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, University of Manchester, U.K.
[2] IBIME Group, ITACA Institute, Universidad Politécnica de Valencia, Valencia, Spain.

Keywords: Quantitative, statistical, noise filtering.

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. With the increasing use of imaging in medicine and sciences it might be considered inevitable that researchers will try to apply such noise filtering schemes in quantitative analysis. In order to do this with confidence we need to develop an understanding of the noise removal process that goes beyond subjective appearance. The purpose of this paper is to develop and test a statistical basis of NLM, in order to try to understand the conditions required for its use. The theory is illustrated on synthetic data and real MR images of the brain.

1 Introduction

MRI de-noising is a common pre-processing step in many MR image processing and analysis tasks, such as segmentation or registration. Many filtering methods are based on the signal averaging principle, which uses the spatial redundancy in the image. Gaussian filters have been widely used in some applications such as fMRI but they have the disadvantage of blurring edges due to averaging of non-similar patterns. In order to avoid this problem many edge preserving filters have been proposed. Probably the best known is the Anisotropic Diffusion Filter (ADF) [1, 2]. Such filters respect edges by averaging pixels in the orthogonal direction of the local gradient. However, they 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 artifacts that can be problematic for clinicians.

The NLM algorithm modifies the intensity g_a of each location of the image (x, y) by comparing many locations within the image and selecting those regions that appear similar on the basis of local context (ie: the patch of surrounding values for the ath patch A and bth patch B). The similarity is defined using a measure such as

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

where $G_\rho(r)$ is a radial Gaussian weighting from the centre of the patch and ρ is the scale parameter. The noise filtered intensity $\hat{g}_a$ is then estimated as a weighted sum of central pixel values g_b for matching locations, typically

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

where h is a scale factor that determines the required degree of similarity. Recently, the NLM algorithm has been shown to be very effective in empirical tests for clinical data [5]. However, it appears that the theoretical basis for the algorithm has not been fully understood in statistical terms. As a consequence, the approach has many free parameters (in particular h and ρ) that influence the behaviour of output data. The application of the approach therefore has to be carefully analysed in order to be considered adequate for clinical applications. For this we need a fundamentally deeper understanding of what the algorithm is doing. As with all estimation tasks, this requires us to identify the assumptions necessary to derive the algorithm from the theory of statistical estimation. In particular we need to understand the patch matching and grey-level weighting processes in terms of conventional statistics.

2 Methods

The NLM algorithm presented in the introduction does not map directly onto statistical estimation processes, and so we simplify it 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.

In the first instance we will modify equation (1) so that it conforms to a standard quantitative χ^2 test

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

which is valid for independent, identically distributed (IID) data. In addition we will investigate a similarity function that allows one degree of freedom for grey level scale (γ) and

minimises

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

The required scaling parameter $\gamma = \alpha/\beta$ is given by (Appendix A)

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

Equation (2) is less straightforward to interpret. This is however clearly an estimation task, and we would expect the weighting factors to be related to the probability that the central grey-level in patch b is drawn from the same distribution as the grey-level from patch a, using the similarity function as the basis for this assessment. In general we can expect this to require us to have knowledge regarding the behaviour of the data in the joint 2D space of grey level residuals and similarity. This interpretation is summarised in Figure 1, which illustrates 2D distributions of the difference between central patch grey-levels plotted against the patch similarity measure ($\sqrt{\chi^2/N^2} = \chi/N$) for corresponding patches ($b = a$) and non-corresponding patches ($b! = a$). Our 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 a variance $1/N$.

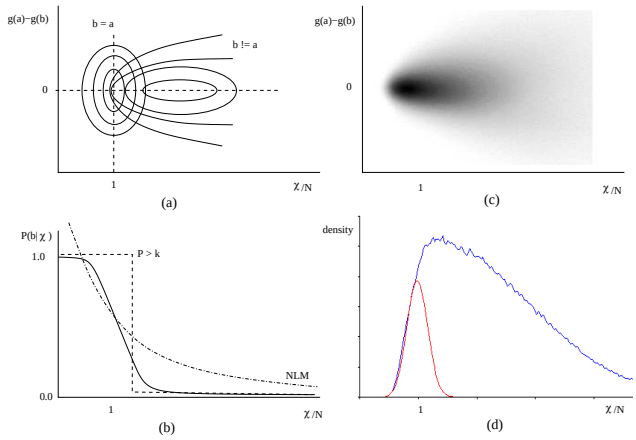


Figure 1: *(a) The 2D distribution of difference between central grey-level values as a function of the patch similarity measure, (b) the marginal projection of the conditional probability of the data being drawn from the matching distribution (b=a), (c) data from the brain image (Figure 4(a)), (d) profile of distribution in (c) (blue) with expected signal distribution (red).*

The process of weighted estimation embodied in equation (2) is directly analogous to 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 Expectation Maximisation requires that a weighted mean is constructed using the conditional probability of classification. Here, this corresponds to the grey-level being drawn from the signal distribution $P(b = a|\chi_{ab}, g_a - g_b)$. The constraint this places on the estimated grey-level must also take account of how well g_b predicts g_a (ie σ_{ab}). Clearly, replacing this 2D distribution with a 1D weight factor, dependent only upon the similarity function as in NLM, (Figure 1(b)) such as

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

rather than

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

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 (as used in equation (6)) 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. When using equation (5), or similar, we can therefore justify putting a cut on the 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 (Figure 1(b)). 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 Figure 1 (c)) 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 equation (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 Figure 1 (d)). 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 ($P_{ab} = P(\chi^2_{ab}, N^2)$) 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 (ie: 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()). This 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. Numerical implementation of the similarity calculation requires some care in order to avoid a six deep nested loop (Appendix B).

2.1 Results

2.1.1 Simulated data

The similarity measure used in the first experiment is given by equation (3), where data is sampled in a region around the central pixel g_b and σ is an estimate of the level of image noise. This sets the scale for our definition of similarity according to the level of evidence in the image.

Here, filtered pixels are estimated using

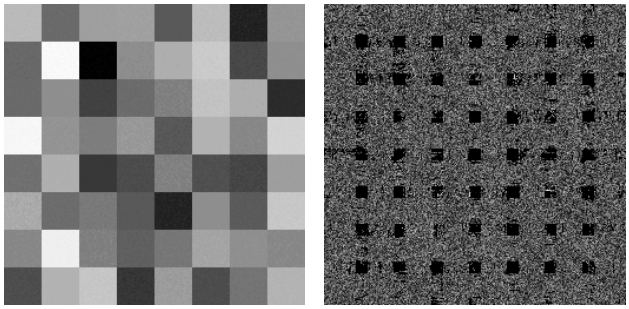
$$\hat{g}_a = \frac{\sum_b g_b P_{ab}/\sigma_{ab}^2}{\sum_b P_{ab}/\sigma_{ab}^2} = \frac{\sum_b g_b P_{ab}/\chi_{ab}^2}{\sum_b P_{ab}/\chi_{ab}^2} \qquad (7)$$

where σ_{ab} is the standard deviation (width) 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.

In order to test this process, simulated data was generated in 32/32 square regions with grey levels drawn randomly from a Gaussian distribution of width 32 grey levels. Independent uniform Gaussian random noise was then added with a variance of 1 grey level (Figure 2(a)). Matching image patches were sought in a 23x23 region about the central pixel g_a. The relative change in grey level produced by filtering using an 11x11 patch similarity calculation is shown in Figure 2(b). Notice that unique regions of the image (corners of each square patch) are generally left unmodified (0 difference) as expected.

In a second simulation, a uniform slope in grey level (1/4 grey-levels per pixel) was applied to the image prior to the addition of noise (Figure 3(a)). For this data more pixel locations are considered statistically unique and left unmodified by the averaging process (Figure 3(b)). This problem is consistent with a spatially varying illumination, as present in many images and field inhomogeneity in MR data. This lack of matching patch data therefore presents a problem with this simple approach to determining correspondence.

[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)

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

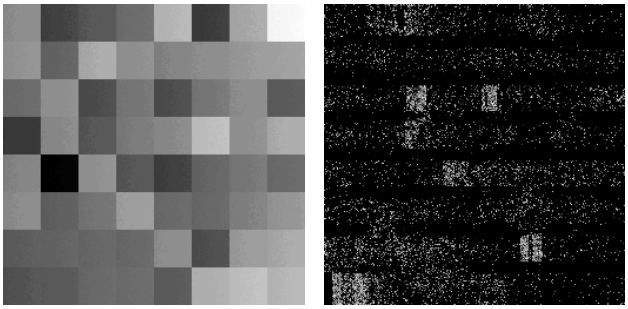
To alleviate this problem and recover more samples, the image regions were instead matched using a free scale parameter, as discussed in the previous section.

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

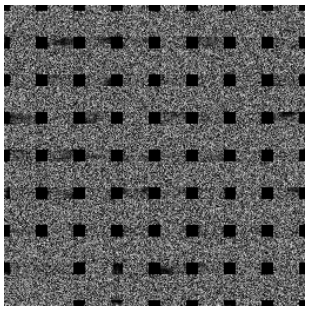
with γ calculated using $|B|/|A|$ for purposes of computational efficiency. This was not found to result in an observable change in the sampled distributions (see below). Our variable is expected to be a χ^2 statistic with $N-1$ degrees of freedom. The variable $\sqrt{\chi^2/(N-1)}$ is drawn from a distribution that is approximately a Gaussian with mean 1.0 and a variance $\kappa/(N-1)$. 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). The hypothesis probability was again be computed using the error function, and weighted averaging was computed using

$$\hat{g}_a = \frac{\sum_b \gamma g_b P_{ab}/(\chi_{ab}^2(1+\gamma^2))}{\sum_b P_{ab}/(\chi_{ab}^2(1+\gamma^2))} \qquad (8)$$

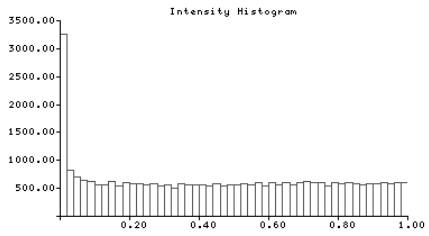
This takes appropriate account of the accuracy of each estimate (γg_b), such that each data point is weighted by the expected difference between the sample and the underlying true value. The difference between filtered values and the original data (Figure 3(a)) is shown in Figure 3(c). Following the scaling modification the filtering results are comparable to the original result (Figure 2(b)), 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 (Figure 3 (d)). The distribution is uniform, as should be the case for hypothesis probabilities [**?**], except for the peak at zero that corresponds to non-equivalent regions. This also validates our choice of κ and the use of the approximation for γ.



(a) (b)



(c)



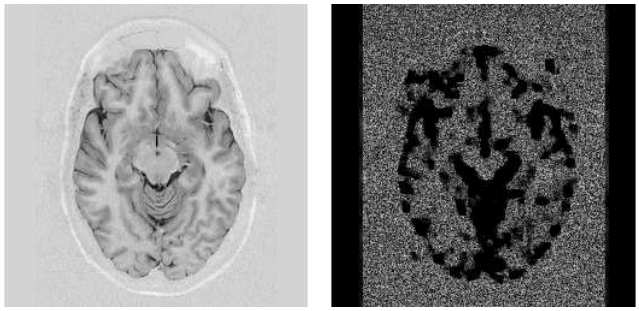
(d)

Figure 3: *Data with illumination variation changing left to right (a) and resulting change due to filtering (b). The distribution of hypothesis probabilities when data is filtered with a scaled matching process (c) using equation (8) is also shown (d). This histogram is uniform for data drawn from the assumed distribution, as expected, with a peak at zero corresponding to unique locations.*

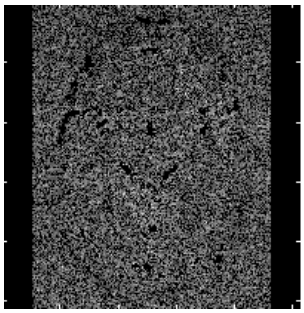
2.1.2 Real Data

The above (scale adjusted) filtering scheme was also applied to a 256x256 MR image of the brain (Figure 4(a)) from a 1.5 Tesla machine. 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 Figure 4(b), and illustrates that as in the simulated data many locations are left unmodified on the basis of statistical uniqueness. This result is typical of clinical data.

A magnified portion of this image and the filtered output is shown in Figure 5(a) and (b). 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) (b)



(c)

Figure 4: *Magnetic resonance data for a normal subject taken at the level of the top of the brainstem (a), and corresponding change in each pixel (b). Figure (c) shows the differences produced with a conventional NLM filter for comparison.*

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 (Figure 4(c)) 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.

3 Discussion and Conclusions

There are several notable differences between the statistically based filtering scheme described above and the original method of NLM. Firstly, we have not used a spatial Gaussian weighting function in the construction of the regional similarity measure. This implies a change in the statistical variables and in order for this formulation to remain within the realm of conventional statistics, (and the principles of quantitative probability as required for scientific validity) the assumption of independent uniform Gaussian noise must be applicable to these new variables. This represents an assumption regarding the differences likely to be obtained between samples of matching patches rather than the intrinsic measurement process. This can only be considered as a justifiable strategy if an effort is made to confirm these distributions empirically and we must expect them to vary between data sets. There is

(a)

(b)

Figure 5: *Magnified region illustrating noise filtering (a) and (b). Despite many areas being left unmodified there is little subjective evidence of under-filtering.*

also a significant difference in the way that weighted estimates are computed (equation [5]). We have explained how an optimal combination of grey level estimates (such as one that minimised the expected error on the filtered value), would be expected to involve a monotonic function of the probability that data is drawn from a signal distribution of equivalent pixels. This would need to be based on some knowledge of possible contamination (distribution overlap) from non-identical patches. If there were no contamination then a simple threshold (such as taking the average for all values above a probability of 99%) would be appropriate. The selection scheme used here is therefore cautious by comparison. Weighting with a value of P_{ab}, allows us to at least eliminate values that cannot be from a statistically equivalent region. When we investigate the behaviour of this process in simulated and real data we find genuine instances that have no matches that can justifiably be considered statistically similar, resulting in extended regions of unmodified pixels. Further, a fully quantitative patch comparison causes immediate problems for the realistic image formation scenario of illumination variation, with many locations unable to identify matching patches. Therefore the basic simple grey level comparison has been modified to include estimation of local patch intensity variation (equation (4)). This has been found to be sufficient to recover many matches, without invalidating the quantitative interpretation of the method.

Even with scale estimation, real data exhibits unmodified regions of pixels, consistent with unique grey-level structures. We can reduce these areas by reducing the region defined for similarity, in other tests we found shrinking the similarity region from 7x7 to 5x5 (which we consider to be a reasonable minimum) reduces the proportion of non filtered pixels within the brain from 30% to 15%. However, this is not necessarily a genuine improvement in the behaviour of the filter as reduction is achieved by degrading the certainty with which we can associate corresponding image structure. The power of the technique to combine data from similar regions would therefore also be diminished. More conventional approaches, using exponential weight factors (i.e. $exp(-\chi^2)$) and spatial Gaussian weighting, do not have regions of unmodified pixels. Our analysis 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. This casts doubt on their optimality in clinical or quantitative applications. A quantitative statistical method leaves these areas alone, but consequently performs no noise filtering. This introduces a further problem, which is that the noise characteristics of the output image are spatially varying.

We believe that a strict demand for identicality between image patches is the root cause the inability to filter some regions. The issue could be alleviated if similar data were processed in a more sophisticated manner, such as constructing an eigenvector model of patch variation using PCA, as done in appearance modelling. This is an avenue that we intend to explore further in future. Software and further documents relating to this work can be found at our web site [6].

There is a simple analogy that can be identified here between the process of NLM noise filtering and the biology of human vision. There seems to be some evidence that micro textures in images are analysed in the brain by looking for common patterns (or templates), and representing the data in images as the conjunction of such patterns. The idea that matching processes occur in early vision has been long established, and this explanation supports tasks such as ‘pop-out’, where unusual image structure is immediately identified. If this is the case, then we can say that the process of representation is, in itself, a noise filtering system analogous to NLM. The main difference here being that the templates might be pre-learned in human vision, and have to be worked out on the fly for NLM. Otherwise, the computational processes required to support

NLM are entirely biologically plausible, ie: computable using local calculations thought to be supported by neurons.

Additional psychophysical support for this hypothesis comes from the observation that humans have no difficulty identifying the noise in noisy images, to the extent that we believe we can tell, just by looking, if the output from a noise filtering process has worked (e.g. Figure 5). The noise filtering schemes that give the best subjective performance will be those, perhaps like NLM, which work in a similar way to the vision system and so produce what we expect to see. We might therefore expect that in clinical applications, such a filter will not destroy information that would have been considered significant to the radiologist.

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 following 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 $\beta = \cos(\theta)$ we can now determine the scaling that minimises the χ^2 function 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 us to determine $cos(2\theta)$ and $sin(2\theta)$.

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

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

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

$$sin(\theta) = sin(2\theta)/2cos(\theta)$$

$$cos(\theta) = \sqrt{(cos(2\theta) + 1)/2}$$

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

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

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

in the form

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

computing the quantities

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

from differences in precomputed integral images[2] of I^2, J^2 and IJ, allows the efficient calculation of $\gamma = \sqrt{B/A}$, for one hypothesised shift of the match patch across the entire image, and all of the other factors required. This allows implementation of the algorithm to be performed within a 4 deep rather than 6 deep nested loop, delivering execution times of minutes rather than hours. 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] 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.

[7] www.tina-vision.net

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