Image Denoising via Learned Dictionaries and Sparse Representations*

Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel

* Joint work with Michal Aharon

IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)
New-York, June 18-22, 2006
Noise Removal?

Our story focuses on image denoising …

- **Important**: (i) Practical application; (ii) A convenient platform (being the simplest inverse problem) for testing basic ideas in image processing.

- **Many Considered Directions**: Partial differential equations, Statistical estimators, Adaptive filters, Inverse problems & regularization, Example-based techniques, Sparse representations, …
Part I:
Sparse and Redundant Representations?
Denoising By Energy Minimization

Many of the proposed denoising algorithms are related to the minimization of an energy function of the form

$$f(x) = \frac{1}{2} \left\| x - y \right\|^2_2 + \Pr(x)$$

- $x$: Unknown to be recovered
- $y$: Given measurements

- This is in-fact a Bayesian point of view, adopting the Maximum-Aposteriori Probability (MAP) estimation.
- Clearly, the wisdom in such an approach is within the choice of the prior – **modeling the images** of interest.
The Evolution Of $\Pr(x)$

During the past several decades we have made all sort of guesses about the prior $\Pr(x)$ for images:

- Mumford & Shah formulation,
- Compression algorithms as priors,
- ... 

\[ \Pr(x) = \lambda \norm{x}_2^2 \]  
\[ \Pr(x) = \lambda \norm{Lx}_2^2 \]  
\[ \Pr(x) = \lambda \norm{L}{x}_W^2 \]  
\[ \Pr(x) = \lambda \rho \{Lx\} \]  

- Energy
- Smoothness
- Adapt+ Smooth
- Robust Statistics

\[ \Pr(x) = \lambda \norm{\nabla x}_1 \]  
\[ \Pr(x) = \lambda \norm{Wx}_1 \]  
\[ \Pr(x) = \lambda \alpha_0^0 \]  
\[ \text{for } x = D\alpha \]  
- Sparse & Redundant

- Total-Variation
- Wavelet Sparsity

- Mumford & Shah formulation,
- Compression algorithms as priors,
- ...

Image Denoising Via Learned Dictionaries and Sparse representations
By: Michael Elad
The *Sparseland* Model for Images

- Every column in the dictionary $D$ is a prototype signal (Atom).

- The vector $\alpha$ is generated randomly with few (say $L$) non-zeros at random locations and with random values.
Our MAP Energy Function

- We $L_0$ norm is effectively counting the number of non-zeros in $\alpha$.

- The vector $\alpha$ is the representation (sparse/redundant).

- The above is solved (approximated!) using a greedy algorithm – the Matching Pursuit [Mallat & Zhang (`93)].

- In the past 5-10 years there has been a major progress in the field of sparse & redundant representations, and its uses.
What Should D Be?

\[ \hat{\alpha} = \arg\min_\alpha \frac{1}{2} \| D\alpha - y \|_2^2 \text{ s.t. } \|\alpha\|_0 \leq L \Rightarrow \hat{x} = D\hat{\alpha} \]

Our Assumption: Good-behaved Images have a sparse representation

\[ \alpha = \hat{\alpha} = \alpha \]

D should be chosen such that it sparsifies the representations

One approach to choose D is from a known set of transforms (Steerable wavelet, Curvelet, Contourlets, Bandlets, …)

The approach we will take for building D is training it, based on **Learning from Image Examples**
Part II: Dictionary Learning: The K-SVD Algorithm
Measure of Quality for $\mathbf{D}$

Each example is a linear combination of atoms from $\mathbf{D}$

$$\min_{\mathbf{D}, \mathbf{A}} \sum_{j=1}^{P} \left\| \mathbf{D} \alpha_j - \mathbf{x}_j \right\|_2^2$$

subject to

Each example has a sparse representation with no more than $L$ atoms

$$\forall j, \left\| \alpha_j \right\|_0 \leq L$$

Field & Olshausen ('96)
Engan et. al. ('99)
Lewicki & Sejnowski ('00)
Cotter et. al. ('03)
Gribonval et. al. ('04)
Aharon, Elad, & Bruckstein ('04)
Aharon, Elad, & Bruckstein ('05)
K- Means For Clustering

Clustering: An extreme sparse representation

- Initialize $D$
- Sparse Coding
  - Nearest Neighbor
- Dictionary Update
  - Column-by-Column by Mean computation over the relevant examples
The K-SVD Algorithm - General

Aharon, Elad, & Bruckstein (’04)

Initialize $D$

Sparse Coding
Use Matching Pursuit

Dictionary Update
Column-by-Column by SVD computation over the relevant examples

$X^T$

$D$
Part III: Combining It All
The K-SVD algorithm is reasonable for low-dimension signals (N in the range 10-400). As N grows, the complexity and the memory requirements of the K-SVD become prohibitive.

So, how should large images be handled?

The solution: Force shift-invariant sparsity - on each patch of size N-by-N (N=8) in the image, including overlaps [Roth & Black (`05)].
What Data to Train On?

Option 1:
- Use a database of images,
- We tried that, and it works fine (~0.5-1dB below the state-of-the-art).

Option 2:
- Use the corrupted image itself !!
- Simply sweep through all patches of size N-by-N (overlapping blocks),
- Image of size $1000^2$ pixels $\Rightarrow \sim 10^6$ examples to use – more than enough.
- This works much better!
\[ \hat{x} = \text{ArgMin}_{x, \{ \alpha_{ij} \}_j, D} \frac{1}{2} \| x - y \|^2_2 + \mu \sum_{ij} \| R_{ij} x - D \alpha_{ij} \|^2_2 \quad \text{s.t.} \quad \| \alpha_{ij} \|^0_0 \leq L \]

- **x = y and D known**
  - Compute \( \alpha_{ij} \) per patch
  \[ \alpha_{ij} = \text{Min}_\alpha \| R_{ij} x - D \alpha \|^2_2 \quad \text{s.t.} \quad \| \alpha \|^0_0 \leq L \]
  using the matching pursuit

- **x and \( \alpha_{ij} \) known**
  - Compute \( D \) to minimize
  \[ \text{Min}_\alpha \sum_{ij} \| R_{ij} x - D \alpha \|^2_2 \]
  using SVD, updating one column at a time

- **D and \( \alpha_{ij} \) known**
  - Compute \( x \) by
  \[ x = \left[ I + \mu \sum_{ij} R_{ij}^T R_{ij} \right]^{-1} \left[ y + \mu \sum_{ij} R_{ij}^T D \alpha_{ij} \right] \]
  which is a simple averaging of shifted patches

Complexity of this algorithm: \( O(N^2 \times L \times \text{Iterations}) \) per pixel. For \( N=8, L=1, \) and 10 iterations, we need 640 operations per pixel.
Denoising Results

Source

Noisy image

\[ \sigma = 20 \]

Result 30.829dB

The obtained dictionary after 10 iterations

The results of this algorithm compete favorably with the state-of-the-art (e.g., GSM+steerable wavelets [Portilla, Strela, Wainwright, & Simoncelli ('03)] - giving \(~0.5-1\text{dB better results}\)
Today We Have Seen …

An energy minimization method

A Bayesian (MAP) point of view

Getting a relatively simple and highly effective denoising algorithm

Using examples from the noisy image itself to learn the image prior

Using an image prior based on sparsity and redundancy

\[ \hat{x} = \text{ArgMin}_{x, \{\alpha_{ij}\}, D} \left( \|x - y\|_2^2 + \mu \sum_{ij} \|R_{ij}x - D\alpha_{ij}\|_2^2 \right) \]

Subject to \( \|\alpha_{ij}\|_0 \leq L \)

More on this in http://www.cs.technion.ac.il/~elad
We refer only to the examples that use the column $d_k$.

Fixing all $A$ and $D$ apart from the $k^{th}$ column, and seek both $d_k$ and the $k^{th}$ column in $A$ to better fit the residual! 

We should solve:

$$\min_{d_k, \alpha} \left\| A_k d_k + E \right\|_F^2$$
K- SVD: Sparse Coding Stage

\[
\min \sum_{j=1}^{P} \left\| D \alpha_j - x_j \right\|_2^2 \quad \text{s.t.} \quad \forall j, \left\| \alpha_j \right\|_p \leq L
\]

D is known! For the j\(^{th}\) item we solve

\[
\min_{\alpha} \left\| D \alpha - x_j \right\|_2^2 \quad \text{s.t.} \quad \left\| \alpha \right\|_p \leq L
\]

Solved by Matching Pursuit