Denoising in digital radiography:
A total variation approach

I. Frosio
M. Lucchese
N. A. Borghese

Images are corrupted by noise…

i) When measurement of some physical parameter is performed, noise corruption cannot be avoided.

ii) Each pixel of a digital image measures a number of photons.

Therefore, from i) and ii)…

…Images are corrupted by noise!

Gaussian noise

(not so useful for digital radiographs, but a good model for learning...)

- Measurement noise is often modeled as Gaussian noise...
- Let x be the measured physical parameter, let μ be the noise free parameter and let σ^2 be the variance of the measured parameter (noise power); the probability density function for x is given by:

$$p(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right]$$

Gaussian noise and likelihood

- Images are composed by a set of pixels, **x** (**x** is a vector!)
- How can we quantify the probability to measure the image **x**, given the probability density function for each pixel?
- Let us assume that the variance is equal for each pixel;
- Let x_i and μ_i be the measured and noiseless values for the i-th pixel;
- Likelihood function, L(**x**|μ):

$$L(\mathbf{x} \mid \boldsymbol{\mu}) = \prod_{i=1}^{N} p(x_i \mid \mu_i) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma} \right)^2 \right]$$

- L(**x**|μ) describes the probability to measure the image **x**, given the noise free value for each pixel, μ.

What about denoising???

- What is denoising then?

<u>Denoising = estimate μ from x.</u>

- How can we estimate μ?
- Maximize $p(\mu|x)$ => this usually leads to an hard, inverse problem.
- It is easier to maximize $p(x|\mu)$, that is => maximize the likelihood function (a "simple", direct problem).
- But… Is maximization of $p(\mu|x)$ different from that of $p(x|\mu)$?

Bayes and likelihood

- Bayes theorem:

$$p(\mathbf{\mu} \mid \mathbf{x})p(\mathbf{x}) = p(\mathbf{x} \mid \mathbf{\mu})p(\mathbf{\mu}) \Rightarrow$$

$$\Rightarrow p(\mathbf{\mu} \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \mathbf{\mu})p(\mathbf{\mu})}{p(\mathbf{x})}$$

Likelihood [$p(\mathbf{x} \mid \mathbf{\mu})$]

A priori hypothesis on the estimated parameters μ. For the moment, let us suppose $p(\mu)$ = cost. [$p(\mathbf{\mu})$]

Probability density function for the data x… Just a normalization factor!!! [$p(\mathbf{x})$]

- In this case, maximizing $p(\mu|x)$ or $p(x|\mu)$ is the same!

So, let us maximize the likelihood...

- Instead of maximizing L(x|μ), it is easier to minize –log[L(x|μ)].
- When the noise is Gaussian, we get:

$$L(\mathbf{x}\,|\,\boldsymbol{\mu}) = \prod_{i=1}^{N} p(x_i \,|\, \mu_i) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i-\mu_i}{\sigma}\right)^2\right]$$

$$f(\mathbf{x}\,|\,\boldsymbol{\mu}) = -\ln[L(\mathbf{x}\,|\,\boldsymbol{\mu})] = -\sum_{i=1}^{N} \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) + \frac{1}{2\sigma}\sum_{i=1}^{N}(x_i-\mu_i)^2$$

Least squares!

Constant!

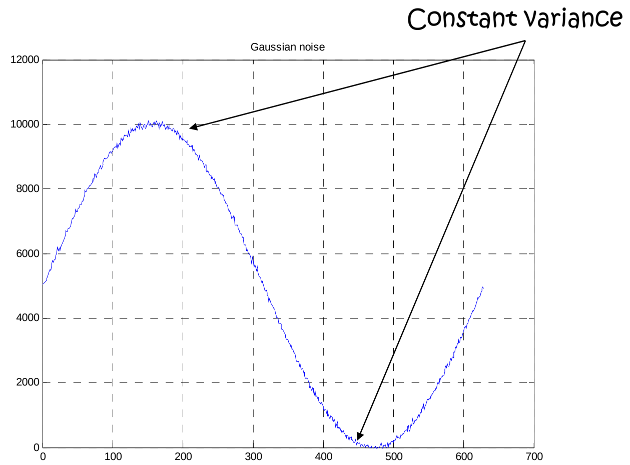
- Maximize L => Least squares problem!

However, what about noise in digital radiography?

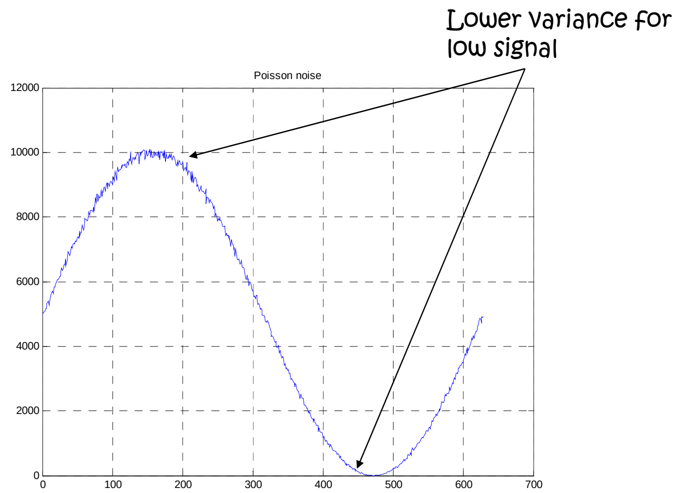
- Noise in digital radiography is Poisson (photon counting noise)!
- Let $p_{n,i}$ be the noisy (measured) number of photons associated to pixel i, and p_i the unnoisy number of photons. Then:

$$p(p_{n,i} \,|\, p_i) = \frac{p_i^{\,p_{n,i}} e^{-p_i}}{p_{n,i}!}$$

Gaussian noise: example



Poisson noise: example



Likelihood for Poisson noise

- Let us write the negative log likelihood for the Poisson case:

$$L(\mathbf{p_n} \mid \mathbf{p}) = \prod_{i=1}^{N} p(p_{n,i} \mid p_i) = \prod_{i=1}^{N} \frac{p_i^{\,p_{n,i}} e^{-p_i}}{p_{n,i}!}$$

$$f(\mathbf{p_n} \mid \mathbf{p}) = -\ln[L(\mathbf{x} \mid \boldsymbol{\mu})] = -\sum_{i=1}^{N}[p_{n,i} \cdot \ln(p_i)] + \sum_{i=1}^{N} p_i + \sum_{i=1}^{N} \ln(p_{n,i}!) =$$

$$= \sum_{i=1}^{N}[p_i - p_{n,i} \cdot \ln(p_i)]$$

- $L(p_n|p)$ is also known as Kullback-Leibler divergence (apart from a constant term, which does not affect the minimization process), $KL(p_n|p)$.

Maximize L!

L is maximized <=> f is minimized;

- Optimization (Gaussian noise) can be performed posing:

$$\frac{\partial f(\mathbf{x} \mid \boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = \mathbf{0} \Leftrightarrow \frac{\partial f(\mathbf{x} \mid \boldsymbol{\mu})}{\partial \mu_i} = 0, \quad \forall i \Rightarrow \frac{\partial \sum_{j=1}^{N} (x_j - \mu_j)^2}{\partial \mu_i} = 0, \quad \forall i \Rightarrow$$

$$\Rightarrow 2(x_i - \mu_i) = 0, \quad \forall i \Rightarrow \boxed{x_i = \mu_i, \quad \forall i}$$

- The noisy image gives the highest likelihood!!!
- This solution is not so interesting… The likelihood approach suffers from a severe overfitting problem.

Maximize L!

L is maximized <=> f is minimized;

- Optimization (Poisson noise) can be performed posing:

$$\frac{\partial f(\mathbf{p_n}|\mathbf{p})}{\partial \mathbf{p}} = \mathbf{0} \Leftrightarrow \frac{\partial f(\mathbf{p_n}|\mathbf{p})}{\partial p_i} = 0, \quad \forall i \Rightarrow \frac{\partial \sum_{i=1}^{N} \left[p_i - p_{n,i} \cdot \ln(p_i)\right]}{\partial p_i} = 0, \quad \forall i \Rightarrow$$

$$\Rightarrow 1 - \frac{p_{n,i}}{p_i} = 0, \quad \forall i \Rightarrow \boxed{p_i = p_{n,i}, \quad \forall i}$$

- The noisy image gives the highest likelihood!!!
- This solution is not so interesting… The likelihood approach suffers from a severe overfitting problem.

Back to Bayes

- Bayes theorem:

Likelihood

$$\Rightarrow p(\mathbf{p}|\mathbf{p_n}) = \frac{p(\mathbf{p_n}|\mathbf{p})\,p(\mathbf{p})}{p(\mathbf{p_n})}$$

A priori hypothesis on the estimated parameters μ.

Probability density function for the data x… Just a normalization factor!!!

- If we introduce a-priori knowledge about the solution μ, we get a <u>Maximum A Posteriori (MAP)</u> solution – $p(p|p_n)$ is maximized!

What do we have to minimize now?

- We want to maximize $p(p|p_n) \sim p(p_n|p)\ p(p)$, that is:

$$-\ln[p(\mathbf{p}\,|\,\mathbf{p_n})] = -\ln[p(\mathbf{p_n}\,|\,\mathbf{p})p(\mathbf{p})] = -\ln\prod_{i=1}^{N}[p(p_{n,i}|p_i)\cdot p(p_i)] =$$

$$-\sum_{i=1}^{N}\ln[p(p_{n,i}|p_i)\cdot p(p_i)] = -\sum_{i=1}^{N}\ln p(p_{n,i}|p_i) - \sum_{i=1}^{N}\ln p(p_i) =$$

$$= -\ln[L(\mathbf{p_n}\,|\,\mathbf{p})] - \sum_{i=1}^{N}\ln p(p_i)$$

Negative log likelihood

Regularization term (a priori information)

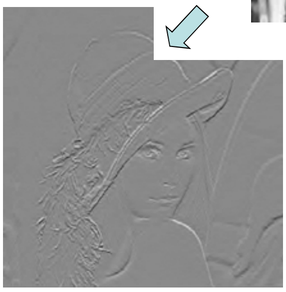
A priori term

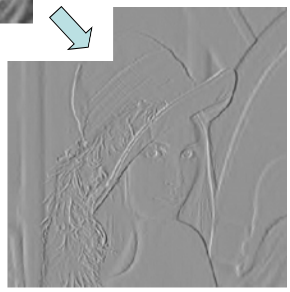
- Let us call p_x and p_y the two components of the gradient of the image.
- These are easily computed, for instance as:
  - $p_x = p(i,j) - p(i-1,j)$;
  - $p_y = p(i,j) - p(i,j-1)$;
- The gradient (a vector!) will be indicated as ∇p;
- $||\nabla p||$ indicates the norm of the gradient.

A priori term – image gradients (no noise)

$p_x = p(i,j) - p(i-1,j)$

$p_y = p(i,j) - p(i,j-1)$

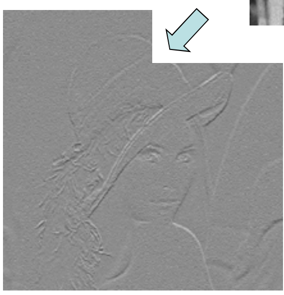


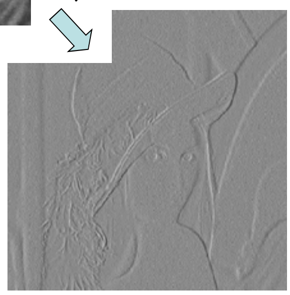


A priori term – image gradients (noise)

$p_x = p(i,j) - p(i-1,j)$

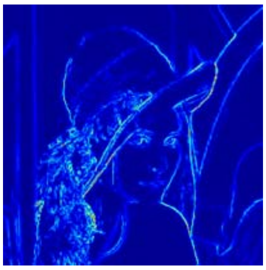
$p_y = p(i,j) - p(i,j-1)$



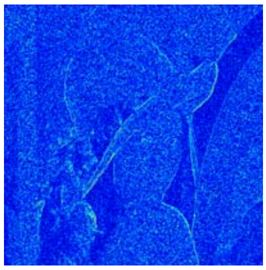


A priori term – norm of image gradient

No noise



Noise



In the real image, most of the areas are characterized by an (almost) null gradient norm;

We can for instance suppose that $||\nabla p||$ is a random variable with Gaussian distribution, zero mean and variance equal to β^2.

[Note that, in the noisy image, the norm of the gradient assume higher values → low $||\nabla p||$ means low noise!]

MAP and regularization theory

- Poisson noise, normal distribution for the norm of the gradient:

$$f(\mathbf{p_n}\,|\,\mathbf{p}) = -\ln[L(\mathbf{p_n}\,|\,\mathbf{p})] - \sum_{i=1}^{N} \ln p(\nabla \mathbf{p_i}) =$$

$$= \sum_{i=1}^{N} [p_i - p_{n,i} \cdot \ln(p_i)] - \sum_{i=1}^{N} \ln\left[\frac{1}{\beta\sqrt{2\pi}} \exp\left(-\frac{1}{2}\frac{\|\nabla \mathbf{p_i}\|^2}{\beta^2}\right)\right] =$$

$$= \sum_{i=1}^{N} [p_i - p_{n,i} \cdot \ln(p_i)] + N \ln\left(\beta\sqrt{2\pi}\right) + \frac{1}{2\beta^2} \sum_{i=1}^{N} \|\nabla \mathbf{p_i}\|^2$$

Const!!!

Negative log likelihood

Regularization term (a priori information)

MAP and regularization theory

- We look for the minimum of f…
- … The likelihood is maximized (data fitting term)…
- … At the same time, the squared norm of the gradient is minimized (regularization term)…
- … The regularization parameter ($1/2\beta^2$) balances between a perfect data fitting and very regular image…

$$f(\mathbf{p_n} \mid \mathbf{p}) = \sum_{i=1}^{N} \left[p_i - p_{n,i} \cdot \ln(p_i) \right] + \frac{1}{2\beta^2} \sum_{i=1}^{N} \left\| \nabla \mathbf{p_i} \right\|^2$$

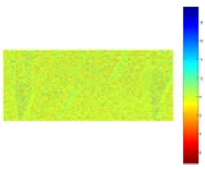
MAP and regularization theory

For ($1/2\beta^2$) = 0 we get the maximum likelihood solution;

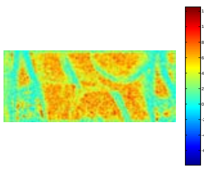
Increasing ($1/2\beta^2$) we get a more regular (less noisy) solution;

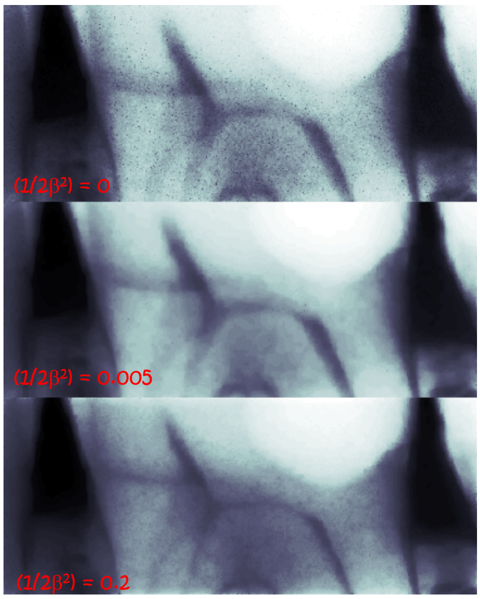
For ($1/2\beta^2$) -> ∞, a completely smooth image is achieved.

Noise reduction.



Noise and edge reduction.





Fix the ideas

- A statistical based denoising filter is achieved minimizing:

  $f = -\ln[L(p_n|p)] - \lambda \cdot \ln[p(p)]$

- The **data fitting** term is derived **from the noise statistical distribution** (likelihood of the data); generally, the choice for this term is unquestionable.

- The **regularization term** is derived from **a-priori knowledge** regarding some properties of the solution; this term is generally user defined.

- Depending on the regularization parameter λ, the first or the second term assume more or less importance. For λ->0, the maximum likelihood solution is obtained.

Gibbs prior

- Up to now, we assumed a normal distribution for the norm of the gradient, → Tikhonov regularization (quadratic penalization).

- A more general framework is obtained considering:

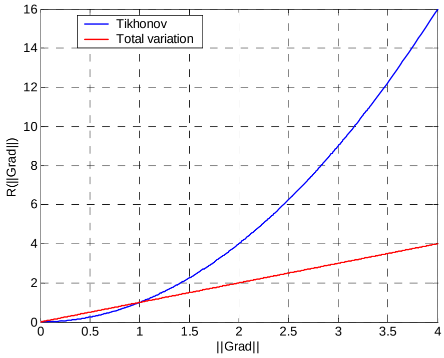
  $p(p) = \exp[-R(p)]$ (Gibb's prior)

- $R(p)$ -> Energy function ~ regularization term (note that $-\ln \exp[-R(p)] = R(p)$!)

- Tikhonov assumes $R(p) = -½ (||\nabla p||/\beta)^2$

Edge preserving denoising?

- Tikhonov term penalizes the image edges (high gradient) more than the noise gradients.
- It is well known that Tikhonov regularization does not preserve edges.
- An edge preserving algorithm is obtained considering $R(p)=-\|\nabla p\|$ [Total variation, TV].

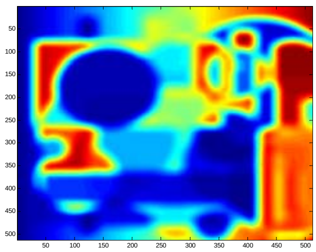


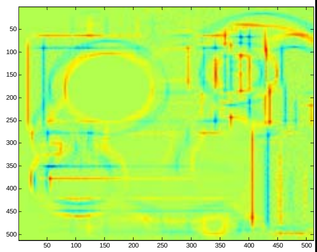
Tikhonov vs. TV (preview)

Filtered image

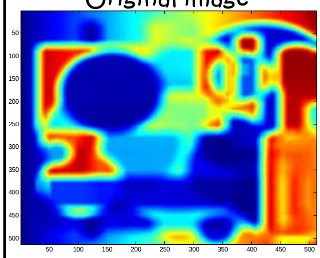
Difference

Tikhonov =>

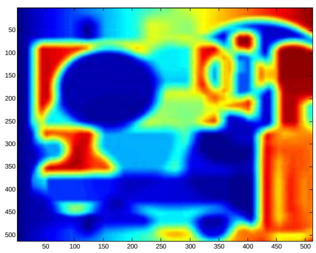


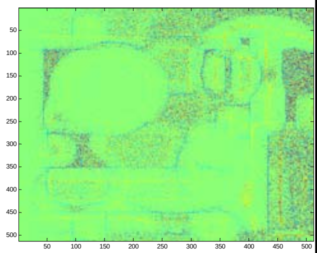


Original image



TV =>



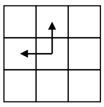


TV in digital radiography: starting point and problems

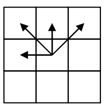
- p_n, noisy image affected by Poisson noise (likelihood => KL);
- p, noise free image (unknown);
- $R(p) = ||\nabla p||$ (Total variation);
- Minimize $f(p|p_n) = KL(p_n,p) + \lambda \cdot \Sigma_{i=1..N}||\nabla p_i||$.

- How to compute $||\nabla p_i||$? => A compromise between computational efficiency and accuracy has to be achieved.
- How to minimize $f(p|p_n)$? => An iterative optimization technique is required.

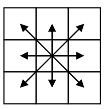
How to compute $||\nabla p_i||$?



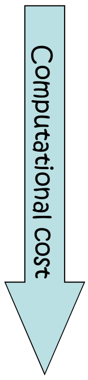
$p_x = p(u,v) – p(u-1,v)$
$p_y = p(u,v) – p(u,v-1)$
$||p_i||_1= |p_x|+|p_y|$ L1 norm
$||p_i||_2= [p_x^2+p_y^2]^{½}$ L2 norm



$||p_i||_1= |p_x|+|p_y|+ |p_{xy}|+|p_{yx}|$
$||p_i||_2= [p_x^2+p_y^2+ p_{xy}^2+p_{yx}^2]^{½}$



$||p_i||_1= |p_x|+|p_y|+ |p_{xy}|+|p_{yx}|+…$
$||p_i||_2= [p_x^2+p_y^2+ p_{xy}^2+p_{yx}^2+…]^{½}$



The computational cost increases with the number of neighbours considered for computing the gradient;

The computational cost is higher for L2 norm with respect to L1 norm;

What about accuracy? => See experimental results!

How to minimize $f(p|p_n)$?

- $f(p|p_n)$ is strongly non linear; solving $df(p|p_n)/dp=0$ directly is not possible => iterative optimization methods.

1) Steepest descent + line search (SD+LS)
2) Expectation – Maximization (damped with line search - EM)
3) Scaled gradient (SG)

Steepest descent + line search (SD+LS)

- $p^{k+1}=p^k-\alpha \cdot df(p|p_n)/dp$ =>
  => $p^{k+1}=p^k-\alpha \cdot df(p|p_n)/dp$
- The damping parameter α is estimated at each iteration to assure convergence ($f^{k+1}<f^k$);

+: easy implementation;

-: slow convergence, the method has been damped (line search) to improve convergence ($\alpha>1$).

EM + line search (EM)

- Consider the pixel i, then:

  $df(\mathbf{p}|\mathbf{p}_n)/dp_i=0$ =>
  => $dKL(\mathbf{p}|\mathbf{p}_n)/dp_i + dR/dp_i = 0$
  => $p_i \cdot \beta \cdot dR/dp_i + p_i - p_{n,i} = 0$ =>
  => $p_i = p_{n,i} / (\beta \cdot dR/dp_i + 1)$ [Fixed poin iteration]

- Damped formula: $p_i = p_i \cdot (1-\alpha) + \alpha \cdot p_{n,I} / (\beta \cdot dR/dp_i + 1)$

- The damping parameter α is estimated at each iteration to assure convergence ($f^{k+1}<f^k$);

+: easy implementation, fast convergence;
-: the method has been damped to assure convergence ($\alpha<1$, what happens when $\beta \cdot dR/dp_i + 1 \to 0$???).

Scaled gradient (SG)

- Consider the gradient method formula;
- Each component of the gradient is scaled to improve convergence (S is a diagonal matrix containing the scaling parameters):

  $\mathbf{p}^{k+1}=\mathbf{p}^k-\alpha \cdot S \cdot df(\mathbf{p}|\mathbf{p}_n)/d\mathbf{p}$

- The matrix S is computed from an opportune gradient decomposition and KKT conditions;

+: easy implementation, fastest convergence; it can also be demonstrated that, for positive initial values, the estimated solution remains positive at each iteration!
-: ???.

Problems with dR/dp_i

- Independently from the optimization method, the term dR/dp_i has to be computed at each iteration far any i;
- We have:

$$dR/dp_i = d[\Sigma_{i=1..N}(||\nabla p_i||_1)]/dp_i$$

XOR

$$dR/dp_i = d[\Sigma_{i=1..N}(||\nabla p_i||_2)]/dp_i$$

Problems with dR/dp_i

- Let us compute it for $||.||_2$ ($R/dp_i = d[\Sigma_{i=1..N}(||\nabla p_i||_2)]/dp_i$)

$$\frac{dR}{dp_i} = \frac{d\sum_{i=1}^{N}\sqrt{p_{x,i}^{\;2} + p_{y,i}^{\;2}}}{dp_i} = \frac{d\left(\sqrt{[p(u,v)-p(u-1,v)]^2 + [p(u,v)-p(u,v-1)]^2}\right)}{dp_i} + \ldots =$$

$$= \frac{2[p(u,v)-p(u-1,v)] + 2[p(u,v)-p(u,v-1)]}{\sqrt{[p(u,v)-p(u-1,v)]^2 + [p(u,v)-p(u,v-1)]^2}} + \ldots = 2\frac{p_{x,i} + p_{y,i}}{\|\nabla p(u,v)\|} + \ldots$$

- To avoid division by zero:

$$\frac{dR}{dp_i} = 2\frac{p_{x,i} + p_{y,i}}{\|\nabla p(u,v)\|} + \ldots \rightarrow 2\frac{p_{x,i} + p_{y,i}}{\sqrt{[p(u,v)-p(u-1,v)]^2 + [p(u,v)-p(u,v-1)]^2 + \delta}} + \ldots$$

Problems with dR/dp_i

- Let us compute it for $||.||_1$ $(R/dp_i = d[\Sigma_{i=1..N}(||\nabla p_i||_1)]/dp_i)$

$$\frac{dR}{dp_i} = \frac{d\sum_{i=1}^{N}\left(|p_{x,i}| + |p_{y,i}|\right)}{dp_i} = \left[\sum_{i=1}^{N} \frac{d}{dp_i} \frac{p(u,v) - p(u-1,v)}{\sqrt{p(u,v)^2 - p(u-1,v)^2}} + \frac{p(u,v) - p(u,v-1)}{\sqrt{p(u,v)^2 - p(u,v-1)^2}}\right] + \ldots =$$

$$= \sum_{i=1}^{N}\left[sign(p_{x,i}) + sign(p_{y,i})\right] + \ldots$$

- Here divisions by zero are automatically avoided – only "sign" is required -> computationally efficient!

Questions

- How many neighbor pixels do we have to consider to achieve a satisfying accuracy at low computational cost?
- Best norm, $||.||_1$ vs $||.||_2$?
- Best optimization method (SD+LS, EM, SG)?

TV in digital radiography…

Research in progress…

Results (answers)

- 75 simulated radiographs with different frequency content, corrupted by Poisson noise (max 15,000 photons).

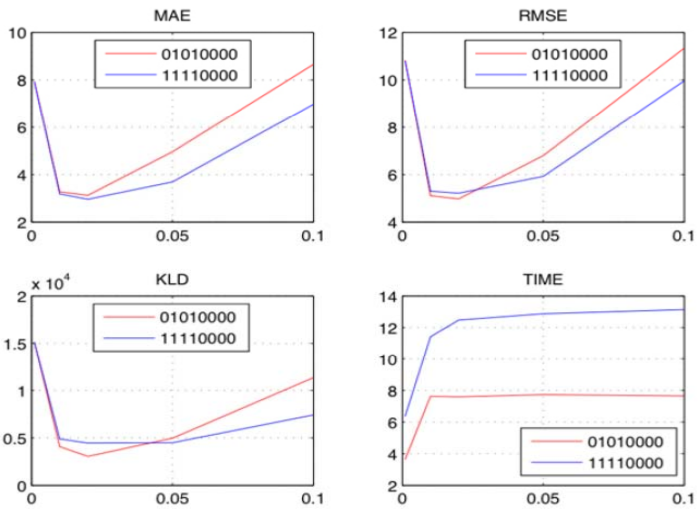
- For any filtered image, measure:

$MAE = 1/N\Sigma_{i=1..N}|p_{i,noisefree}-p_{i,filtered}|$

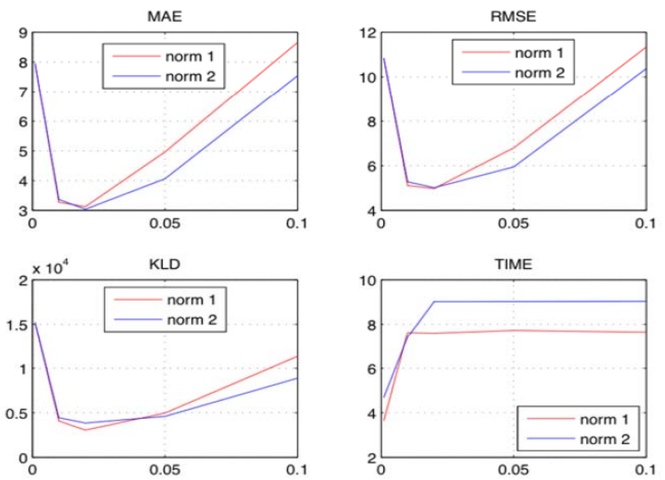
$RMSE = [1/N\Sigma_{i=1..N}(p_{i,noisefree}-p_{i,filtered})^2]^{1/2}$

$KL = \Sigma_{i=1..N}[p_{i,noisefree}\cdot\ln(p_{i,noisefree}/p_{i,filtered})+p_{i,noisefree}-p_{i,filtered}]$

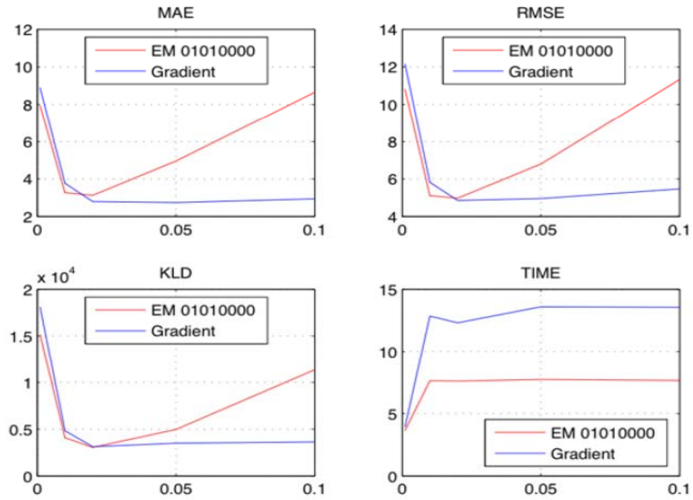
2 neighbors (01010000) vs. 4 neighbors (11110000)



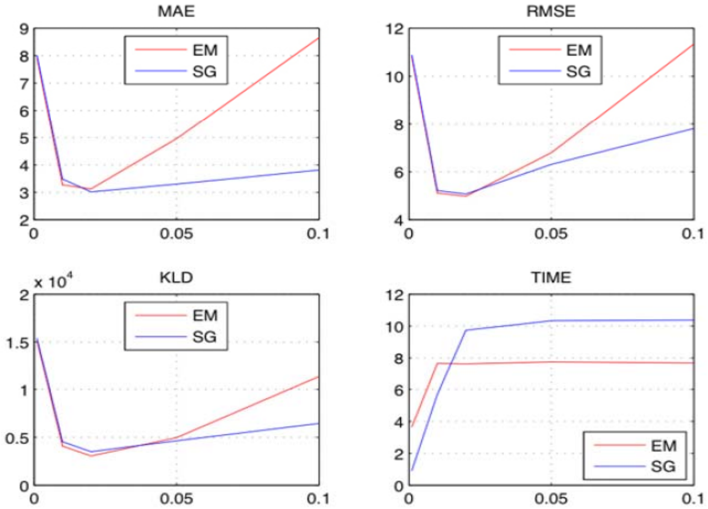
$\|.\|_1$ vs. $\|.\|_2$



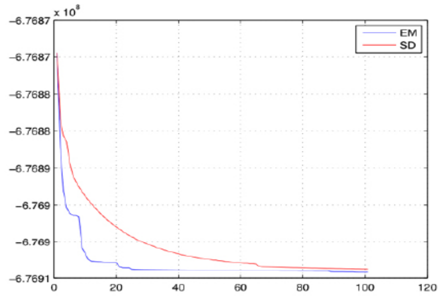
EM vs. SD+LS

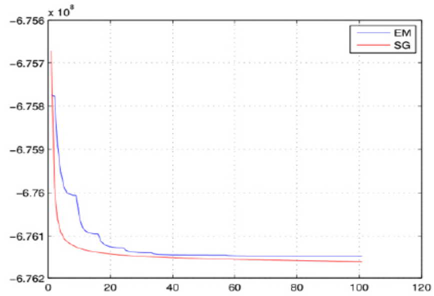


EM vs. SG



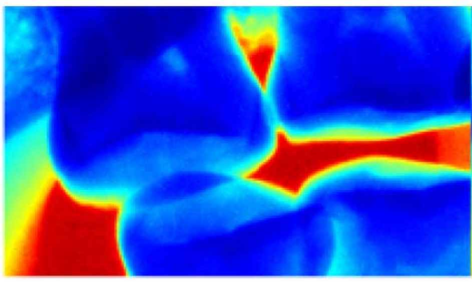
Convergence and iterations



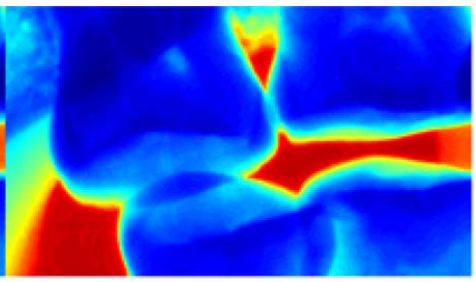


Filter effect

Original



Filtered



Filter effect: before filtering

Filter effect: after filtering

Conclusion

- Effective edge preserving filter;
- 2 neighbors, $||.||_1$ and EM achieve the best compromise between accuracy and computational cost;
- SD achieves results better then EM when the regularization parameter is not correctly selected.

- Adaptive regularization parameter;
- GPU (CUDA) implementation;
- Expanding the likelihood model
  - Mixture of Poisson, Gaussian and Impulsive noise;
  - Include the sensor point spread function.