A Quantitative Theory of the Non-Local Means Algorithm

N.A. Thacker[a*], P.A. Bromiley[a] and J.V. Manjon[b]

[a] Imaging Science and Biomedical Engineering, University of Manchester, U.K.
[b] IBIME Group, ITACA Institute, Universidad Politécnica de Valencia, Valencia, Spain.

Abstract. Noise filtering is a common step in image processing, and is particularly effective in improving the subjective quality of images. A number of techniques have been developed, many of which concentrate on the problem of removing noise without damaging small structures, such as edges. One recent approach demonstrating empirical merit is Non-Local Means (NLM). However, an understanding of the statistical basis of NLM is required before it can be used in quantitative image analysis. In this paper we investigate this basis in order to understand the conditions required for the use of NLM, testing the theory on simulated data and MR images of the normal brain.

1 Introduction

De-noising is a common pre-processing step in many MR image analysis tasks, such as segmentation and registration. Many filtering methods are based on the signal averaging principle, which uses the spatial redundancy in the image. Gaussian filters, for example, have been widely used in applications such as fMRI, but have the disadvantage of blurring edges due to averaging non-similar data. Edge-preserving filters that average data in the direction orthogonal to the local image gradient, such as anisotropic diffusion [1, 2], have been proposed to avoid this problem. However, such filters can erase small features and may change image statistics. Wavelet-based filters have also been applied to MRI de-noising [3], but tend to introduce characteristic artefacts that can be problematic for clinicians.

Non-Local Means (NLM) avoids destructive image modification by averaging only over data that is statistically similar on the basis of local context. The ath patch of N^2 pixels A surrounding each voxel g_a is compared to the bth patches B in other locations in the image, and the filtered value $\hat{g}_a$ computed by averaging over the central pixels g_b

$$\hat{g}_a = \sum_b g_b w(a,b) \propto \sum_b g_b exp(-d(a,b)/h^2) \tag{1}$$

in patches that are similar on the basis of some measure $d(a,b)$, typically

$$d(a,b) = \sum_i^N \sum_j^N G_\rho(r)(A_{ij} - B_{ij})^2 \tag{2}$$

where $w(a,b)$ are weighting factors, $G_\rho(r)$ is a radial Gaussian weighting from the centre of the patch with scale ρ, and h is a scale factor that determines the required degree of similarity. The effectiveness of NLM has been demonstrated empirically on clinical data [5], but its statistical foundations have not been investigated. One consequence is the presence of free parameters (e.g. h,ρ). Whilst this situation might be acceptable in applications where success is evaluated empirically, the application of NLM has to be carefully analysed in order to be considered adequate for quantitative or clinical tasks. An understanding of the algorithm as a statistical estimation task must be developed; in particular, the patch-matching and intensity-weighting processes must be related to conventional statistics.

2 Methods

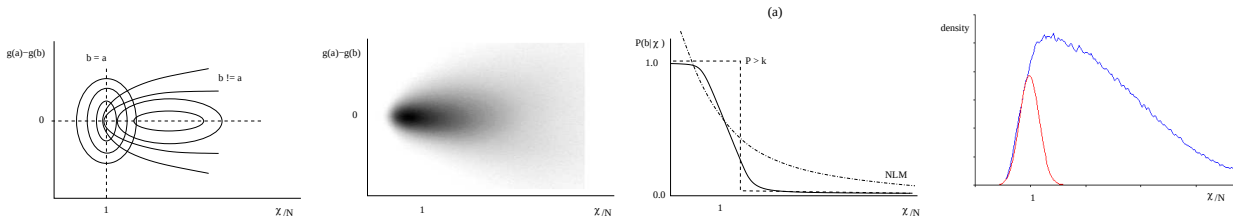
Eqs.1 and 2 do not map directly onto statistical estimation processes, and so we simplify them in order to make such as association. The aim here is not to evaluate the modified version in an algorithmic "shoot-out", but to determine the statistical characteristics of the original NLM algorithm through comparison with the modified version. First, Eq.1 is modified to conform to a standard χ^2 test for independent, identically distributed (IID) data

$$d(a,b) = \chi^2_{ab} = \frac{1}{2\sigma^2} \sum_i^N \sum_j^N (A_{ij} - B_{ij})^2 \tag{3}$$

where the definition of similarity is scaled by the standard deviation of the image noise σ. In addition we investigate a similarity function that allows one degree of freedom for intensity scale γ (see Appendix A)

$$\chi^2 \propto \sum_i^N \sum_j^N (\alpha A_{ij} - \beta B_{ij})^2 \quad \text{s.t.} \quad \alpha^2 + \beta^2 = 1 \quad \text{and} \quad \gamma = \frac{\alpha}{\beta} \approx \frac{2|A||B|}{(|A|^2 - |B|^2) + (|A|^2 + |B|^2)} = \frac{|B|}{|A|} \tag{4}$$

*E-mail: neil.thacker@man.ac.uk



(a) (b) (c) (d)

Figure 1. (a) The 2D distribution of difference between central intensity values as a function of the patch similarity measure, and (b) the equivalent plot for the brain image Fig.3a. (c) The marginal projection of the conditional probability of the data being drawn from the matching distribution (b=a), and (d) the marginal projection of the brain image data (blue) with the expected signal distribution (red).

Eq.2 is less straightforward to interpret. However, it is clearly an estimation task and so the weighting factors should be related to the probability of the intensities g_b being drawn from the same distribution as the intensity g_a, based on the similarity function. This implies that knowledge of the behaviour of the data in the joint space of intensity residuals and similarity is required. Fig.1 summarises this, showing difference between central patch intensities plotted against the patch similarity measure ($\sqrt{\chi^2/N^2} = \chi/N$) for corresponding ($b = a$) and non-corresponding ($b! = a$) patches. The χ^2 is a standard chi-square variable with mean N. Therefore, for the distribution $b = a$, the variable χ/N is drawn from a distribution that is well approximated by a Gaussian with mean 1.0 and variance $1/N$.

The process of weighted estimation embodied in Eq.2 is directly analogous to the parameter estimation (or Maximisation step) in the Expectation Maximisation algorithm [4]. The location of the mean value for $b = a$ would be the likelihood estimate of the noise-free data. The condition for convergence of EM requires that a weighted mean is constructed using the conditional probability of classification. Here, this corresponds to the intensity being drawn from the signal distribution $P(b = a|\chi_{ab}, g_a - g_b)$. The constraint this places on the estimated intensity must also take account of how well g_b predicts g_a (i.e. σ_{ab}). Clearly, replacing this 2D distribution with a 1D weight factor (Fig.1c), dependent only upon the similarity function as in NLM, such as

$$\hat{g}_a = \frac{\sum_b g_b P(b = a|\chi_{ab})/\sigma_{ab}^2}{\sum_b P(b = a|\chi_{ab})/\sigma_{ab}^2} \quad \text{rather than} \quad \hat{g}_a = \frac{\sum_b g_b P(b = a|\chi_{ab}, g_a - g_b)/\sigma_{ab}^2}{\sum_b P(b = a|\chi_{ab}, g_a - g_b)/\sigma_{ab}^2} \tag{5}$$

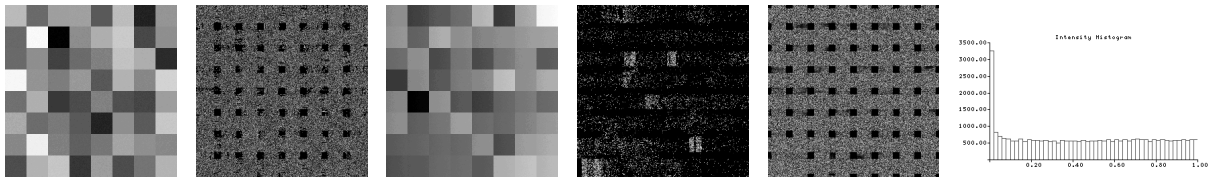
involves making a number of additional assumptions regarding the relationship between the full 2D distribution and its marginal projections. In particular, a 2D conditional probability would select data as a function of both variables and be likely to reject data with large $g_b - g_a$ due to the increased breadth of the non-match distribution. We can therefore justify putting a cut on this difference variable, as is often done in NLM implementations in the interest of computational efficiency (Appendix B).

Returning now to the standard NLM formulation, there are two simple cases we should consider as an alternative to exponential weighting (Fig.1c). If there is no overlap between the two distributions then using a single weighting factor of 1 for all data below a threshold is valid. However, data drawn from an approximately uniform continuum of possible patches (such as Fig.1b) has overlap and we can only say that the weight factor should be 1 for high patch similarity and drop to 0 for statistically dissimilar patches. In either case we can conclude that the exponential distribution embodied in Eq.2 is suboptimal for both low and high similarity scores.

In general, without knowledge of the non-match distribution ($b! = a$) we cannot estimate $w(b, a)$. However, knowledge of the expected match distribution $P(\chi_{ab}|b = a)$ and the total sample distribution $P(\chi_{ab})$ would allow us to compute this factor $P(b = a|\chi_{ab}) = P(\chi_{ab}|b = a)/P(\chi_{ab})$ (red curve divided by the blue in Fig.1d). We may assume (as in NLM) that the overlap between true signal and background is similar across the image, or estimate the actual distributions present at each location. In this work we take the χ^2 hypothesis probability ($P_{ab} = P(\chi_{ab}^2, N^2)$) as the weighting factor, as an approximation that is consistent with what we expect of the required function, and errs on the side of caution (i.e. it is a less efficient estimator) for situations of no overlap

$$w(a, b) \propto \frac{P_{ab}}{\sigma_{ab}^2}$$

This can be approximated using error function (erf()), and has the advantage that the weighting factor reduces to approximately zero for patches that are not equivalent within the expected variation due to noise. Calculation of this quantity also supports testing of the assumed noise distribution. Numerical implementation of the similarity calculation requires some care in order to avoid a six-deep nested loop (Appendix B).



(a) (b) (c) (d) (e) (f)

Figure 2. Simulated data without (a) illumination change, and the intensity change produced by filtering with Eq.6 (b). Simulated data with (c) illumination change and the intensity changes produced by filtering with Eq.6 (d), and with Eq.7 (e) using a scaled matching process. The distribution (f) of hypothesis probabilities for the scaled filtering is uniform for data drawn from the assumed distribution, as expected, with a peak at zero for unique locations.

3 Results

Three experiments were performed to compare the original and statistically based versions of NLM. First, a simulated image of 256x256 pixels containing 32x32 pixel regions with random intensities drawn from a Gaussian distribution of width 32 grey levels was constructed. Independent Gaussian random noise with a standard deviation of 1 grey level was added (Fig.2a). NLM was applied using Eq.3 as the similarity function, and filtered values estimated using

$$\hat{g}_a = \frac{\sum_b g_b P_{ab}/\sigma_{ab}^2}{\sum_b P_{ab}/\sigma_{ab}^2} = \frac{\sum_b g_b P_{ab}/\chi_{ab}^2}{\sum_b P_{ab}/\chi_{ab}^2} \tag{6}$$

where σ_{ab} is the standard deviation of the difference distribution for $g_b - g_a$. For matching patches we expect $\sigma_{ab} = \sigma$. The use of an estimate of local variance derived from patch similarity reduces the dependency of the algorithm on the assumed value of global image noise σ. Regions that do not match take no part in the weighted average and $P_{aa}/\chi_{aa} = 1$ by definition. Pixels that do not change their value upon application of this filter ($g_a - \hat{g_a} = 0$)[1] represent statistically unique locations. Matching image patches were sought in 23x23 locations about the central pixel g_a. The relative change in intensity produced by filtering using an 11x11 patches is shown in (Fig.2b). Notice that unique regions of the image (corners of each square patch) are generally left unmodified as expected.

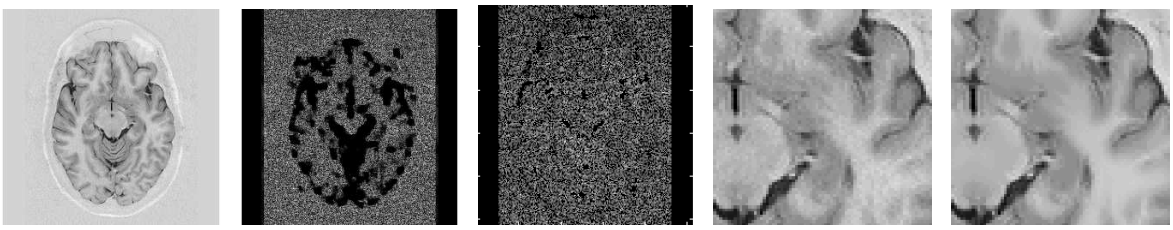
A second simulated image was then constructed in the same way, but adding a uniform intensity slope (1/4 grey-levels per pixel) prior to noise addition (Fig.2c). This simulates the processes of spatially varying illumination or field inhomogeneity in MR. Fig.2d shows the differences between the data and the result of the statistically based NLM filter; many more pixel locations are statistically unique and are left unmodified. Therefore, a free intensity scale parameter was introduced in the matching process, as discussed in the previous section

$$\hat{g}_a = \frac{\sum_b \gamma g_b P_{ab}/(\chi_{ab}^2(1+\gamma^2))}{\sum_b P_{ab}/(\chi_{ab}^2(1+\gamma^2))} \quad \text{where} \quad \chi_{ab}^2 = \frac{1}{\sigma^2(1+\gamma^2)} \sum_i^N \sum_j^N (\gamma A_{ij} - B_{ij})^2 \tag{7}$$

with γ calculated using $|B|/|A|$ for purposes of computational efficiency. χ_{ab}^2 is expected to be a χ^2 statistic with $N-1$ degrees of freedom. The variable $\sqrt{\chi^2/(N-1)}$ is drawn from a distribution that is approximately a Gaussian with mean 1.0 and a variance $\kappa/(N-1)$. A value of $\kappa = 0.9$ was set empirically to adjust for the extra instability introduced into the variable due to error in γ (a similar problem to that accommodated by the Student t-test). The weighted averaging takes appropriate account of the accuracy of each estimate (γg_b), such that each data point is weighted by the expected difference between the sample and the underlying true value. Fig.2e shows the intensity changes produced by filtering: addition of the scale parameter produces results comparable to (Fig.2b), with only unique locations being left unmodified. The quantitative validity of the method can be tested by observing the histogram of hypothesis probabilities for the data (Fig.2e). The distribution is uniform, as expected for a hypothesis probability [6], except for the peak at zero that corresponds to non-equivalent regions. This also validates our choice of κ and the approximation for γ.

The third experiment involved application of the scale-adjusted filtering scheme to a real 256x256 pixel MR image of the brain (Fig.3a). Noise was directly estimated from the image in the region of the brain and patch similarities computed over a 7x7 region. The magnitude of pixel change due to filtering is shown in Fig.3b and illustrates that, as in the simulated data, many locations are left unmodified on the basis of statistical uniqueness. A magnified portion of this image and the filtered output is shown in Fig.3d and e. This process can be considered a safe strategy for

[1]Final results are represented as integer values so that this definition identifies unique locations as a lack of change at a level much less than intrinsic image noise.



(a) (b) (c) (d) (e)

Figure 3. MR image of a normal subject taken at the level of the top of the brainstem (a), and the change in each pixel (b) produced by statistically based filtering with scale estimation. (c) shows the changes produced by a conventional NLM filter for comparison. Magnified regions illustrating noise filtering are shown in (d) and (e). Despite many areas being left unmodified there is little subjective evidence of under-filtering.

noise filtering in these images. Data is only modified when each estimate of the central pixel comes from a region that satisfies the hypothesis test. A less conservative weighting, such as weighting with the likelihood $exp(-\chi^2)$, as suggested in the original papers, removes noise even in unique regions (Fig.3c) even without adjusting for local grey-level scale changes. This implies that filtered estimates must necessarily include data (γg_b) from patches that cannot justifiably be expected to be from the same distribution as the central pixel.

4 Discussion and Conclusions

The statistically based filtering scheme described here differs from the original NLM method in several ways. First, spatial Gaussian weighting is not used in the similarity measure, implying a change in the statistical variables. In order for the new formulation to remain within the realm of conventional statistics (and the principles of quantitative probability required for scientific validity) the assumption of uniform independent Gaussian noise must be applicable to these new variables. This is an assumption regarding the differences between matching patches, rather than the intrinsic measurement process, and can only be justified if an effort is made to confirm the distributions, which may vary between data sets, empirically. Second, there is a significant difference in the way that weighted estimates are computed (Eq.5). Weighting with P_{ab} eliminates data that cannot be statistically equivalent and produces extensive regions of unmodified pixels, where no equivalent data can be found, in both simulated and real images. The situation is exacerbated by illumination variation. However, incorporating estimation of intensity variation into the intensity comparison proved sufficient to recover many matches without invalidating the quantitative interpretation of the method.

Even using scale estimation, real data exhibits unmodified regions of pixels, consistent with unique intensity structures. These areas can be reduced by relaxing the similarity requirements, for example by reducing the region over which similarity is computed. In previous work we found that the proportion of non-filtered pixels in the brain image could be reduced from 30% to 15% by reducing the region from 7x7 to 5x5 pixels. However, this is achieved at the expense of reducing the power of the technique to combine statistically similar data, and so does not necessarily represent an improvement in performance. More conventional approaches using exponential weight factors (i.e. $exp(-\chi^2)$) and spatial Gaussian weighting, have fewer regions of unmodified pixels. Our analysis therefore implies that conventional NLM combines non-equivalent data: intensities are modified by averaging over data from different distributions, potentially introducing bias and making the conventional technique suboptimal for clinical or quantitative applications. The presence of unfiltered pixels in all of the NLM versions studied also implies that the noise distribution in the output data will be spatially varying. The root cause of this is the demand for strict identicality between image patches. This could be alleviated by processing the data in a more sophisticated manner, such as constructing an eigenvector model of patch variation using PCA, as done in appearance modelling. We intend to investigate this in future work.

An analogy can be identified between NLM and biological vision. There is evidence that the brain analyses micro-textures by identifying common patterns (templates) and representing image data as the conjunction of these. The idea that matching processes occur early in vision is well established, and this explanation supports tasks such as "pop-out", where unusual image structure is immediately identified. If this is the case, then the representation is itself a noise-filtering process analogous to NLM, although perhaps with some pre-learning of the templates. Otherwise, the computational processes required by NLM are biologically plausible i.e. computable using the local calculations though to be supported by neurons. The observation that humans can easily identify the image noise (c.f. Fig.5) provides additional psychophysical support for this hypothesis. The noise filtering schemes that provide the best empirical performance will therefore be those that, perhaps like NLM, correspond closely to the visual system. In clinical situations, therefore, such filters will not destroy information that would be considered significant by a radiologist.

Appendix A: Estimating Normalisation

Application of the variational principle to the problem of matching two scaled noisy image patches I and J gives

$$\chi^2 \propto \sum_n (\alpha I_n - \beta J_n)^2 \quad \text{s.t.} \quad \alpha^2 + \beta^2 = 1$$

Putting $\alpha = sin(\theta)$ and $\beta = \cos(\theta)$ we can determine the scaling that minimises the χ^2 using

$$\frac{\partial \chi^2}{\partial \theta} \propto 2 \sum_n (cos(\theta) I_n + sin(\theta) J_n)(sin(\theta) I_n - cos(\theta) J_n) = 0$$

so that the minimum is defined according to

$$tan(2\theta) = \frac{2 \sum_n I_n J_n}{\sum_n I_n^2 - \sum_n J_n^2}$$

The denominator and numerator of this expression can be considered as the sides of a right-angled triangle, allowing us to determine $cos(2\theta)$ and $sin(2\theta)$.

$$cos(2\theta) = \frac{\sum_n (I_n^2 - J_n^2)}{\sqrt{4(\sum_n I_n J_n)^2 + (\sum_n I_n^2 - J_n^2)^2}} \quad \text{and} \quad sin(2\theta) = \frac{2 \sum_n I_n J_n}{\sqrt{4(\sum_n I_n J_n)^2 + (\sum_n I_n^2 - J_n^2)^2}}$$

The factor we need to rescale image I by in order to minimise the χ^2 is $\gamma = sin(\theta)/cos(\theta)$. Using the identities $sin(\theta) = sin(2\theta)/2cos(\theta)$ and $cos(\theta) = \sqrt{(cos(2\theta) + 1)/2}$ we get

$$\gamma = \frac{sin(2\theta)}{cos(2\theta) + 1} = \frac{2|I||J|}{|I|^2 - |J|^2 + \sqrt{4|I|^2|J|^2 cos^2\phi + (|I|^2 - |J|^2)^2}}$$

where $\phi \approx 0$ for similar patches.

Appendix B: Numerical Details

Implementing the patch similarity measure using

$$\chi_{ij}^2 = \frac{1}{\sigma^2(1+\gamma^2)} \sum_n^N (\gamma I_n - J_n)^2 \Rightarrow \chi_{ij}^2 = \frac{1}{\sigma^2(1+\gamma^2)} \left[\gamma^2 \sum_n I_n^2 + \sum_n J_n^2 - 2\gamma \sum_n I_n J_n \right]$$

and computing the quantities

$$A = \sum^n I_n^2 \quad B = \sum_n J_n^2 \quad \text{and} \quad C = \sum_n I_n J_n$$

from differences in precomputed integral images[2] of I^2, J^2 and IJ, allows the efficient calculation of $\gamma = \sqrt{B/A}$ (and all other required quantities) for one hypothesised shift of the match patch across the entire image. This allows implementation as a 4- rather than 6-deep nested loop, reducing execution time from hours to minutes. Additional efficiency improvements were gained by rejecting scale values that do not lie in the range $0.5 < \gamma < 2.0$, and also by rejecting new estimates γg_i to the weighted sum when $|\gamma g_i - g_j|/\sigma > 8.0$.

References

1. G. Gerig, O. Kubler, R. Kikinis, and F. A Jolesz. Nonlinear Anisotropic Filtering of MRI Data. IEEE Trans. Med Imag., vol. 11, pp. 221-232, 1992.
2. A. Samsonov and C. Johnson. Noise-Adaptive Nonlinear Diffusion Filtering of MR Images With Spatially Varying Noise Levels. Magnetic Resonance in Medicine 52:798-806, 2004.
3. R. D. Nowak, Wavelet-based Rician noise removal for magnetic resonance imaging, IEEE Transactions on Image Processing, 8:10, 1408 -1419, 1999.
4. A.P. Dempster, N.M. Laird & D.B. Rubin, "Maximum Likelihood from Incomplete Data via EM Algorithm", *Journal of the Royal Society*, **39**, pp.1-38, 1977.
5. J.V. Manjon, M. Robles, N.A. Thacker, Multispectral MRI De-noising Using Non-local Means. Proc. MIUA 2007, 41-46, Aberystwyth, 2007.
6. P.A. Bromiley, N.A. Thacker and P. Courtney, Non-Parametric Image Subtraction for MRI. Proc. MIUA 2001, 105-108, Birmingham, 2001.

[2] Stored in double precision to avoid numerical issues.