A Statistical Interpretation of Non-Local Means

N.A. Thacker, J.V. Manjon and P.A. Bromiley

Last updated
18/09/2009

TiNA
WWW.TINA-VISION.NET

Imaging Science and Biomedical Engineering Division,
Medical School, University of Manchester,
Stopford Building, Oxford Road,
Manchester, M13 9PT.

A Statistical Interpretation of Non-Local Means

N.A. Thacker[1], J.V. Manjon[2] and P.A. Bromiley[1]

1: Imaging Science and Biomedical Engineering Division, Medical School,
University of Manchester, Manchester, M13 9PT, UK

2: ITACA Institute, Universidad Politécnica de Valencia, Valencia, Spain.

neil.thacker@manchester.ac.uk

Abstract

Noise filtering is a common step in image processing, and is particularly effective in improving the subjective quality of images. A large 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 that demonstrates empirical merit is the non-local means (NLM) algorithm. However, in order to use noise filtering algorithms in quantitative or clinical image analysis tasks an understanding of their behaviour that goes beyond subjective appearance must be developed. The purpose of this paper is to investigate the statistical basis of NLM in order to attempt to understand the conditions required for its use. The theory is illustrated on synthetic data and clinical MR images of the brain.

1 Introduction

Noise filtering is a common pre-processing step in many magnetic resonance (MR) image processing and analysis tasks, such as segmentation or registration. A wide variety of noise filtering algorithms have been proposed [5], many of which operate by weighted averaging across local data on the basis that, if the noise on each pixel is uncorrelated or weakly correlated, then the spatial scale of variation in the underlying signal is typically larger than that of the noise; averaging thus removes noise whilst preserving signal. Gaussian filters fall into this category and have been widely used in some applications such as functional MRI [20]. However, the assumption of slow spatial variation in the signal will be violated at points such as edges, and filters that are insensitive to this will destructively modify the image at such points due to averaging across data drawn from different intensity distributions. The characteristic blurring of edges resulting from Gaussian filtering is an example of this effect. Edge-preserving filters have been proposed to avoid this problem, the best known of which is the anisotropic diffusion filter (ADF) [15]. Such filters respect edges by averaging in the direction orthogonal to the local image gradient. However, they can erase small features and may change image statistics. Wavelet-based filters have also been applied to MRI [13] but tend to introduce characteristic artifacts [5] that can be problematic for clinicians.

In order to mitigate destructive image modification, algorithms have been proposed that incorporate explicit tests of similarity between the data to be averaged. This may be implemented either in a feature space, as in the mean shift algorithm [6], or in the image domain, as in bilateral filtering [19], in which the averaging process incorporates weighting using both the spatial distance between pixels and the difference in intensity. The mean shift, anisotropic diffusion and bilateral filtering algorithms have been shown to be closely related [1]. The NLM algorithm [5] develops this concept further by defining similarity on the basis of local context i.e. the intensities in patches surrounding each pixel, rather than using only the intensities of the pixels themselves. Furthermore, in NLM the pixel redundancy is not restricted to be local i.e. pixels are not penalised on the basis of geometrical distance as happens for example in the bilateral filter. The intensity g_a of each pixel is modified by comparing the surrounding patch $a = A_{i=1...N,j=1...N}$ to other patches $b = B_{i=1...N,j=1...N}$ across the image, using a similarity measure such as

$$d(a,b) = \sum_i^N \sum_j^N G_\rho(r)(A_{ij} - B_{ij})^2 \tag{1}$$

where $G_\rho(r)$ is a radial Gaussian weighting dependent on the distance r from the centre of the patch, ρ is the scale parameter, N^2 is the size of the patch, and i and j are iterators over the x and y dimensions of the image. The noise-filtered intensity $\hat{g}_a$ is then estimated as a sum of the intensity g_b of the central pixel in each of the other

patches B, weighted by their similarity

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

where h is a scale factor that determines the required degree of similarity and w is the weighting factor. The weights are normalised to satisfy the usual conditions of $0 \le w(a,b) \le 1$ and $\sum_b w(a,b) = 1$. Implementations of NLM typically introduce two corrections in order to avoid over-weighting effects. During the calculation of $d(a,b)$, the radial Gaussian weighting for the central intensity in each patch is set equal to that for intensities one pixel away. Furthermore, $w(a,a)$ is set equal to the maximum of the $w(a,b)$ for the calculation of each filtered intensity.

The original description of NLM calls for the averaging process to incorporate all pixels in the image. However, this leads to impractical processor time requirements. Therefore, on the basis that similar patches are more likely to be found in the immediate vicinity of the pixel being filtered, the averaging process can be limited to an $S \times S$ search window where S is smaller than the dimensions of the image [7]. Furthermore, the majority of the computational effort is taken up in the calculation of $d(a,b)$; some authors (e.g. [10]) have proposed limiting this calculation to patches that satisfy a threshold on local intensity mean and gradient.

It has been found previously that NLM performs significantly better than other state-of-the-art noise filtering algorithms in empirical tests on natural images, and that the performance of the algorithm is related to how similar the noise is to Gaussian noise [5]. This implies that NLM should be applicable to noise filtering in medical images such as MR, where the noise is expected to be Gaussian or Rician [9], and the algorithm has been shown to be very effective in empirical tests on clinical data [11, 12]. However, it appears that the theoretical basis of the algorithm has not been fully understood in statistical terms. One consequence is the presence of several free parameters (in particular h and ρ) that affect the results. This is not a significant issue in applications where the visual assessment of the results is the dominant consideration. However, a fundamentally deeper understanding of the algorithm is therefore required before it can be used with confidence in quantitative or clinical applications. As with all estimation tasks, this requires identification of the assumptions necessary to derive it from statistical estimation theory. In particular, the patch similarity and intensity weighting processes must be understood in terms of conventional statistics.

The remainder of this paper is organised as follows. A statistical interpretation of the NLM algorithm is developed in Section 2 by drawing analogies between the patch similarity calculation and a standard χ^2 similarity test, and between the intensity weighting process and a χ^2 probability calculation. Section 3 describes the methodology adopted to evaluate the statistically motivated version of the NLM algorithm. Results from the evaluation are presented in Section 4, and our conclusions and suggestions for further work are presented in Section 5.

2 Statistical Interpretation of NLM

The NLM algorithm embodied in Eqs. 1 and 2 does not map directly onto statistical estimation processes, but can be simplified in order to make such an 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. The purpose of Eq. 1 is to measure the similarity between image patches a and b, and it is immediately clear that the standard quantitative χ^2 test provides the statistical equivalent

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

which is valid for independent, identically distributed (IID) data. In addition, an equivalent function that allows the overall intensity scale to vary between the patches

$$\chi^2 \propto \sum_i^N \sum_j^N (\alpha A_{ij} - \beta B_{ij})^2 \quad s.t. \quad \alpha^2 + \beta^2 = 1 \tag{4}$$

will also be considered. The required scaling parameter $\gamma = \alpha/\beta$ is given by (Appendix A)

$$\gamma \approx \frac{2|A||B|}{(|A|^2 - |B|^2) + (|A|^2 + |B|^2)} = \frac{|B|}{|A|}$$

Equation 2 is less straightforward to interpret. However, the NLM algorithm is clearly an estimator of the noise-free intensity of each pixel. This implies that the weighting of the central intensity in patch b should be related
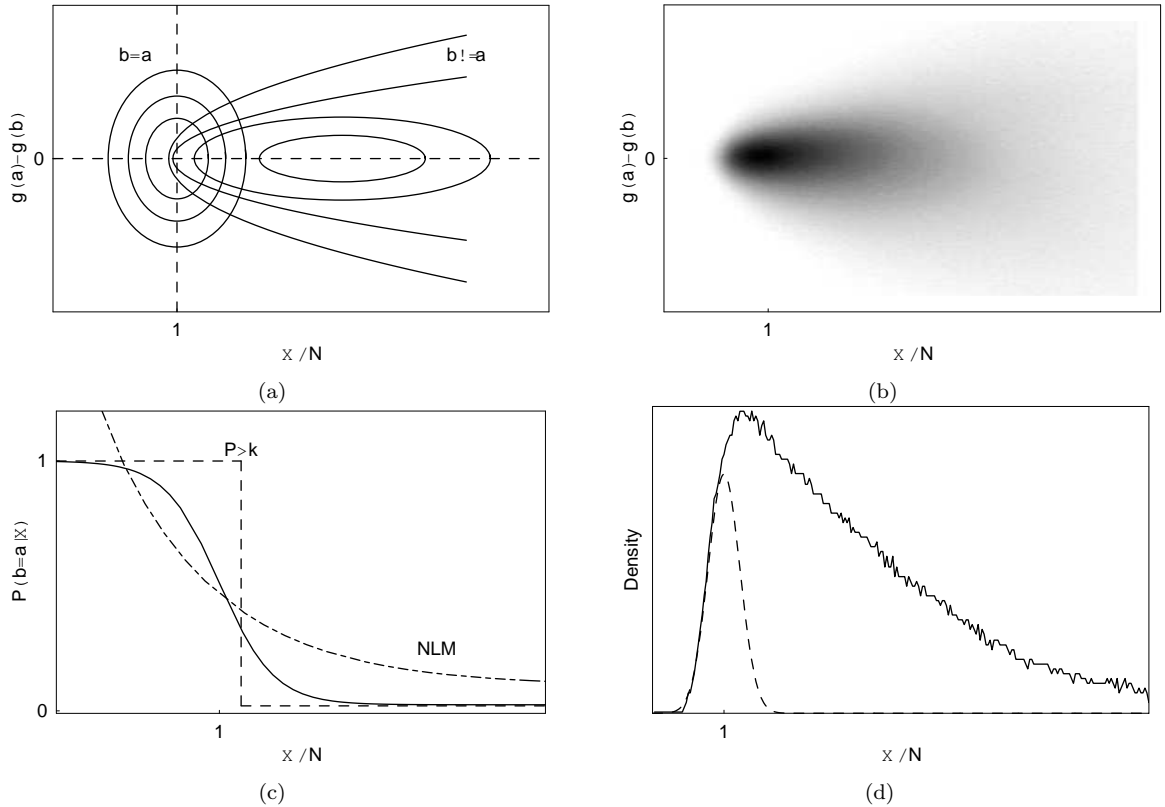
Figure 1: (a) The 2D distribution of differences between central intensity values as a function of the patch similarity measure, and (b) the equivalent plot for the MR brain image Fig. 6a. (c) The marginal projection of the conditional probability of the data being drawn from the matching distribution (b=a, solid line), showing the exponential weighting process embodied in NLM and the binary thresholding applicable in cases of no overlap between the matching and non-matching distributions. (d) The marginal projection of the brain image data (solid curve) with the expected signal distribution (dashed).

to the probability that it was drawn from the same intensity distribution as the central intensity in patch a, calculated using the similarity measure. Therefore, the calculation of this probability in the general case would require knowledge of the distributions both for matching patches (i.e. patches drawn from equivalent structures) and non-matching patches in the joint space of intensity (or, more specifically, intensity residual $g_a - g_b$) and similarity. Fig. 1a illustrates the concept; the χ^2 similarity measure is a standard chi-squared variable with N^2 degrees of freedom, since this is the number of pixels in each patch, and so for matching patches the variable $\sqrt{\chi^2/N^2} = \chi/N$ is drawn from a distribution that is approximately Gaussian with mean 1.0 and variance $1/N$ if $N \gtrsim 5$ [2]. The distribution in $g_a - g_b$ for matching patches will be identical to the noise distribution. However, the distributions for non-matching patches will be dependent on image structure and so are, in general, unknown.

If Eq. 2 is replaced with an estimator of $\hat{g}_a$ constructed from a weighted mean using the conditional probability of classification $P(a = b|data)$, then the result is directly analogous to the parameter estimation (maximisation) step of the Expectation-Maximisation (EM) algorithm [8]; the result is therefore a maximum-likelihood estimate of the noise-free intensity. As the aim is to test the equivalence of the central pixel intensities in each patch, the data used in this process should in theory take account of both the patch similarity measure and the central pixel intensity difference. However, the original specification of NLM uses a weight that is determined from the similarity measure alone i.e. is analogous to

$$\hat{g}_a = \frac{\sum_b g_b P(b = a|\chi_{ab})/\sigma_{ab}^2}{\sum_b P(b = a|\chi_{ab})/\sigma_{ab}^2} \tag{5}$$

rather than a probability of classification conditioned on both similarity and intensity residual

$$\hat{g}_a = \frac{\sum_b g_b P(b = a|\chi_{ab}, g_a - g_b)/\sigma_{ab}^2}{\sum_b P(b = a|\chi_{ab}, g_a - g_b)/\sigma_{ab}^2} \tag{6}$$

where σ_{ab} expresses how well g_b predicts g_a i.e. the width of the conditional probability distribution. Using this marginal projection, rather than the full 2D conditional probability, involves making a number of additional

assumptions about their relationship. In particular, the 2D conditional probability would give a lower weight to patches with a high intensity residual due to the increase in the width of the distribution in $g_a - g_b$ as a function of similarity (Fig. 1b).

Several methods for calculating $P(b = a|\chi_{ab})$ could be envisaged as alternatives to the exponential weighting used in NLM. If there was no overlap between the match and non-match distributions, then a simple binary threshold would suffice (Fig. 1c). However, as shown in Fig. 1d, the distributions overlap in clinical data, implying that the weight should be 1 for high patch similarity and drop to zero for statistically dissimilar patches; in either case exponential weighting is sub-optimal at both high and low similarity. In general the optimal weighting cannot be determined without knowledge of both the match and non-match distributions. Knowledge of the expected match distribution $P(\chi_{ab}|b = a)$ and the total sample distribution $P(\chi_{ab})$ would allow computation of $P(b = a|\chi_{ab}) = P(\chi_{ab}|b = a)/P(\chi_{ab})$ (the dashed curve divided by the solid curve in Fig. 1d). Use of this as a weighting scheme would require either the assumption that the overlap between the true signal and the background was similar across the image, which is implicit in the original formulation of NLM, or estimation of the actual distributions present at each point in the image. In order to avoid this, in the work presented here the standard χ^2 hypothesis probability P_{ab} [2] is used

$$w(a,b) \propto \frac{P_{ab}}{\sigma_{ab}^2} \quad where \quad P_{ab} = \int_0^{\chi^2} P(\chi_{ab}^2|N^2)d\chi^2 = \int_0^{\chi^2} \frac{2^{-N^2/2}}{\Gamma(N^2/2)} \chi^{N^2-2} e^{-\chi^2/2} d\chi^2$$

where $\Gamma(x)$ is the standard gamma function. This is an approximation that is consistent with the expected behaviour of the required function and is a less efficient estimator (i.e. errs on the side of caution) for situations of no overlap. This weighting can be approximated using the error function (erf()) [16] in order to avoid explicit numerical integration of the distribution, and has the advantage that the weighting factor reduces to approximately zero for matching patches that are not equivalent within the expected variation due to noise. Calculation of this quantity also supports testing of the assumed noise distribution. Its numerical implementation requires some care in order to avoid a six-deep nested loop (Appendix B).

3 Evaluation Methodology

3.1 Simulated Data

Several stages of evaluation were performed, using implementations of both the original and statistically motivated variants of NLM within the TINA open-source medical image analysis software (www.tina-vision.net). The first used simulated data sets in order to illustrate the effects of including a free scale parameter in the statistically-motivated version of NLM. First, a 256x256 pixel simulated image was generated, consisting of 32x32 pixel regions with intensities drawn randomly from a Gaussian distribution with a standard deviation of 32 grey-levels. A Gaussian random noise field with a standard deviation of 1 grey level was then added; the resulting image is shown in Fig. 2a. The statistically motivated version of NLM was applied to this image using the similarity measure given by Eq. 3, with σ set to the standard deviation of the image noise in order to scale the definition of similarity according to the level of evidence available in the image. Filtered intensities were estimated using

$$\hat{g}_a = \frac{\sum_b g_b P_{ab}/\sigma_{ab}^2}{\sum_b P_{ab}/\sigma_{ab}^2} = \frac{\sum_b g_b P_{ab}/\chi_{ab}^2}{\sum_b P_{ab}/\chi_{ab}^2} \tag{7}$$

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 $N^2 P_{aa}/\chi_{aa}^2 = 1$ 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 of 11x11 pixels were sought in a 23x23 pixel region about the central pixel g_a.

The similarity measure given by Eq. 3 represents a simple approach to determining correspondence. However, many natural images exhibit spatially varying illumination, which interferes with similarity measures based on exact matching of intensities. A similar process occurs in the majority of MR images in the form of intensity inhomogeneity introduced by variations in the imposed (B_0) magnetic field. Therefore, a second and more realistic data set was produced by adding a uniform slope in intensity (1/4 grey-levels per pixel) to the image prior to the

---

[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.
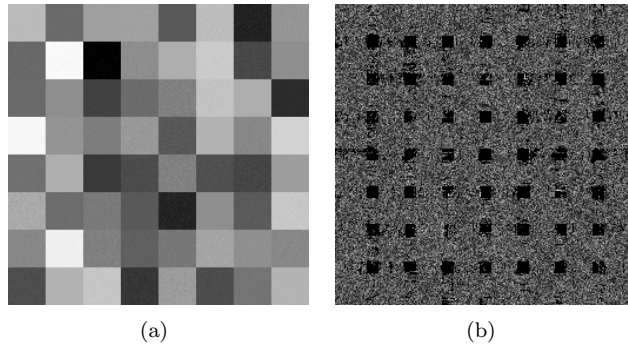
Figure 2: (a) Simulated data with no illumination variation and (b) consequent change in input values following filtering with Eq. 7.
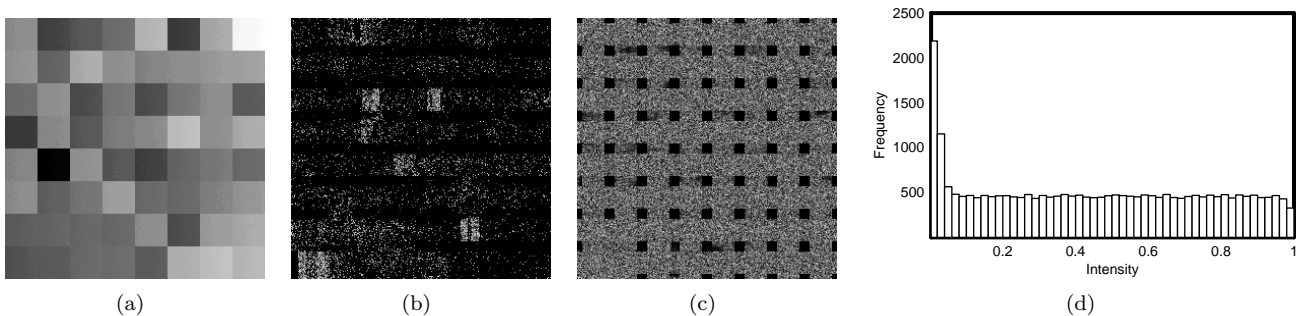

Figure 3: (a) Data with illumination variation changing left to right and (b) the resulting change due to filtering with Eq. 7. (c) The resulting change due to filtering with Eq. 9, incorporating scale-adjustment. (d) The distribution of hypothesis probabilities for (c) is uniform, as expected, for data drawn from the assumed distribution, with a peak at zero corresponding to unique locations.

addition of noise; the result is shown in Fig. 3a. In order to compensate for the intensity slope, image regions were matched using the similarity measure incorporating a free scale parameter as described above

$$\chi_{ab}^2 = \frac{1}{\sigma^2(1+\gamma^2)} \sum_i^N \sum_j^N (\gamma A_{ij} - B_{ij})^2 \tag{8}$$

with γ calculated using $|B|/|A|$ for purposes of computational efficiency. This was not found to result in an observable change in the sampled distributions (see below). This variable is expected to be a χ^2 statistic with $N^2 - 1$ degrees of freedom. The variable $\sqrt{\chi^2/(N^2-1)}$ is drawn from a distribution that is approximately a Gaussian with mean 1.0 and a variance $\kappa/\sqrt{(N^2-1)}$. The value of $\kappa = 0.9$ was set to adjust empirically for the extra instability introduced into the variable due to error in γ (a similar problem to that accommodated by the Student T-test). Weighted averaging was computed using

$$\hat{g}_a = \frac{\sum_b \gamma g_b P_{ab}/(\chi_{ab}^2(1+\gamma^2))}{\sum_b P_{ab}/(\chi_{ab}^2(1+\gamma^2))} \tag{9}$$

This 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.

3.2 Clinical MR Data

The differences between the behaviour of the original NLM algorithm represented by Eqs. 1 and 2 and the scale-adjusted, statistically motivated variant were demonstrated using a clinical MR image volume of the normal brain acquired with informed consent and subject to local ethics committee approval. A 1.5-T system (ACS-NT, with PowerTrack 6000 gradient subsystem; Philips Medical Systems, Hamburg, Germany) with a birdcage head coil receiver was used. A fast spin-echo inversion-recovery (IRTSE) image volume (repetition time, 6850 msec; echo time, 18 msec; inversion time, 300msec; echo train length, 9) was obtained in contiguous 3-mm thick sections throughout the brain, with an in-plane resolution of $0.89mm^2$ (matrix, 256×204, field of view, $230 \times 184mm$),

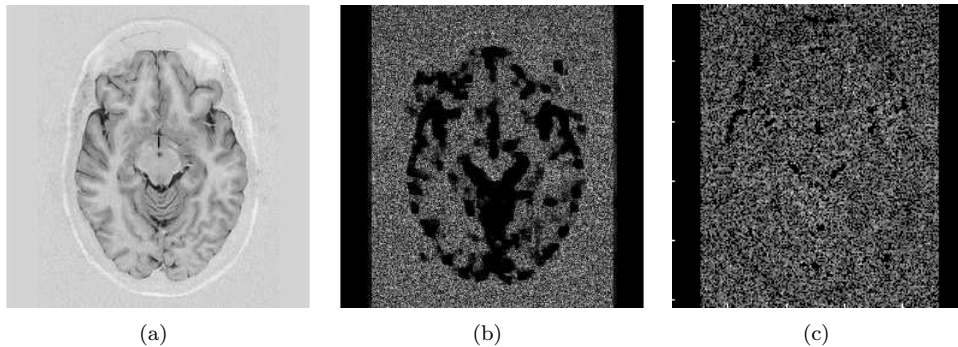<div style="text-align:center">(a)   (b)   (c)</div>

Figure 4: (a) IRTSE MR image and (b) intensity changes due to filtering with Eq. 9. (c) The intensity changes due to filtering with the original NLM algorithm for comparison.

and real image reconstruction was performed. A single structure-rich slice from the volume, shown in Fig. 4a, was chosen for analysis and both the original NLM algorithm and the statistically motivated variant incorporating scale adjustment were applied. Noise was estimated directly ([14]) from the image in the region of the brain and patch similarities computed over a 7x7 region (shown in previous work [11] to provide an optimal balance between accuracy and required processor power for the original NLM algorithm).

3.3 Quantitative Evaluation of Non-Local Means

The final stage of evaluation focused on quantitative analysis of the performance of NLM. Three additional MR image volumes, a variable echo proton density (VE(PD)), a variable echo T2 (VE(T2)) and a fluid-attenuated inversion recovery (FLAIR), were acquired from the same subject and scanner described above; again, single structure-rich slices were chosen from equivalent positions in each volume for analysis. The images are shown in Fig. 6. The original and scale-adjusted, statistically motivated variants of NLM were applied to all four images. Two performance measures were applied to the results. First, a Monte-Carlo stability analysis [18], which measured the relative change in the output image intensities produced by the addition of a small amount of noise to the input images, was applied to estimate the fraction of noise remaining after filtering. Second, the number of reconstructed pixels whose intensity was modified by more than three standard deviations of the image noise was counted, after compensating for local field inhomogeneity using the algorithm described by [17]. This measure is referred to as the residual outlier measure (ROM) and, since the expected value for a perfect noise filter can be calculated as the two-tailed integral of the noise distribution beyond 3σ, any deviation from this expected value quantifies the number of pixels to which inappropriate smoothing is applied. The local image noise was estimated independently of the Monte-Carlo stability analysis using a technique based upon the distribution of local derivatives [14]. In order to provide a benchmark for the evaluation, it was also applied to three conventional noise filtering schemes. The first, tangential filtering [4], is the simplest possible version of anisotropic diffusion and applies averaging over three pixels (one central and two either side) along the normal to the direction of maximum local image gradient. In the absence of noise, the gradient along this normal is expected to be zero in any image composed of smooth, continuous regions. Since many medical image modalities produce images that conform to this behaviour, tangential smoothing is theoretically the least destructive (i.e. best edge-preserving) of the simple noise filtering schemes (where simple is used in the sense that all pixels are treated equally). Gaussian filtering using a kernel with a standard deviation of 1 pixel and median filtering over the local neighbourhood of 9 pixels were also used.

4 Results

4.1 Simulated Data

Figures 2a and 3a show the simulated data set without and with a smoothly varying intensity change respectively; the intensity changes introduced by smoothing with the statistically motivated variant of NLM without a free scale parameter, using the similarity measure described by Eq. 3, are shown in Figs. 2b and 3b. Note that in Fig. 2b the unique regions of the image (corners of each square patch) are generally left unmodified (0 difference) as expected i.e. no similar patches can be found on the basis of the similarity measure, and so no smoothing is applied at these locations. In Fig. 3b this behaviour is observed across the entirety of the image; the similarity measure described by Eq. 3 and based on exact matching of intensity values is incapable of finding matching patches in the presence
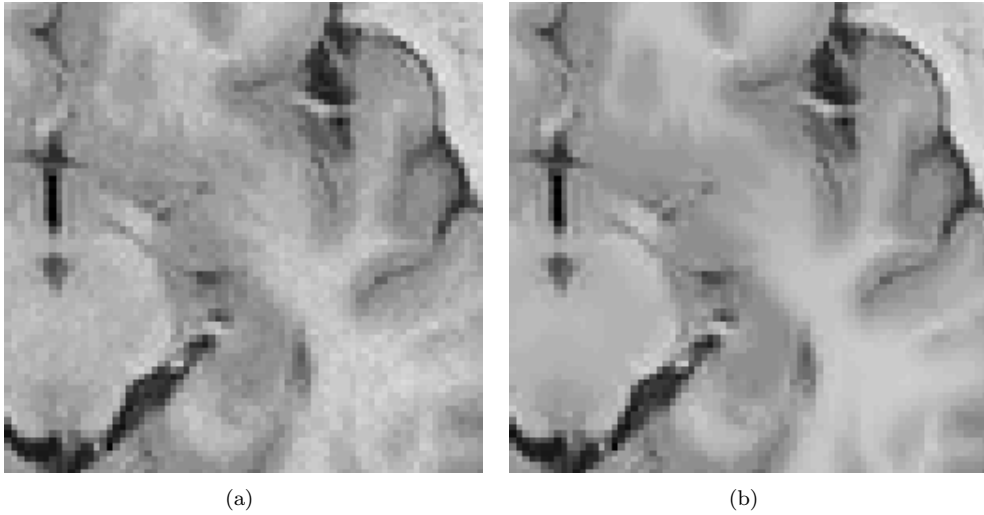
<center>(a)   (b)</center>

Figure 5: Magnified regions from Fig. 4 (a) before filtering and (b) after filtering with Eq. 9. Despite many areas being left unmodified there is little subjective evidence of under-filtering.

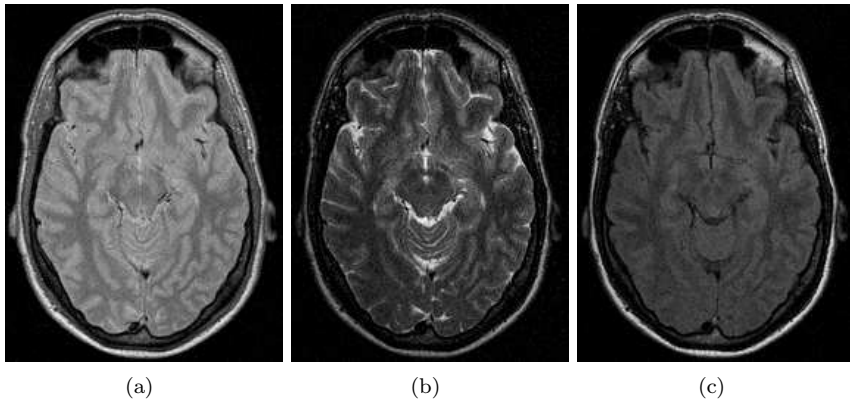<center>(a)   (b)   (c)</center>

Figure 6: Clinical MR data used in addition to Fig. 4a: (a) VE(PD); (b) VE(T2); (c) FLAIR.

of the uniform intensity variation, and so little smoothing is applied at any location. The differences introduced by smoothing Fig. 3a with the statistically motivated variant of NLM incorporating scale adjustment, using the similarity measure described by Eq. 8, is shown in Fig. 3c. Following the scaling modification the filtering results are comparable to the original result Fig. 2b, with only unique locations being left unmodified. The quantitative validity of the method can be tested by observing the sample distribution (histogram) of hypothesis probabilities for this data, shown in Fig. 3d. The distribution is uniform, as should be the case for hypothesis probabilities [3], except for the peak at zero that corresponds to non-equivalent regions. This also validates the choice of κ and the use of the approximation for γ.

4.2 Clinical MR Data

Figure 4a shows the structure-rich slice from the IRTSE MR image volume chosen for analysis; the intensity change introduced by filtering with the scale-adjusted, statistically motivated version of NLM is shown in Fig. 4b. As in the simulated data, many locations are left unmodified on the basis of statistical uniqueness i.e. no matching patches can be found using the similarity measure; these tend to be located around the tissue boundaries as expected. This result is typical of clinical MR data. Magnified regions of Fig. 4a are shown in Fig. 5, illustrating the noise removal. This process can be considered a safe strategy for 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. However, the intensity changes introduced through application of the original NLM algorithm are shown in Fig. 4c. The less conservative exponential weighting embodied in this algorithm results in noise removal even in unique regions; few locations in the image are left unmodified, even without the incorporation of a free scale parameter. This implies that filtered estimates must necessarily include data from patches that cannot justifiably be expected to be from
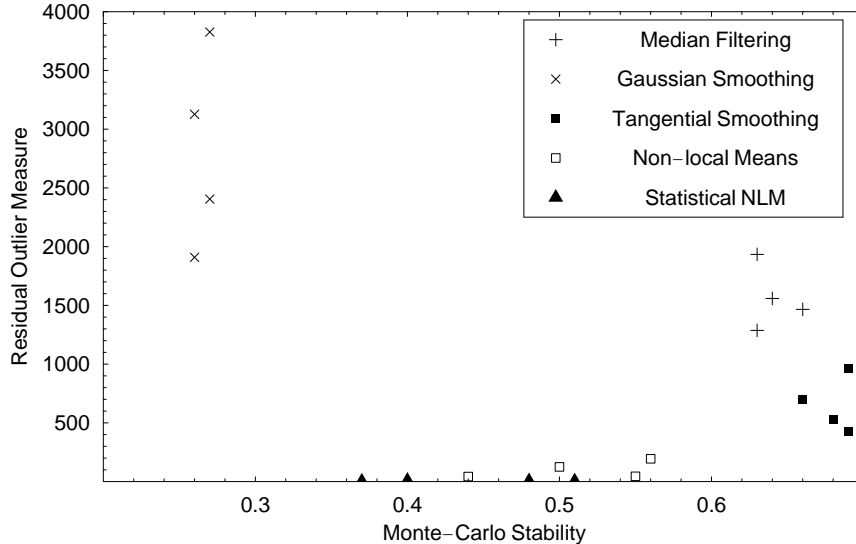
Figure 7: The proportion of noise removed (Monte-Carlo stability) and number of pixels destructively modified (ROM) during smoothing of the clinical MR images shown in Fig. 4a and Fig. 6. NLM was applied using a patch size of 7x7. The results of tangential smoothing over the three pixels along the normal to the direction of maximum local image gradient, Gaussian smoothing with σ=1 pixel and median filtering over 9-pixel patches are also shown.

	Median	Gaussian	Tangential	NLM	Stat. NLM
IRTSE	0.64 (1559)	0.27 (2405)	0.66 (698)	0.50 (125)	0.51 (4)
VE(PD)	0.66 (1466)	0.26 (3127)	0.68 (530)	0.55 (45)	0.37 (2)
VE(T2)	0.63 (1287)	0.26 (1909)	0.69 (426)	0.44 (43)	0.48 (5)
FLAIR	0.63 (1934)	0.27 (3827)	0.69 (966)	0.56 (194)	0.40 (9)

Table 1: Numerical results from Fig. 7, showing the Monte-Carlo stability and the ROM (in brackets) for each of the noise filtering techniques on each of the MR images.

the same distribution as the central pixel.

4.3 Quantitative Evaluation of Non-Local Means

Figure 7 shows the ROM and Monte-Carlo stability for noise filtering of the MR images shown in Fig. 4a and Fig. 6 with the original and scale-adjusted, statistically motivated NLM algorithms, Gaussian filtering, tangential smoothing and median filtering. The results are provided numerically in Table 1. The ROM should be equal to 0.27% of the pixels which, given the size of the images, is approximately 60 pixels. Tangential smoothing, which theoretically should be the least destructive of the simple noise filters, removes approximately 25% of the noise but applies inappropriate smoothing to approximately 2.7% of the pixels due to destabilisation of the gradient calculations by the image noise. Median filtering removes more noise but is approximately twice as destructive, whilst Gaussian smoothing removes approximately 75% of the noise at the expense of degrading the images significantly. The original NLM algorithm removes approximately 50% of the noise, and so falls mid-way between tangential filtering and Gaussian filtering. It has an average ROM of approximately 100 pixels, about 50% higher than the value expected for a perfect noise filter, again indicating that some pixels are modified inappropriately. The scale-adjusted, statistically motivated variant of NLM provides slightly better noise removal whilst achieving a ROM consistent with zero, although the differences between the two NLM variants are not statistically significant on either measure. This behaviour is expected for a strict test of similarity; pixels lying in the tails of the noise distribution represent uncommon intensities by definition; few pixels with similar intensities will be found by the similarity metric, and so these pixels will be left largely unmodified. An implementation of NLM with a strict test of similarity should therefore achieve a ROM of zero by definition.

5 Discussion and Conclusions

The statistically motivated variant of NLM differs from the original algorithm in two main ways. First, radial Gaussian weighting has been removed from the similarity metric such that it becomes a standard χ^2 test of patch similarity. There appears to be no conventional statistical analogue of the radial Gaussian weighting process. Second, the averaging process has been modified so that the intensities are weighted by a conditional probability of classification, through analogy with the Expectation-Maximisation algorithm, such that it becomes a maximum-likelihood estimator of the noise-free intensities. In general it is not possible to compute the required conditional probability without knowledge of the distributions both for matching and non-matching patches in the joint space of intensity and similarity. The χ^2 hypothesis probability used here is a less strict test of similarity by comparison.

In tests on both simulated and clinical data the modified algorithm finds instances of data, representing unique structures in the images, that have no statistically similar patches, resulting in extended areas of unmodified pixels. The presence of illumination variation in the simulated images or intensity inhomogeneity in MR images causes this behaviour to extend over the entire image. However, even the incorporation of a free scale parameter to estimate local patch intensity variation, whilst recovering many matches, leaves areas of unmodified pixels. In other tests it was found that reduction of the patch size from 7x7 to 5x5 in the MR brain images reduced the proportion of non-filtered pixels from 30% to 15%, although this does not necessarily represent an improvement in performance since it reduces the power of the technique to discriminate between matching and non-matching patches. By comparison, the original specification of NLM with exponential weight factors and spatial Gaussian weighting produces far fewer unmodified pixels. The analysis presented here therefore suggests that conventional NLM methods may be modifying data inappropriately, by combining data that cannot be considered equivalent at the level of the information present, potentially introducing an image structure dependent bias on the filtered values. Such behaviour must be carefully considered in clinical applications. All variants of NLM result in some unmodified pixels, implying that the noise on the filtered images is spatially varying and correlated with image structure; this would have to be taken into account in any subsequent analysis procedure. The demand for strict identicality between image patches appears to be the root cause of these problems, and so it may be possible to alleviate them through more sophisticated processing, such as constructing an eigenvector model of patch variation using principal component analysis.

The quantitative evaluation performed here suggests that NLM removes on average about half of the noise present; by comparison, tangential filtering (the simplest possible instantiation of ADF) removes about a quarter of the noise, whilst Gaussian filtering removes about three quarters. However, NLM achieves a significantly lower ROM than alternative techniques, indicating that the incorporation of an explicit test of similarity is effective in preventing destructive modification of the images e.g. blurring of edges. The original NLM algorithm does not achieve the theoretical optimum ROM, confirming that some pixels are modified inappropriately i.e. the exponential weighting is not sufficiently strict. The scale-adjusted, statistically motivated variant of NLM achieves slightly better noise removal, probably due to the presence of the scale adjustment; the inhomogeneity correction algorithm applied to the images does not remove all inhomogeneity, and the scale adjustment allows the algorithm to detect more matching patches in the presence of any residual inhomogeneity. In addition, it achieves an ROM of almost exactly zero in all cases, as would be expected with a strict test of similarity. It should also be noted that the original paper on NLM [5] suggested incorporating a test on the output data such that, if pixels were modified by more than the expected amount given the noise, the smoothed value should be averaged with the original, noisy intensity to provide some degree of correction; a method for performing this averaging was also derived. This introduces an explicit correction for pixels not well described by the implicit image model assumed by the filter, and for which the filter therefore gives erroneous results. Since the ROM was designed to directly quantify the number of such pixels, applying this correction prevents subsequent ROM-based evaluation. However, an alternative measure based on the Komolgorov-Smirnov distance between the intensity residuals and the expected noise distribution could be applied. The correction is very effective in improving the empirical performance of all noise filtering algorithms.

A simple analogy can be identified between NLM noise filtering and biological vision in humans. There is some evidence that the brain analyses microtextures in images by identifying common patterns, or templates, and representing image data as the conjunction of such patterns. The idea that matching processes occur in early vision has long been established, and supports tasks such as "pop-out", where unusual structure is immediately identified. If this is the case then the process of representation is in itself a noise-filtering system analogous to NLM; the main difference is that the templates might be pre-learned in human vision, but are obtained on-the-fly in NLM. This suggests that a possible improvement to the NLM filter as presented here could be the use of patch dictionaries for specific classes of images, allowing the removal of noise from unique locations in such images by simple patch substitution, taking scale differences into account. Otherwise, the computational processes required to support NLM are entirely biologically plausible i.e. consistent with those thought to be supported by neurons. Additional psychophysical support for this hypothesis is provided by the observation that humans can visually

identify image noise (e.g. Fig. 5). The noise-filtering algorithms that give the best subjective performance will therefore be those, perhaps like NLM, which replicate the human vision system and so provide the expected output. Such a filter could therefore be used in clinical applications without destroying information that a clinician would consider significant.

Appendix A: Estimating Normalisation

Application of the variational principle to the problem of matching two scaled noisy image patches I and J, results in the optimisation function

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

Putting $\alpha = sin(\theta)$ and $\beta = cos(\theta)$ the scaling that minimises the χ^2 function can be determined as follows

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

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 which can be considered as the sides of a right-angled triangle allowing $cos(2\theta)$ and $sin(2\theta)$ to be determined

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

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

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

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

Appendix B: Numerical Details

The main approach used here to avoid lengthy execution times was to expand the patch similarity measure as

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

computing the quantities

$$A = \sum_n^n I_n^2 \quad B = \sum_n J_n^2 \quad and \quad C = \sum_n I_n J_n$$

from differences in precomputed integral images[2] of I^2, J^2 and IJ, allowing the efficient calculation of $\gamma = \sqrt{B/A}$ for one hypothesised shift of the patch across the entire image, and all of the other factors required. This allows implementation of the algorithm within a 4- rather than 6-deep nested loop, reducing execution time from hours to minutes. Additional improvements in efficiency can be 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] D Barash. A fundamental relationship between bilateral filtering, adaptive smoothing and the nonlinear diffusion equation. *IEEE Trans. PAMI*, 24(6):844–847, 2002.

---

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

[2] R J Barlow. *Statistics - A Guide to the Use of Statistical Methods in the Physical Sciences*. John Wiley & Sons Ltd., U.K., 1989.

[3] P A Bromiley, N A Thacker, and P Courtney. Non-parametric image subtraction for MRI. In *Proc. MIUA 2001*, pages 105–108, Birmingham, 2001.

[4] P A Bromiley, N A Thacker, and P Courtney. Non-parametric image subtraction using grey level scattergrams. *Image and Vision Computing*, 20:609–617, 2002.

[5] A Buades, B Coll, and J M Morel. A review of image denoising algorithms, with a new one. *Multiscale Model. Simul.*, 4:490–530, 2005.

[6] D Comaniciu and P Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. PAMI*, 24(5):603–619, 2002.

[7] P Coupe, Pierre Yger, and C Barillot. Fast non local means denoising for 3D MR images. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2006*, volume 4191/2006 of *Lecture Notes in Computer Science*, pages 33–40. Springer Berlin/Heidelberg, 2006.

[8] A P Dempster, N M Laird, and D B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Society*, 39:1–38, 1977.

[9] H Gudjbartson and S Patz. The Rician distribution of noisy MRI data. *Magnetic Resonance in Medicine*, 34(6):910–914, 1995.

[10] M Mahmoudi and G Sapiro. Fast image and video denoising via nonlocal means of similar neighbourhoods. *IEEE Signal Processing Letters*, 12(12):839–842, 2005.

[11] J V Manjon, J Carbonell-Caballero, J J Lull, G Gracian-Marti, L Marti-Bonmati, and M Robles. MRI denoising using non-local means. *Medical Image Analysis*, 12:514–523, 2008.

[12] J V Manjon, M Robles, and N A Thacker. Multispectral MRI de-noising using non-local means. In *Proc. MIUA 2007*, pages 41–46, Aberystwyth, 2007.

[13] R. D. Nowak. Rician noise removal for magnetic resonance imaging. *IEEE Transactions on Image Processing*, 8(10):1408–1419, 1999.

[14] S I Olsen. Estimation of noise in images: an evaluation. *CVGIP: Graphical Models and Image Processing*, 55:319–323, 1993.

[15] P Perona and J Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. PAMI*, 12(7):629–639, 1990.

[16] W H Press, B P Flannery, S A Teukolsky, and W T Vetterling. *Numerical Recipes in C*. Cambridge University Press, New York, 2nd edition, 1992.

[17] N A. Thacker, A J Lacey, and P A Bromiley. Validating MRI field homogeneity correction using image information measures. In *Proc. BMVC'02*, pages 626–635, 2002.

[18] N A Thacker, M Pokrić, and D C Williamson. Noise filtering and testing illustrated using a multi-dimensional partial volume model of MR data. In *Proc. BMVC*, pages 909–919, Kingston, London, 2004.

[19] C Tomasi and R Manduchi. Bilateral filtering of gray and color images. In *Proceedings of the Sixth IEEE International Conference on Computer Vision*, pages 839–846, Bombay, India, January 4-7, 1998.

[20] E Vul, C Harris, P Winkielman, and H Pashler. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, 4(3):274–290, 2009.

Response to IET Computer Vision Reviewers' Comments

We include, in the following sections, the Reviewers' comments received when this paper was submitted to IET Computer Vision (in italics) and our responses (in normal text), as we believe that they are illuminating both in regard to the main text and the reviewing process itself (the latter with particular reference to the comments we received from Reviewer 3). Comments relating to minor syntax errors have been omitted.

Comments to Editor

Thank you for your e-mail providing the Reviewers' comments on our paper "A Statistical Interpretation of Non-Local Means". We have implemented the majority of the changes suggested by Reviewers 1 and 2 directly, and provide a detailed list of these modifications in an accompanying document entitled "A Statistical Interpretation of Non-Local Means: Response to IET Computer Vision Reviewers' Comments".

We would like to raise some concerns over the comments made by Reviewer 3. Their review consists of just four sentences:

- *Textual elaborateness with insufficient citations suggestive of author bias in presentation. Justify.*

- *Manuscript presentation could have created better impact with evidence than descriptive summaries.*

- *Is it possible to test the proposed method using any other strict statistical tool? Justify.*

- *Suggested that an acceptable sample population is chosen and used for testing the proposed method. Visual, Graphical and Tabular evidence be given to support exhaustive testing.*

None of the points made are supported by examples, or provide practical suggestions for alterations to the paper. For example, with reference to the first point, at which locations in the paper is the text over-elaborate? Which citations should be added to the paper? Which statements in the paper are biased? The fourth point seems to suggest that the Reviewer would like to see the entire evaluation replaced with a different set of experiments, although again they do not define what is meant by an "acceptable sample population", or state what is wrong with the existing evaluation.

In the face of such a lack of specific comments, we have found it hard to respond to this review. We have done so mostly by listing changes made in response to specific comments from the other Reviewers, where they seem to be appropriate. We therefore hope that you will take the poor quality of this review into consideration when deciding whether to accept the paper.

Reviewer 1

- *I felt the organisation of the paper would be improved if the statistical NLM interpretation (sect 2.1) was in a dedicated section (ie sect 2.), allowing sect 2.2-2.4 to be in a sect 3 entitled Evaluation Methodology or similar. Fig 2 could also be moved back, and possibly combined with Fig 5 (fig 2a and 5a appear to be the same?).*

  We agree with all of these suggestions, and have implemented them directly. One additional benefit of this reorganisation is that the subsections in the new Evaluation Methodology section correspond exactly to those in the Results section. Therefore, we have also modified the subsection headings so that they are the same across these two sections.

  Fig. 2a and Fig. 5a are the same, but we believe that Figs. 2 and 5 should be separate, as Fig. 2 shows the clinical MR data used in testing and Fig 5 shows results (the input data is included in Fig. 5 in order to allow a direct visual comparison with the results). Therefore, we have moved Fig. 2 back as suggested, so that it is contained within the new evaluation methodology section, and removed Fig. 2a so that Fig. 2 now shows the additional MR images used together with the IRTSE image (and mentioned this in the Figure caption).

- *The abstract is concerned with noise filtering in general whereas line 1 of the introduction is specifically focused on MR imaging. I think that it would be helpful for the reader if the introduction could briefly describe the noise types for which the NLM algorithm performs well (the last para on pg 11 could be moved here also). For example, additive Gaussian noise is used in the simulations and therefore it is appropriate for images*

characterised by Gaussian noise (of which MRI is a good example) rather than impulsive or multiplicative noise?

We have moved the sentence describing previous empirical evaluations of NLM from the last paragraph of page 11 to the start of the 4th paragraph of the introduction, and added the requested material at this point i.e. Buades concluded that the more similar the noise is to genuine white noise, the better NLM works (this is due to the restrictions of the similarity test i.e. the more structure the noise contains, the more it begins to look like genuine image structure, and so the more NLM tends to leave it unaltered). This also helps to put our focus on medical imaging into context, and so we have expanded this point a little. Finally, we have altered one sentence in the abstract from "quantitative image analysis tasks" to "quantitative or clinical image analysis tasks" to make it clear that our main focus is on clinical imaging.

- *Although a comparison with the original NLM algorithm is performed for the MR data, it is not included for the simulated data. I would like to see this comparison included in sect 3.1. As the noise-free intensities are available here, they can be used in a quantitative comparison and a simple measure such as SNR used to quantify the improvement gained by using the statistically motivated NLM.*

We take the point that a quantitative evaluation could be performed using the simulated data, and that knowledge of the gold standard result (i.e. the noise-free simulated data) would allow some simple metrics to be applied. However, we strongly feel that this material should not be added to the paper, for two reasons.

First, the aim of this section of the paper (i.e. Sections 3.1 and 4.1 in the revised manuscript) is to illustrate two specific points. The simulated data was constructed to have a grid of unique locations (i.e. locations for which the surrounding patch of intensities is unlike any other patch of intensities in the image). This grid can be seen in the result (Figs. 2b and 3c). Therefore, the first point illustrated using this data is that NLM tends to leave such unique locations unmodified, as there are no similar patches of intensities in the rest of the image that can be used in the averaging process. This is in turn used to explain the results shown in Fig. 4. The second point illustrated using this data is that, when a smooth intensity variation is added (as might be generated by illumination effects in natural images or an inhomogeneity artefact in clinical MR images), all locations in the image become unique. This motivates the addition of a free scale parameter to the statistically motivated variant of NLM; without this illustration it is not obvious that a free scale parameter is required. We feel that attempting to turn this section of the paper into a quantitative evaluation section would deemphasise these points, and the intended message would be lost.

Second, we feel that performing the proposed evaluation would not add anything to the existing evaluation. Simulated data can always be criticised on the basis that it does not provide a realistic test i.e. does not contain the full range of features, artefacts etc. that would be expected in natural or clinical images. This is particularly true in the case of the simulated data set used in the paper which, being designed to provide a regular grid of unique locations in order to illustrate the points described above, is highly unrealistic i.e. very different from the clinical data upon which the algorithm was actually tested. Our stated aim was to investigate the statistical basis of NLM, in order to allow its use in quantitative or clinical tasks, where its behaviour must be completely understood. Therefore, the quantitative evaluation section of the paper used clinical images, so that the data could not be criticised as being unrealistic. Therefore, any quantitative testing on simulated data is redundant: it would simply provide a less-convincing evaluation of algorithmic performance than that already present in the paper.

Reviewer 2

- *Lack of references*

  - *you refer to anisotropic diffusion (AD) algorithm but the original Perona's and Malik's paper is not cited (Perona, P.; Malik, J. "Scale-space and edge detection using anisotropic diffusion", IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 12, Issue 7, Jul 1990 Page(s):629 - 639)*

  - *I would also mention the mean shift (MS) algorithm (it can also be used for filtering, see D. Comaniciu, P. Meer: Mean Shift: A Robust Approach toward Feature Space Analysis, IEEE Trans. Pattern Analysis Machine Intell., Vol. 24, No. 5, 603-619, 2002)*

  - *An interesting relationship between AD, MS, and BF is derived in: D. Barash "A Fundamental Relationship between Bilateral Filtering, Adaptive Smoothing and the Nonlinear Diffusion Equation" , IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 24, Issue 6, Page(s):844 - 847, Jun 2002.*

- – *When you mention "... Gaussian filters fall into this category and have been widely used in some applications such as functional MRI." Please, mention a reference related to.*

  We have added all of the suggested references to the paper.

- *At the end of Introduction Section you have to explain the organisation of the paper.*

  A paragraph explaining the layout of the paper has been added at this point.

- *Some terms of relationships or formulas are not explained (perhaps I have not seen them).*

  - – *Formula (2): what is gb?*

    gb is the intensity of the central pixel in patch B: this is defined in the line of text above Eq. 2. We have altered the text at this point to clarify the definition of gb.

  - – *what is $erf()$?*

    $erf()$ is the error function i.e. $erf(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp^{-t^2} dt$. It is closely related to the incomplete gamma functions and also to the Gaussian distribution i.e. $erf(\frac{x}{\sqrt{2}\pi})$ is the integral of a Gaussian distribution with mean 0 and standard deviation σ from $-x$ to x. The point is that these indefinite integrals cannot be performed analytically, but most numerical software libraries (e.g. math.h if you are using C) will provide a function that gives a numerical approximation. We have added a reference to avoid this confusion in future.

- *Section 2.3 Why did you use a region of 7x7 and not a region of 5x5 or other size? Please, explain it.*

  Manjon et al. evaluated the effect of this parameter in a previous paper (MRI denoising using Non-Local Means, Medical Image Analysis 12 p. 514, 2008; cited in the paper), concluding that a 7x7 similarity patch provided an optimal balance between accuracy and required processor power. We have added a statement to this effect to the paper.

- *Section 2.4 Why have you compared the filter against Gaussian filtering and median filtering and you have not used AD, MS or Wavelets?*

  As stated in the Introduction, the aim of this paper was not to perform a large-scale comparison of many different noise filtering algorithms (in any case, this has already been done by Manjon et al. and Buades et al.: see the references in the paper), but to try to understand the NLM algorithm in a statistical sense. Therefore, all evaluation stages were focused towards this aim. In the comparison to other noise filtering algorithms, the underlying aim was to quantify how much noise was removed by NLM and how destructive it was to image content, and to compare this between the original and statistically motivated versions of the algorithm. The other algorithms included in this section were present only to provide some context for the results. Therefore, they needed to be simple algorithms, familiar to the readers of the paper (i.e. widely used), whose properties would be known in advance (e.g. the fact that median filtering is not very destructive to image content but also does not remove much noise, and that Gaussian filtering provides a high degree of noise removal but tends to blur the edges in the image). In this context, we feel that algorithms like mean shift and wavelet smoothing are too complex to provide an easily understandable context.

  We also note that, again with a view to providing an easily understandable context for the NLM results, we tried to keep the settings of free parameters as low as possible, and did not evaluate the effects of changing them (e.g. using multiple different Gaussian filters with different kernel sizes). In this context, anisotropic diffusion is present in the comparison in the form of the tangential smoothing algorithm, which is the simplest possible implementation of anisotropic diffusion, as mentioned in Section 3.3.

  In retrospect, entitling the relevant sub-sections "Comparison to Other Noise Filtering Algorithms" was a mistake in the light of the above comments, and was bound to lead to this confusion. Therefore, was have changed the title of these sub-sections (both in the Evaluation Methodology and Results sections) to "Quantitative Evaluation of NLM".

- *Results. A table indicating rates will show the advantages of the filter and it would also improve the quality of the paper.*

  We have added a table of numerical results as suggested.

15

Reviewer 3

- *Textual elaborateness with insufficient citations suggestive of author bias in presentation. Justify.*

  The Reviewer does not specify the points at which the text is over-elaborate, specify which papers should be cited, specify at which points the paper is biased, or make any suggestions for alterations to remedy any of these issues. In the absence of this information, it is difficult to interpret the Reviewer's comment or decide which changes should be made to the paper. However, we have added more references to the paper in response to a comment from Reviewer 2 (see above), and have also clarified the text at several points in response to specific comments made by the other Reviewers (again, see above).

- *Manuscript presentation could have created better impact with evidence than descriptive summaries.*

  The Reviewer does not specify what additional evidence is required to support the conclusions we draw in the paper. We contend that the paper incorporates sufficient quantitative evidence to support our main conclusion i.e. that the original NLM algorithm incorporates a similarity test that, in comparison to a strictly statistical test of similarity when applied to clinical MR images, is not sufficiently strict to avoid any destructive image modification. This conclusion is clearly stated in the second paragraph of the Conclusions section (see the response to the final comment, below).

- *Is it possible to test the proposed method using any other strict statistical tool? Justify.*

  We note that the Reviewer does not suggest alternative metrics, or describe why they would be superior to the ones used in the paper. There are only two important measures of the performance of any noise filtering algorithm: how much of the noise does it remove, and how much does it degrade image structure. We have directly evaluated these two quantities in the quantitative evaluation section of the paper, and contend that they do constitute strictly statistical metrics. Numerous different measures could no doubt be devised: for example, the noise removal efficiency could be quoted in terms of SNR rather than as a percentage of noise. However, such measures would contain the same basic information as those used in the paper.

- *Suggested that an acceptable sample population is chosen and used for testing the proposed method. Visual, Graphical and Tabular evidence be given to support exhaustive testing.*

  The reviewer does not specify what would constitute an "acceptable sample population", or specify how the evaluation present in the paper is insufficient. They seem to be suggesting that the statistically motivated variant of NLM should be tested simply by applying it to a large image library. However, it is difficult to see what this would add to the paper.

  Our stated aim was not to perform a wide-ranging algorithmic "shoot-out", but to determine the statistical foundations of NLM, in order to determine if it can be used in clinical tasks where any destructive modification of the images would be unacceptable, as it might remove information that a clinician would consider important. The approach taken was to identify an analogous method based on standard statistics, and then to compare the performance of the original and statistically motivated variants of NLM. We achieved this in two stages. First, tests on simulated data qualitatively illustrate the behaviour of the algorithm i.e. the tendency to leave unique image regions unmodified, and the subsequent need for a free scale parameter. Second, tests on clinical data were used in a quantitative evaluation, and showed that the original NLM variant, whilst less destructive than simple noise filters like Gaussian smoothing (used in some fMRI applications, as stated in the Introduction) still modifies some pixels that are left unmodified by the statistically motivated variant. We contend that this provides sufficient evidence to support our main conclusion, that the original NLM algorithm incorporates a similarity test that, in comparison to a strictly statistical test of similarity when applied to clinical MR images, is not sufficiently strict to avoid any destructive image modification. This conclusion is clearly stated in the second paragraph of the Conclusions section.