by sparse 3D transform-domain collaborative filtering," *IEEE Transactions on Image Processing*, vol. 16, no. 8, pp. 2080 - 2095, 2007.

[3] K. Egiazarian, V. Katkovnik, and J. Astola, "Local transform-based image de-noising with adaptive window size selection," *Proc. SPIE Image and Signal Processing for Remote Sensing VI*, vol. 4170, 4170-4, Jan. 2001.

[4] Katkovnik, V., K. Egiazarian, and J. Astola, "Adaptive window size image de-noising based on intersection of confidence intervals (ICI) rule", *J. of Math. Imaging and Vision*, vol. 16, no. 3, pp. 223-235, 2002.

[5] Katkovnik, V., A. Foi, K. Egiazarian, and J. Astola, "Directional varying scale approximations for anisotropic signal processing", *Proc. XII European Signal Proc. Conf., EUSIPCO 2004,* Vienna, pp. 101-104, September 2004.

[6] Katkovnik, V., A. Foi, K. Egiazarian, and J. Astola, "Anisotropic local likelihood approximations", *Proc. of Electronic Imaging 2005*, 5672-19, January 2005.

[7] Katkovnik, V., K. Egiazarian, J. Astola, *Local approximation techniques in signal and image processing*. SPIE PRESS, Bellingham, Washington, 2006.

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

[9] Astola, J. and L. Yaroslavsky (eds.) *Advances in Signal Transforms: Theory and Applications*. EURASIP Book Series on Signal Processing and Communications, vol. 7, Hindawi Publishing Corporation, 2007.

[10] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE PAMI*, vol. 6, no. 6,pp. 721–741, 1984.

[11] M. Bertero and P. Boccacci. *Introduction to inverse problems in imaging*. IOP Publishing Ltd, 1998.

[12] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Phys. D,* 60 2, pp. 259-268, 1993.

[13] L. Rudin, S. Osher, and E. Fatemi, "Nonlinear algorithms," *Phys. D*, vol. 60, pp. 259-268, 1992.

[14] S. Roth and M. J. Black, "Fields of experts: a framework for learning image priors, " *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, San Diego, CA, 2005, pp. 860–867.

[15] S. Roth and M. J. Black, "Fields of experts," *International Journal of Computer Vision,* vol. 82, no. 2, pp. 205-229, 2009.

[16] S. Kindermann, S. Osher, and P.W. Jones, "Deblurring and denoising of images by nonlocal functionals," *SIAM Multiscale Modeling & Simulation*, vol. 4, no. 4, pp. 1091-1115, 2005.

[17] G. Gilboa and S. Osher, "Nonlocal linear image regularization and supervised segmentation," *SIAM Multiscale Modeling and Simulation*, Vol. 6, No. 2, pp. 595-630, 2007.

[18] Elad, M., and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries", *IEEE Trans. Image Processing*, vol. 15, no. 12, pp. 3736-3745, 2006.

[19] Guleryuz, O., "Weighted averaging for denoising with overcomplete dictionaries", *IEEE Trans. Image Processing*, vol. 16, no. 12, pp. 3020-3034, 2007.

[20] M. Aharon, M. Elad, and A.M. Bruckstein, "The K-SVD: An algorithm for designing of overcomplete dictionaries for sparse representation", *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311-4322, 2006.

[21] Buades, A., B. Coll, and J.M. Morel, "A review of image denoising algorithms, with a new one," *SIAM Multiscale Modeling and Simulation*, vol. 4, pp. 490–530, 2005.

[22] Buades, A., B. Coll, and J.M. Morel, "Nonlocal image and movie denoising", *Int. J. Computer Vision*, 2007.

[23] Kervrann, C., and J. Boulanger, "Local adaptivity to variable smoothness for exemplar-based image regularization and representation", *Int. J. Computer Vision*, vol. 79, no. 1, pp. 45-69, 2008.

[24] Dabov, K., A. Foi, V. Katkovnik, and Egiazarian, K., "A nonlocal and shape-adaptive transform-domain collaborative filtering," in *Proceedings 2008 Int. Workshop on Local and Non-Local Approximation in Image Processing, LNLA 2008*, Lausanne, Switzerland, August 2008.

[25] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "BM3D Image Denoising with Shape-Adaptive Principal Component Analysis," in *Proceedings Signal Processing with Adaptive Sparse Structured Representations* (SPARS'09), Saint-Malo (France), April, 2009.

[26] Lansel, S., D. Donoho, and T. Weissman, "DenoiseLab: a standard test set and evaluation method to compare denoising algorithms", *http://www.stanford.edu/~slansel/DenoiseLab/.*

[27] V. Katkovnik, A. Foi, K. Egiazarian, "Mix-distribution modeling for overcomplete denoising," *Proc. 9th workshop on Adaptation and Learning in Control and Signal Processing (ALCOSP'07)*, St. Petersburg, Russia, August, 29-31, 2007.