Tina Memo No. 2016-002
Internal Report, IMI work extended version of 2009-004

Histogram-based Image Quality Assessment

N.A. Thacker and H. Ragheb.

Last updated
27 / 06 / 2015

WWW.TINA-VISION.NET

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

Histogram-based Image Quality Assessment

N.A. Thacker and H. Ragheb
Imaging Science and Biomedical Engineering Division
Medical School, University of Manchester
Manchester, M13 9PT, UK
neil.thacker@manchester.ac.uk

Abstract

Although image processing is often taught as an “image-in image-out” process with no thought as to data characteristics, medical image analysis algorithms should be based on statistical principles. This requires us to make assumptions concerning the image formation process, the structure of the intensity histogram, or other statistical properties of the input data. Application of such algorithms to image data that do not fit these assumptions will produce unreliable results. Unfortunately these problems are often subtle and are not easily identified using visual inspection. Reliable use of automated analysis therefore requires additional tests in order to confirm the suitability of data. Basic tests include evaluation of signal to noise ratios and conformity to expected ranges, but the distribution of signal is also an important factor. Data may be inappropriately scaled, quantised or distributed.

This report describes a technique for the automatic identification of images that do not have histogram structure consistent with that expected. The approach is based upon a component analysis followed by statistical testing. Experiments validate its use in the identification of quantisation problems and unexpected image structure. It is intended that this test will form one component of a quality control assessment, to aid in the use of sophisticated statistical image analysis software by non-expert users.

1 Introduction

Complex image processing techniques, such as segmentation, registration and parametric image generation, have been shown to have utility in clinical applications. However, these techniques are often based on specific assumptions about the image formation process, the structure of the intensity histogram, or other statistical properties of the images. Considerable insight on the part of the end users may be required in order to avoid inappropriate application of such techniques to input data that do not fit these assumptions, and thus to avoid invalid results. Further, it may not be practical to provide adequate levels of training to end-users to enable them to assess the numerical or statistical stability of an algorithmic process on specific data. This introduces a requirement for automatic data quality assessment prior to the main analysis (such as signal-to-noise checks [3]).

For CT and MR images, the DICOM header file may be used to check acquisition parameters such as temporal resolution, spatial resolution, weighting factors, and the presence or absence of contrast enhancements. However, many subtle statistical effects are not immediately apparent to a human observer, and such simple checks may not suffice to identify all possible image quality issues. In addition, the goal of automatic quality assessment software should be to provide end users with useful feedback and possible solutions when an input dataset fails a quality check.

Many problems with data analysis occur when images are not consistent with those on which an an algorithm was developed. For many observers, it might be difficult to ascertain any difference. In our experience, simple visual inspection is unable to identify all issues even for well trained staff. We are particularly bad at identifying; spatial distortions, image quantisation, image non-uniformity and more general quantitative changes. All of which can generate significant challenges to data analysis. However, a lot of the relevant information relates to the specific data values and is visible in distributions, i.e. a histogram. Here, we use a histogram-based model of the data to ensure the valid use of statistical approaches. The approach can be thought of as similar to using a spell checker to check a document, where a library a previously correctly spelled words are compared to new text, but here we compare a model of histogram distributions, paying particular attention to allowable variation. The database of starting distributions would be intialised with data on which the analysis methods were originally tested, and subsequently extended following successful analysis of new data not found to match these examples. We train the algorithm using a variety of compatible images. To obtain a robust model, we divide each image into several non-overlapping windows and use all corresponding histograms in the process of training.

Our approach is based on fitting a combination of appropriate density functions to data throughout the corresponding bin intervals of each histogram. The density model is built on simple assumptions regarding expected distibutions in MR and CT data, which are commonly used as the basis for general purpose image segmentation. It includes components for both pure tissue and partial volume voxels. We make use of weighting parameters to model a set of histograms based on the proportions of the density functions. These weighting parameters are updated using Bayes theory to estimate the components for an independent components analysis (ICA). This generates a compact non-parametric model. A cost function is then computed based on the log-likelihood or the Matusita measure. Finally, the cost function is optimised using the EM algorithm to obtain the density parameters. These parameter estimates, together with the weighting parameter estimates, should provide a sufficient description of the image histogram data. Statistical tests are then sensitive to the same quantitative distribution variations which can impede data analysis but might be visually undetectable.

Below we should how quantitative statistical tests can be used to identify similar and dis-similar image content, and to identify the more subtle problem of grey-level quantisation. The data used are a random selection of MR images, with distribution modifications made where necessary.

2 Algorithm

The idea is to train the algorithm using a number of acceptable images using a histogram-based model. This model is later used to assess the quality of new images by computing a measure for the goodness of fit.

2.1 Training Phase

The input image used for training is divided into J non-overlapping windows of equal size. This gives J different data histograms f_j $(j = 1, 2, ..., J)$ to which a unique histogram model is fitted. The model consists of I components $(i = 1, 2, ..., I)$ where each component is a density function $p(g|v_i)$ defined based on knowledge of the corresponding tissues. While the tissue parameters are identical for all histograms and are learnt through the optimisation of a global cost function, each histogram has specific weighting parameters α_{ij} which are updated using the Bayes theory.

The data density model is based on that originally proposed by Santago and Gage [4], and includes components for both pure tissue and partial volume components. It is based on three assumptions: that, in the absence of artefacts and noise, each pure tissue has a well-defined signal intensity; that in partial volume voxels the constituent tissues contribute proportionately to the intensity of the voxel (i.e. that the image formation process is linear); and that there is no correlation between voxel boundaries and tissue boundaries. Pure tissues therefore generate a delta function in the intensity histogram, convolved with a noise distribution that is assumed to be Gaussian, giving

$$G(g; \sigma, \mu) = \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(g-\mu)^2}{2\sigma^2} \tag{1}$$

where σ is the standard deviation of the image noise and μ is the mean intensity of the tissue. In the original Santago and Gage model, partial volume contributions are modelled using uniform distributions lying between the means of each pair of pure tissues that share a common boundary, again convolved with a Gaussian distribution representing the acquisition noise. In this work, the model is refined by dividing the uniform distribution into a pair of complementary triangular distributions. Each triangular distribution represents the volumetric contribution of one of the pure tissues to the partial volume voxels. If the triangular distribution is defined using the line equation $y = kx + c$ then its convolution with the Gaussian distribution is given by

$$\int_a^b (kt + c) \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(t-g)^2}{2\sigma^2} dt \tag{2}$$

Note that the mean parameter has no effect on the convolution process. Hence, by solving the integral we find [1]

$$-\frac{(kg+c)}{2}\{erf[\frac{g-b}{\sqrt{2}\sigma}] - erf[\frac{g-a}{\sqrt{2}\sigma}]\} - \frac{k\sigma}{\sqrt{2\pi}}\{\exp[-\frac{(g-b)^2}{2\sigma^2}] - \exp[-\frac{(g-a)^2}{2\sigma^2}]\} \tag{3}$$

The parameters a and b represent the non-zero range of the distribution. It is straightforward to find the intercept c and the slope k parameters of the line that defines the triangle, for instance, by normalising the area under the distribution function. However normalisation is not necessary at this stage and it is sufficient to assume that the maximum height of the distribution function is constant, or simply is equal to unity. Then the slope parameter for a left-angled triangular distribution is $k = -1/(b-a)$ and its intercept is $c = -kb$. Similarly, the parameters of a right-angled triangular distribution using identical range are $k = 1/(b-a)$ and $c = -ka$. In the work described here, the total model consisted of five pure tissue components and eight partial volume components.

For partial tissue volumes we use a convolution of a triangular distribution $v_i(g; a, b)$ with a Gaussian distribution $G(g; \sigma, \mu = a)$ to get proper density functions. Here $\mu = a$ means that each pure tissue Gaussian model is attached from the right-hand side to a left-angled triangular model and from the left-hand side to a right-angled triangular model. These triangular models account for the neighbouring mixed tissues within two different ranges which come together at the point a. The point b for each of the triangular models is identical to the point $\mu = a$ corresponding to the neighbouring Gaussian distribution. Consequently, there are one right-angled and one left-angled triangular models between each two neighbouring Gaussian distributions which are hard locked to their means at $\mu_1 = a$ and $\mu_2 = b$. This is to say, there are four triangular models (two right-angled and two left-angled) with one side fixed at the mean point of each pure Gaussian distribution. When normalized, the summation of each two triangular distributions in the same range initially makes a rectangular distribution. Obviously the left-most Gaussian distribution has no triangular models attached to it from its left-hand side. Similarly, the right-most Gaussian distribution has no triangular models attached to it from its right-hand side. Our density functions which are represented by $p(g|v_i)$ are equivalent to our ICA components. Here our model consists of five Gaussians and eight corresponding triangular density functions between them. This makes four pairs of (a, b) together with an identical σ for all components. However as parameter b for each range is identical to parameter a for the neighbouring range, only six parameters are sufficient to account for all the model components. These are simply the five mean parameters of the five Gaussians plus the single σ parameter. Initial values used for these tissue parameters could correspond to five equal partitions of the widest existing histogram range.

The next step is to determine all weighting parameters α_{ij} for histograms f_j and components $p(g|v_i)$ from the EM algorithm. We approximate our data histogram as a linear combination of all density f functions defined so that

$$f_j \approx \sum_i \{\alpha_{ij} p(g|v_i)\} \tag{4}$$

The process of estimating the weighting parameters is iterative. Here the iterated quantity for weighting parameters is called α'_{ij}. The iterated quantity is estimated using a summation over the whole histogram for bin value of every pixel grey level g and its corresponding probability of belonging to each of the components. Hence we can write

$$\alpha'_{ij} = \sum_g \{f_{gj} P(v_i|g)\} \tag{5}$$

Note that these probabilities are computed using the density functions and current weighting parameters α_{ij} based on Bayes theory. Specifically we have

$$P(v_i|g) = \frac{\alpha_{ij} p(g|v_i)}{\sum_i \{\alpha_{ij} p(g|v_i)\}} \tag{6}$$

The initial values used for α_{ij} to converge to stable values could be unity. The process involving Equations (6) and (5) is iterated until $\alpha'_{ij} \approx \alpha_{ij}$. To speed up the process of optimisation using the EM algorithm, updated values of α_{ij} are used after each optimisation step as new initial values. Alternatively, we can just set the number of iterations to a small number that is usually decided empirically.

Once updated values of α_{ij} are in hand, it is straightforward to compute the cost function L_j for the histogram f_j. The appropriate cost function to use can be derived from the probability of getting the observed sample using Poisson assumptions. This result in the conventional likelihood function

$$L_j = -\sum_g \{f_{gj} \log f_j\} = -\sum_g \{f_{gj} \log[\sum_i \{\alpha_{ij} p(g|v_i)\}]\} \tag{7}$$

Equation (7) is correct subject to a fixed normalisation of the model $f_j = \sum_i \{\alpha_{ij} p(g|v_i)\}$ (in accordance with use of Extended Maximum Likelihood). We therefore perform normalisation on each model histogram so that the area under each model becomes equal to the number of corresponding data points.

As this expression is proportional to the joint probability, the optimisation of this function is valid for parameter estimation. However, the unknown scale factor makes the measure unsuitable as an absolute estimate of fit quality (see below). The total cost function when summed over all image regions is

$$M_v = \sum_j \{L_j\} \tag{8}$$

This expression is optimised using the downhill simplex method of Nealder and Meade [2], with restarts in order to avoid local minima. Convergence of the process to a global minimum implies that optimum estimates have been found for all model parameters, i.e. the weighting parameters and the tissue parameters.

2.2 Test Phase

Once an approximate model is obtained, the optimisation process does not need to be executed again for the test data. Instead, for each new test image, it is sufficient to build the J data histograms with specifications similar to those used in the training phase. The difficulty here is that since the grey level values stored in image files from different imaging equipments may correspond to different scale factors, we have no way of knowing equivalent grey levels from different images. Hence we apply a scale factor that is varied in the range [0.5, 2.0] to find the best fit of the input data to the model. Obviously, using the model histogram specifications some of the scales may result in overflow and/or underflow in the data histograms. Such scale factors cannot correspond to the best fit and are ignored. To allow the majority of the test data points to be included in the histogram, we add some 10% tolerance on the model histogram range during the training phase. In order to obtain an absolute measure of similarity, the out-of-fit measure is then computed using the Matusita measure [5, 6]

$$M_v = \frac{1}{4JH} \sum_{j,g} [\sqrt{\sum_i \alpha_{ij} p(g|v_i)} - \sqrt{f_{gj}}]^2 \tag{9}$$

where H is the number of bins for each histogram. This can be considered as a χ^2 test, (i.e. the $\sqrt{f_{gj}}$ values are assumed to have a Gaussian distribution with a standard deviation of 1/2).

As the search for the best corresponding scale is an optimisation with one parameter it is amenable to direct search. We set the scale step to 0.02 and compute the out-of-fit measure at 76 scales in the range [0.5, 2.0] (this involves no more evaluations than would be expected if using a conventional optimisation). The minimum out-of-fit measure may then be considered as the best fit for the corresponding scale value. One may proceed further by interpolating the minima from a quadratic equation to three points for increased accuracy. These are the point at the global minimum and the two side points with identical scale difference from the centre.

3 Experiments

We set the number of bins to 108 and divide each image into 4 by 4 windows which makes 16 corresponding histograms. We trained the model using a single MR brain image (anatomy4: slice-12) shown in Figure 1. The algorithm was converged with an out-of-fit measure at 0.6172. The corresponding 16 data histograms are shown in Figure 3 with the model superimposed. We set the initial parameters so that $\sigma = 3.0$ and the five Gausian mean parameters were located at the bin points $\mu_1 = 18.0$, $\mu_2 = 36.0$, $\mu_3 = 54.0$, $\mu_4 = 72.0$ and $\mu_5 = 90.0$. After the convergance of the EM algorithm using simplex with maximum error at 0.00001, the final standard deviation parameter of the Gaussians is given at $\sigma = 2.4227$ and the five mean parameters of the Gaussians are given at $\mu_1 = 10.1972$, $\mu_2 = 41.1133$, $\mu_3 = 60.2390$, $\mu_4 = 70.07797$ and $\mu_5 = 86.9204$.

To study how the out-of-fit measure behaves, we have also varied the number of windows as listed in Table 1. As expected, the larger the number of histograms the smaller the out-of-fit measure, and so more accurate fits are obtained. Of course, increasing the number of histograms to some extent is advantageous. However, with too many histograms used during the training phase, there will be disadvantages mainly during the test phase. Having too many histograms means small numbers of data points in each histogram, which leads us to loosing the histogram shape corresponding to the tissues of interest. Hence the main disadvantage is lowering the ability of the model in recognising between valid and invalid test images. Another disadvantage is that the training process takes much longer.

3.1 Valid test data

We tested 9 new MR images (anatomy4: slice-10 to slice-19) against the model. The 10 image slices we used to test the model are shown in Figures 1 and 2. The results are listed in Table 2. It is clear from this experiment that the out-of-fit measure in all cases is close to its value for training data. Since the deviation from the typical measure value is small for the whole set, this means that the data-set consists of valid data. Alternatively, one may investigate training using several different images and using an average model. In what follows we perform further tests using different data to evaluate the algorithm.

3.2 Re-scaled test data

One issue of data quality that frequently occurs is that data is under-quantised during acquisition or following an image conversion for file storage. This often has negative effects on sophisticated analysis processes, particularly

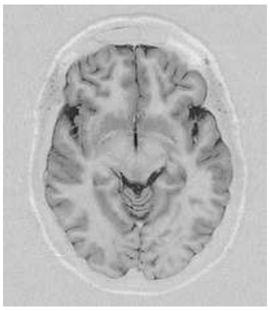

Figure 1: Image slice-12 used for training (from anatomy4 MR brain images).

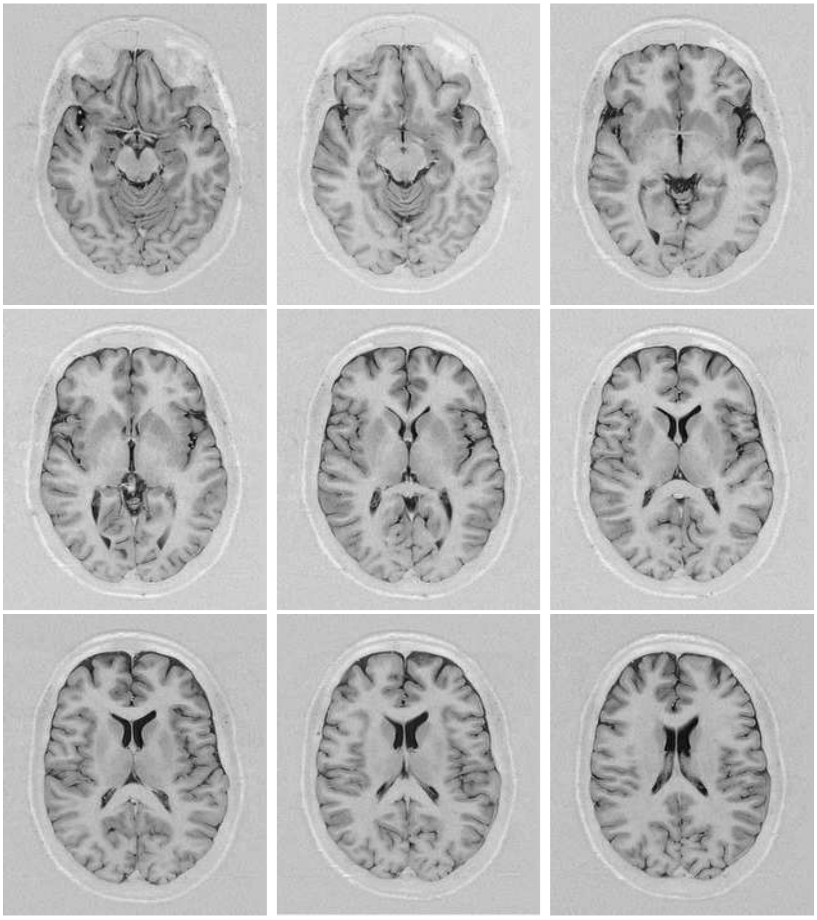

Figure 2: Image slices used for testing (from anatomy4 MR brain images); slice numbers from left-to-right: 10, 11, 13 (top), 14, 15, 16 (middle), 17, 18 and 19 (bottom).

number of windows (X * Y)	out-of-fit measure
4 (2 , 2)	1.2104
6 (2 , 3)	1.0070
9 (3 , 3)	0.8412
12 (3 , 4)	0.7113
16 (4 , 4)	0.6172
20 (4 , 5)	0.4856
25 (5 , 5)	0.3782
100 (10 , 10)	0.1386

Table 1: Out-of-fit measures for the training phase using slice-12 for different number of image windows (histograms).

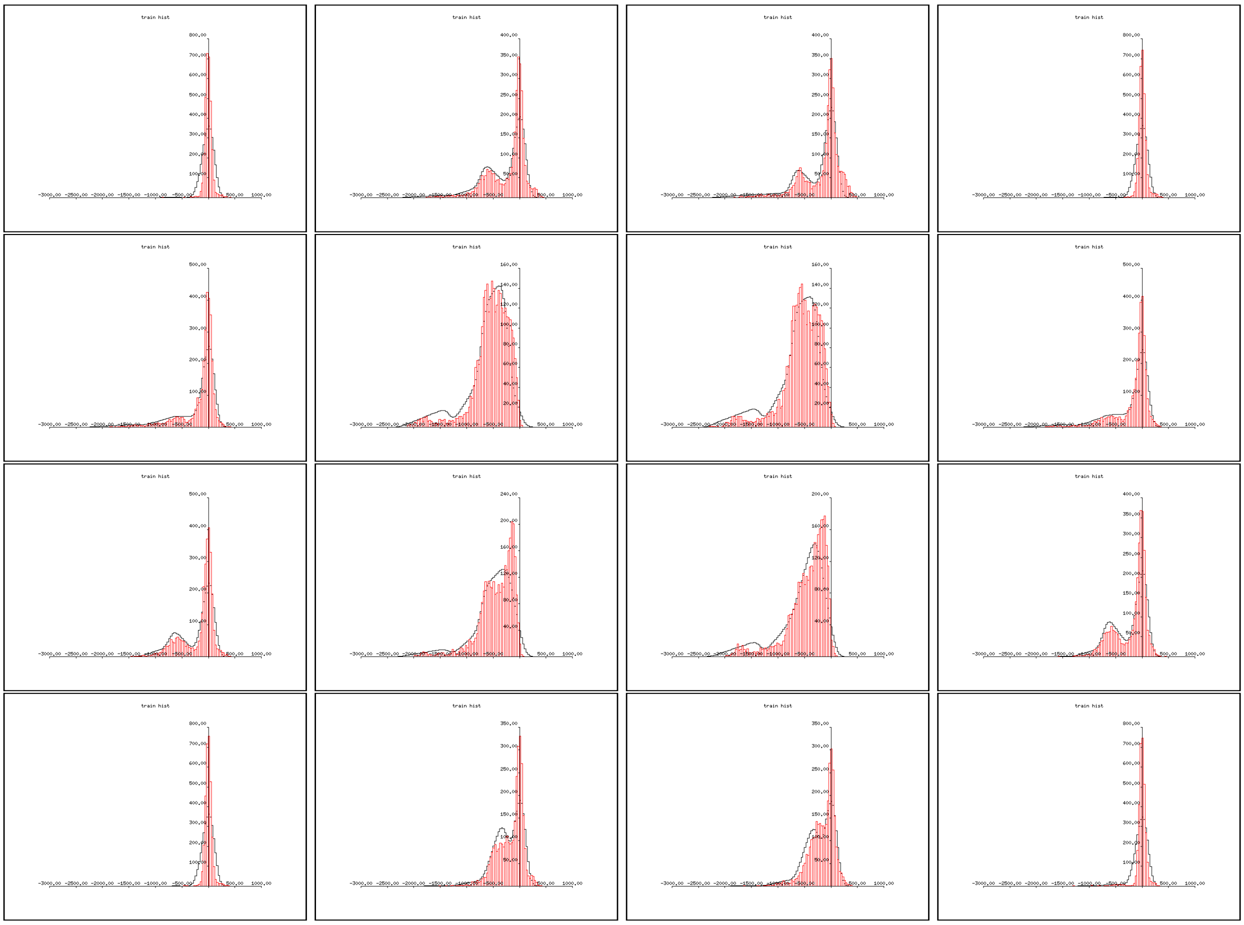

Figure 3: Data histograms (red) with the model superimposed (black) for the image slice-12 (16 windows) at the end of training; the histogram locations on the figure map spatially to their corresponding windows on the image.

slice number	scale factor	out-of-fit measure
10	1.1451	0.6714
11	1.1000	0.5625
12	1.1200	0.5093
13	1.2400	0.5143
14	1.1795	0.5317
15	1.1600	0.5484
16	1.2000	0.5525
17	1.2000	0.6043
18	1.2000	0.6793
19	1.2200	0.8694

Table 2: Test results on original data for the algorithm trained using slice-12 (16 windows): columns from left to right refer to the image slice number, the scale factor giving the best fit, and the corresponding out-of-fit measure.

those that involve data density modelling or require spatial derivatives. Such a process directly modifies the structure of the image histogram and should be detectable via our quality checking process. We have quantised the test images of Figures 1 and 2 leading to smaller numbers of non-zero values (≈ 32), so that gaps appear between bins in their corresponding histograms. We expect the out-of-fit measure to increase dramatically. These results are shown in Table 3.

slice number	scale factor	out-of-fit measure
10	1.0049	1.4151
11	1.0105	1.3337
12	1.0081	1.3078
13	1.0115	1.3588
14	1.0067	1.3522
15	1.0113	1.3805
16	1.0104	1.4244
17	1.0107	1.4710
18	1.0049	1.5720
19	1.2400	1.7863

Table 3: Test results on re-scaled data for the algorithm trained using slice-12 (16 windows): columns from left to right refer to the image slice number, the scale factor giving the best fit, and the corresponding out-of-fit measure.

3.3 Invalid test data

To test using some MR images of different imaging parameters or different tissues, we processed the MR images shown in Figure 4 so that their histograms range fall in the range used when training the algorithm. While this process does not change the visual appearance of the images, it provides a better test of the sensitivity of our method to identify new image structure. The corresponding results are shown in Table 4.

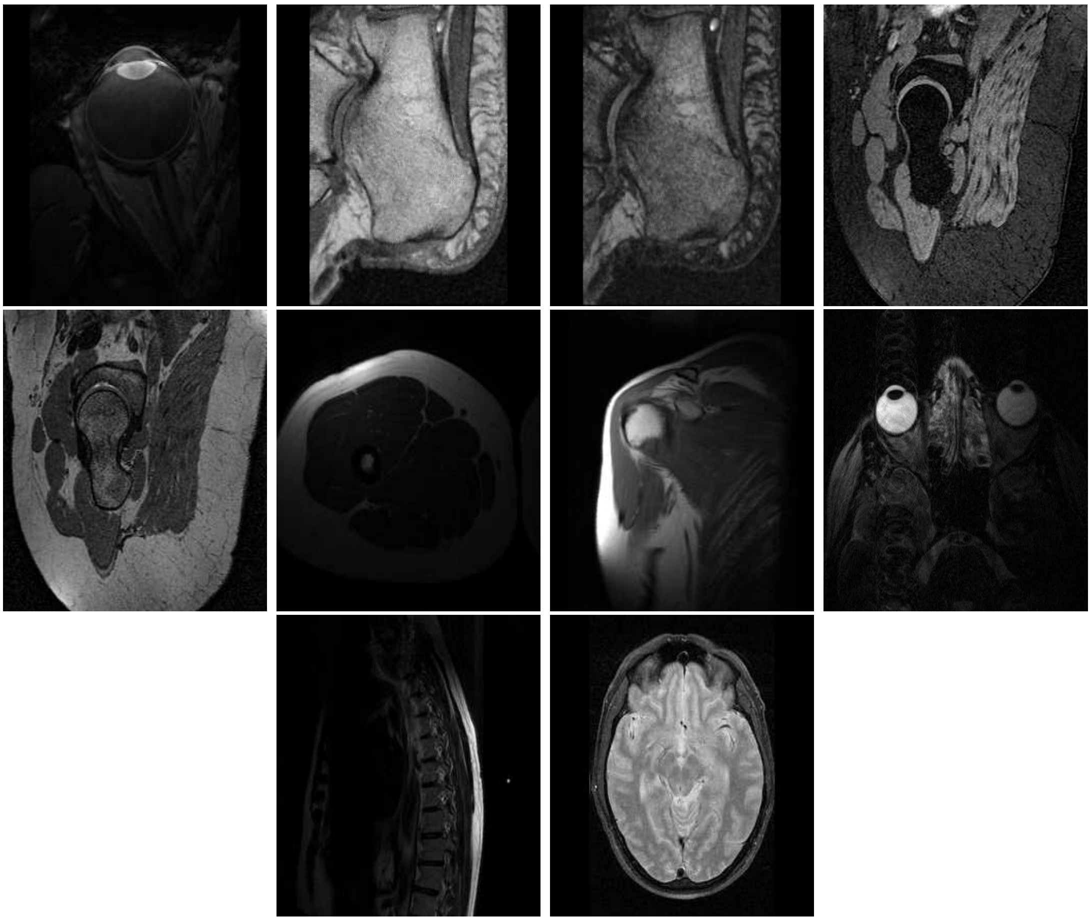

Figure 4: Image slices used for testing (from coil MR brain images); from left-to-right: (1) eye, foot0, foot1, hip1, (2) hip2, hip3, shoulder, skin and (3) spine and brain-pd.

slice number	scale factor	out-of-fit measure
eye	0.5200	12.4878
foot0	0.7601	6.4304
foot1	0.5402	8.3516
hip1	0.5531	9.6117
hip2	0.5200	6.9464
leg	0.5200	13.8013
shoulder	0.5920	11.1210
skin	0.5358	11.5766
spine	0.5200	13.3478
brain-pd	0.6831	9.9263

Table 4: Test results on re-scaled invalid data for the algorithm trained using slice-12 (16 windows): columns from left to right refer to the image slice number, the scale factor giving the best fit, and the corresponding out-of-fit measure.

3.4 Conclusions

We have identified the problem of use of algorithms on data that is not suitable for such processing when analysis software is used as a measurement tool. Conventional approaches to the issue of quality control involve checking imaging parameters or signal to noise. Such tests are often inappropriate for identifying more subtle problems, particularly when obtaining data from alternative imaging equipment. Unfortunately, such problems are often difficult to identify without significant technical knowledge and access to appropriate investigative tools. This level of expertise is expected to be beyond that avaliable to many users. In order to deal with this problem we have suggested a supplementary statistical test based upon the construction of a component model, trained on sub-regions of images known to be suitable for analysis. We have shown how this technique will identify not only quantisation effects, but also novel histogram structure arising from different biological structures. Our suggestion is to use such tests as one component of a quality control system, which advises users as to the likely suitability of datasets.

References

[1] P A Bromiley and N A Thacker. Multi-dimensional medical image segmentation with partial volume and gradient modelling. *Annals of the BMVA*, 2008(2):1–22, 2008. www.bmva.org/annals/2008/2008-0002.pdf.

[2] 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.

[3] K Rank, M Lendl, and R Unbehauen. Estimation of image noise variance. *IEE Proc Vis Image Signal Process*, 146(2):80–84, 1999.

[4] P Santago and H D Gage. Quantification of MR brain images by mixture density and partial volume modelling. *IEEE Trans Med Imaging*, 12:566–574, 1993.

[5] N A Thacker, F Ahearne, and P I Rockett. The Bhattacharryya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 34(4):363–368, 1997.

[6] N A Thacker and P A Bromiley. The effects of a square root transform on a Poisson distributed quantity. Technical Report TINA Memo no. 2001-010, The University of Manchester, 2001. www.tina-vision.net/doc/memos/2001-010.