NeuroImage 178 (2018) 668–676

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

A supervised learning approach for diffusion MRI quality control with minimal training data

Mark S. Graham *, Ivana Drobnjak, Hui Zhang

Centre for Medical Image Computing & Department of Computer Science, University College London, UK

A B S T R A C T

Quality control (QC) is a fundamental component of any study. Diffusion MRI has unique challenges that make manual QC particularly difficult, including a greater number of artefacts than other MR modalities and a greater volume of data. The gold standard is manual inspection of the data, but this process is time-consuming and subjective. Recently supervised learning approaches based on convolutional neural networks have been shown to be competitive with manual inspection. A drawback of these approaches is they still require a manually labelled dataset for training, which is itself time-consuming to produce and still introduces an element of subjectivity. In this work we demonstrate the need for manual labelling can be greatly reduced by training on simulated data, and using a small amount of labelled data for a final calibration step. We demonstrate its potential for the detection of severe movement artefacts, and compare performance to a classifier trained on manually-labelled real data.

Introduction

Quality control (QC) involves ensuring a dataset meets a certain set of standards before the dataset is given the clearance for inclusion in subsequent analyses. In MRI there are a large number of potential artefacts that need to be identified, to enable problematic images to either be excluded or accounted for in further processing and analysis. The gold standard for identification of these is visual inspection of the data.

There are a number of challenges with manual QC. For a typical study, which may involve hundreds of subjects, the process can be extremely time-consuming. This is especially true in diffusion MRI (DW-MR) where many volumes might be acquired for every subject, and there are numerous artefacts that each volume must be screened for, such as intra-volume movement, radiofrequency spikes, chemical shifts, and ghosting. The current trend towards acquiring increasingly large datasets means the time required for human QC is becoming prohibitive. The HCP (Essen et al., 2012) acquired data for 1200 subjects with almost 300 DW-MR volumes per subject and the UK Biobank will eventually acquire imaging data for 100,000 subjects with over 100 volumes per subject (Miller et al., 2016). Manual QC is also subjective. Each rater has their own sensitivity and specificity which cannot be easily altered, meaning the data is either QCed by a single rater, leading to a single standard but requiring large amounts of time, or many raters look at the data, which requires less time but means variable standards are applied across the dataset. Some artefacts can also be hard to detect with manual QC, such as ghosting artefacts which require the careful examination of every slice in a volume. These challenges have led to an increased interest in automated methods for QC.

Automated methods for QC fall into two classes. The first kind extracts tailored features from the datasets and applies hand-tuned cutoffs to determine whether each volume contains artefacts (Liu et al., 2010; Oguz et al., 2014). The second kind are supervised learning approaches. These involve extracting features from the datasets and then using a training set, obtained from manual QC of a proportion of the data, to learn the mapping between these features and the classification of each volume as passing or failing QC. Recently these approaches have used support vector machines (SVMs), random forest classifiers (Esteban et al., 2017) and ensembles of classifiers (Alfaro-Almagro et al., 2017). Whilst promising, both types of approach report performance significantly below that of a human rater.

Recently, deep-learning based convolution neural networks (CNNs) (Goodfellow et al., 2016) have been demonstrated to provide near-human levels of accuracy for identifying motion artefacts in structural (Iglesias et al., 2017) and DW-MR data (Kelly et al., 2016). Unlike other supervised approaches, CNNs learn features from the data during training, rather than requiring them to be hand-crafted and supplied as input. CNNs tend to have many parameters requiring optimisation — often in the millions — meaning they typically require large, labelled datasets for training.

* Corresponding author.
E-mail address: mark.graham.13@ucl.ac.uk (M.S. Graham).

https://doi.org/10.1016/j.neuroimage.2018.05.077
Received 9 January 2018; Received in revised form 22 May 2018; Accepted 31 May 2018
Available online 5 June 2018

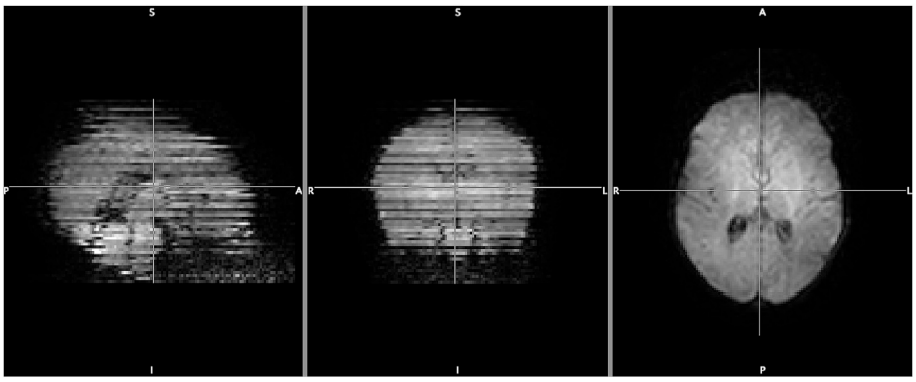


Fig. 1. Example of the intra-volume movement artefact. For interleaved acquisitions, which are common in DW-MR, the movement causes jagged edges perpendicular to the EPI plane, here the coronal and sagittal views. This misplacement of the signal is different to signal dropout, also caused by intra-volume movement, which leads to loss of the signal — an example of dropout can be seen in one of the most inferior slices of the coronal view in this subject.

Obtaining training datasets in medical imaging can be challenging. The acquisition of the data is time-consuming and expensive, and once acquired the data can be subject to ethical considerations or anonymisation requirements that prevent that data being shared freely. Labelling of such data is also challenging. For the case of QC, labelling requires a human rater to manually inspect each image volume and flag any that contain artefacts. The process is subjective, and the accuracy of the trained classifier will depend on the quality of the labelled dataset, so often a number of raters are used and their classifications combined in order to get a more reliable 'ground-truth' for the dataset. Furthermore, a tool trained on a specific set of training data may not generalise well to datasets acquired with different protocols or hardware, meaning new training datasets may need to be labelled for each new study.

One potential way to address these issues is to use simulated data. Simulation could circumvent the need for human labelling by producing realistic datasets, along with ground-truth labels, for training machine learning tools on. In the case of QC, a simulator that was capable of producing datasets containing artefacts, such as motion, could be used to produce a training set. Little research has been done to investigate the feasibility of a simulation-based approach to training supervised learning tools. Until recently, DW-MR simulations have not been capable of producing realistic data complete with artefacts, but these have both been demonstrated by a recently proposed framework (Graham et al., 2016), enabling these new simulation-driven approaches to be investigated.

In this work, we aim to investigate the feasibility of a supervised-learning approach to QC that uses simulated data and a small amount of real data for calibration; thus greatly reducing the amount of training data required when compared to classifiers trained on real data. As a first step, we focus on the problem of detecting intra-volume movement in DW-MR data. Intra-volume movement refers to both signs of head movement and the signal dropout that this gives rise to — see Fig. 1. We focus on this artefact because they are often not tackled in QC, as checking every volume in a dataset can be extremely time-consuming. Volumes containing this artefact typically need to be identified in QC so that they can either be removed, or information about them can be used as confounds in later statistical analysis (Yendiki et al., 2014). Whilst post-processing techniques have been proposed to correct for intra-volume movement (Oubel et al., 2012; Marami et al., 2017; Andersson et al., 2017), it has been reported these fail for severe cases and an initial QC needs to be performed to remove especially bad volumes before processing (Kelly et al., 2016). We compare the performance of QC classifiers trained on real and simulated data. We aim to investigate how close a simulation-trained classifier can get to the state-of-the-art, a real-trained classifier, for QC of movement artefacts (Kelly et al., 2016).

Table 1
Information on the dHCP subjects used in the study.

Subject	M/F	Gestational age at birth/weeks	Gestational age at scan/weeks
CC00069XX12	M	39.14	39.57
CC00099AN18	F	37.43	37.71
CC00117XX10	M	41.57	42.14
CC00122XX07	M	37.43	38.29
CC00126XX11	M	38.14	38.29
CC00138XX15	F	41.43	41.57
CC00162XX06	M	40.14	40.86
CC00164XX08	M	38.71	38.86
CC00168XX12	M	40.14	43.86
CC00170XX06	M	37.86	38.43

Methods

This section details the data, both real and simulated, used in this work and describes the classifier that was trained on these data.

Data

Real

Ten subjects were taken from the developing Human Connectome Project (Hughes et al., 2016) (dHCP), which contains MRI data acquired in neonates. Table 1 shows the age and sex of the subjects. These were chosen because neonatal scans tend to contain large amounts of movement. The data was acquired on a 3T Philips Achieva, consisting of a spherically optimized set of directions on 4 shells ($b = 0\,s\,mm^{-2}$: 20, $b = 400$: 64, $b = 1000$: 88, $b = 2600$: 128) split into four subsets, each with a different phase-encoding (PE) direction. It was acquired using a multiband acceleration factor of 4, SENSE factor 1.2 and partial fourier 0.86, TR/TE 3800/90 ms. The reconstructed data has matrix size 128×128, with 64 slices per volume resolution $1.17 \times 1.17 \times 1.5$ mm (dHCP Consortium, 2017).

For this study, the $b = 2600\,s\,mm^{-2}$ volumes were removed as they contained very little signal, which caused even manual QC to be challenging. This left 172 volumes per subject. Manual QC was performed by visual inspection, with one rater assigning a label of either acceptable or unacceptable to each volume. The rater classified the whole dataset twice, on two separate occasions, to provide an estimate of intra-rater agreement.

Simulated

Data was generated using the simulation framework described in Graham et al. (2016). In brief, the simulator uses an input object that

Table 2
Tissue parameters used for the simulations in this chapter. Proton density ρ is in arbitrary units.

Tissue	T_1/ms	T_2/ms	ρ
Grey matter	2200	200	0.8
White matter	2850	250	0.8
CSF	3700	280	0.8

describes the proton density of WM, GM and CSF at each voxel, as well as the T_1 and T_2 parameters for each tissue type. This simulator also takes as input a pulse sequence describing the RF pulses and gradients to be applied, a description of any artefacts (such as motion), and a representation of the diffusion-weighting at every voxel. The Bloch equations are solved for every voxel in the input object, enabling realistic data along with its artefacts to be generated (Drobnjak et al., 2006, 2010).

Simulated data was designed to be visually similar to the dHCP data. Data was simulated using the same *b*-values and directions as the dHCP. Voxel size and FOV were selected to minimise computation time: 2.5 mm isotropic and $72 \times 86 \times 55$ voxels — this results in lower resolution data than the dHCP, but to simulate data at a resolution of $1.17 \times 1.17 \times 1.5$ mm would be computationally prohibitive. Seven subjects were simulated using input objects derived from different subjects from the HCP, according to the process described in Graham et al. (2016). Neonatal DW-MR images have different contrast to adult data, with much reduced contrast between GM, WM and CSF. MR parameter values were modified to increase the visual similarity between simulated and real data — T_1 parameters were taken from Williams et al. (2005), T_2 parameters from Leppert et al. (2009) and then adjusted further to maximise visual similarity with the dHCP datasets — the final set of parameters used are shown in Table 2.

The simulated data contained motion and eddy-current artefacts. Known motion was injected into the datasets during simulation. The traces describe the object's translations along and rotations about each of the three axes, and are discretised in time so that they provide these six parameters for each slice that is to be acquired. The traces were synthesised to contain sudden 'jerks' of the head that consist of the head moving to a certain location and then back to its original location in a period less than T_R, the repetition time — these give rise to the intra-volume movement artefact shown in Fig. 1. These 'jerks' were chosen randomly so that there was a 40% chance of any given volume containing movement, and movement was equally likely to be a translation or rotation around any of the three axes. In addition to this a slow drift component was also added so that the overall position of the head changed across the simulation of all volumes for a subject. In practice the traces of the 'jerks' were approximated as Gaussians with time period $0.2*T_R$ and the slow drift was a Cosine with a period in the range $5*T_R$-$15*T_R$. Fig. 2 shows an example trace for all 172 volumes of a subject, and Algorithm 1 describes more precisely how the traces were generated. Interleaved slice-ordering was simulated, without multiband, so that these intra-volume movement spikes produced the characteristic zig-zag edge pattern as seen in Fig. 1. Signal dropout was also simulated. In the dHCP data signal dropout is often, but not always, present in volumes that show other signs of severe intra-volume movement. To reflect this, dropout was added to a volume containing significant motion with a probability of 70%. Dropout is applied by directly reducing the signal of slices in k-space, rather than simulating the effect of the interaction between movement and the diffusion gradients. This simpler approach still produces realistic-looking dropout artefacts (Andersson et al., 2016). The full details of how dropout was added is described in Algorithm 2. Eddy-current artefacts were included in the data using the method described in Graham et al. (2016).

Algorithm 1 Synthesising movement traces for the simulated data

for Each volume to be simulated **do**
  Set volume's movement trace for T_x,T_y,T_z,R_x,R_y,R_z to 0 at every timepoint
  Draw number from random-number generator (RNG)
  if number $\leq$ 0.4 **then**
    for Each movement trace T_x,T_y,T_z **do**
      Draw number from RNG
      if number $\leq$ 1/6 **then**
        Generate Gaussian motion spike with height randomly selected between 0 and 10 mm, standard deviation 0.2*repetition time (T_R)
        Add Gaussian to volume's trace, centred on a randomly selected timepoint
      end if
      Select time period in interval [$5*T_R$, $15*T_R$]
      Add cosine with this period to movement trace to simulate slow drift
    end for
    for Each movement trace R_x,R_y,R_z **do**
      Draw number from RNG
      if number $\leq$ 1/6 **then**
        Generate Gaussian motion spike with height randomly selected between 0 and 10°, standard deviation $0.2*T_R$
        Add Gaussian to randomly selected location in volume's trace
      end if
      Select time period in interval [$5*T_R$, $15*T_R$]
      Add cosine with this period to movement trace to simulate slow drift
    end for
  end if
end for

Algorithm 2 Adding signal dropout to simulated data

```
for Each volume to be simulated do
    if Volume's motion trace has any translations ≥ 2.5 mm or rotations ≥ 2.5° then
        Draw number from RNG
        if Number ≤ 0.70 then
            for Every slice in volume do
                Draw number from RNG
                if Number ≤ 0.85 then
                    Draw number from RNG
                    Multiply signal in slice by number drawn
                end if
            end for
        end if
    end if
end for
```

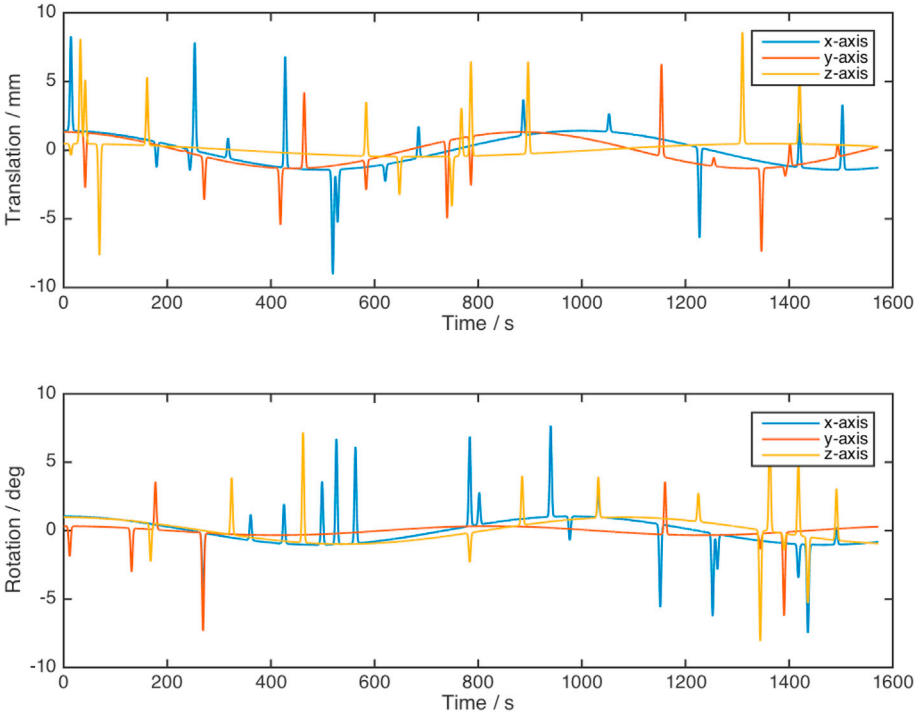


Fig. 2. Example simulated motion trace for all 172 volumes in a dataset.

Labels were assigned to each volume using the following scheme. The amount of intra-volume movement for each volume was calculated for each of the three rotations and three translations. If all of the translations were less than 1 mm and rotations less than 1°, the volume was given a label 0, for acceptable. If any of the translations were greater than 1 mm and less than 2.5 mm, or rotations greater than 1° and less than 2.5° the volume was assigned 1, for moderate. If any translations were greater than 2.5 mm or rotations greater than 2.5°, the volume was assigned a 2. It was observed that inclusion of volumes with moderate movement (label 1) with the volumes with severe movement (label 2) made it much more difficult to train the classifier. These volumes included very subtle signs of motion that were challenging to QC when visually inspected. By contrast, we found that most volumes of the real data were much more straightforward to classify; it was usually very obvious whether or not they contained movement artefacts. For this reason it was decided to remove the 108 volumes with moderate motion from the simulated dataset. This left a total of 1096 volumes, 732 without movement and 364 with.

Classifier

We based our classifier on a type of neural network called a convolutional neural network. These networks have provided state-of-the-art performance in computer vision tasks in recent years (Russakovsky et al., 2015), making them sensible candidates for the task of identifying motion-artefacts in scans. A drawback of such networks is that they contain millions of parameters, and so they typically require large amounts of training data and large amounts of computational power to successfully train. To circumvent this we adopted a transfer learning approach (Pan and Yang, 2010), which consists of taking a classifier trained to perform on a certain task and re-training a small number of parameters using a small amount of data to perform well on another, often similar, task.

Our transfer learning approach here is similar to that described in Kelly et al. (2016), where they successfully trained a classifier on real data to detect motion artefacts. We used the pre-trained InceptionV3 network as the base network, which has achieved state-of-the-art

performance in the classification of natural images (i.e. photos taken on standard cameras of everyday objects such as dogs, boats, cars) (Szegedy et al., 2015). To finetune InceptionV3, the top layer of the network was removed, and replaced with a fully connected layer with 16 neurons, followed by a prediction layer with 2 neurons for the two classes in our problem (motion-corrupted or normal). All parameters were fixed apart from those in the newly added layers, vastly reducing the number of parameters required for training.

The classifier was trained by passing it sagittal slices through the brain along with ground-truth labels. We trained two classifiers: one on real data, and one on simulated data, using seven subjects for training as in Kelly et al. (2016). Five subjects were used for training, and two for validation. We chose to use three sagittal slices from each volume at training time, though in principle using more might provide better results. One of the slices was taken from the central plane of the volume and the other two from either side of this central plane, towards the edges of the brain. For the real data, these side slices were 14 slices away from the middle slice on either side, for the simulated data, these were 16 slices away, as the simulated brains were slightly larger.

For both classifiers, images were zero-padded along the shorter dimension to make them square, resized to 299 by 299 pixels, and replicated three times for the three channels of the network (a fixed requirement of the InceptionV3 network). Each image was scaled so that its intensity lay between −1 and 1. Each classifier was trained for 30 epochs using the Adam optimizer with a learning rate of 0.001 (Kingma and Ba, 2014) and a cross entropy loss function. The classifiers were implemented in Keras (Chollet et al., 2015). Training took less than 20 min on a Titan X Pascal GPU.

Testing was performed on the three reserved dHCP subjects. To assign a label to each volume, a certain number of slices were extracted from the volume and classified; if the mean of these scores was greater than a certain threshold, t, the volume was labelled as containing motion. Each brain had approximately 60 saggital slices, so we chose to score alternate slices to strike a balance between dense sampling and GPU memory constraints, giving a total of 30 scores per volume. For the real-trained classifier we used the natural threshold of $t = 0.5$. Despite attempts to match the simulated and real data in appearance, there are remaining differences (such as the noise patterns caused by the multiple coil receivers in the real data, and the presence of artefacts not simulated such as susceptibility) which the simulation-trained classifier has not 'seen' before, which affects the scores that it assigns to real data. This means the threshold for the simulation-trained classifier had to be calibrated using a single subject from the real training set. We experimented with two methods for determining the optimal threshold from this subject. In the first, the threshold that maximized the F1-score between the true and predicted labels for this subject was chosen for use at test-time. The F1 score is defined as:

$$F1 = 2*\frac{precision*recall}{precision + recall} \tag{1}$$

with precision and recall defined as:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

$$recall = \frac{TP}{TP + FN} \tag{3}$$

where TP are true positives, TN true negatives, FP false positives and FN false negatives. The rationale for this method was finding a threshold that balanced precision and recall. In the second, the greatest threshold that gave >95% recall for motion-corrupted volumes in this subject was chosen. The rationale for this method was that it may be preferable to ensure the majority of corrupted volumes are found, even if this means rejecting some false positives. Classification took 28 ms per volume, or 48 s for the 172 volumes in each dHCP dataset.

Experiment design

We aim to investigate how close a simulation-trained classifier can get to the state-of-the-art, a real-trained classifier, for QC of movement artefacts. Given the real-trained classifier has been claimed to approach the performance of a human-rater (Kelly et al., 2016), we first compare our real-trained classifier to the intra-rater variability in order to establish whether our implementation of it serves as a suitable benchmark. We then compare our simulation-trained classifier to the performance of the real-trained. As part of the evaluation of the simulation-trained classifier we compare the two proposed choices for evaluating the optimal threshold described earlier.

Results

Simulated and real data is shown in Fig. 3. Both simulation-trained and real-trained classifiers fit their validation sets well — the simulation-trained achieved 95% accuracy on the simulated validation set, and the real-trained achieved 93% accuracy on the real validation set.

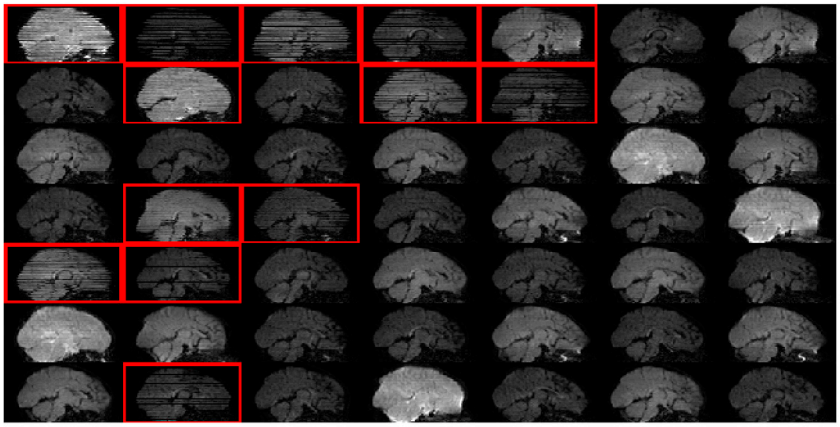
We determined the optimal threshold for the simulation-trained classifier using each of the seven real subjects in the training + validation set, in order to get a sense for the extent to which the threshold depends on choice of subject. Thresholds determined from the F1-score criterion were more tightly clustered (from 0.86 to 0.97) then those determined from the sensitivity criterion (0.71–0.96), indicating the F1 criterion is a more reliable way of determining a threshold. We decided to use the F1 criterion thresholds for the test dataset.

Both classifiers were tested on the three held-back dHCP subjects. Fig. 4 shows the precision-recall curve for the two classifiers. Whilst it was decided to use the F1 criterion for the results, thresholds for the sensitivity criterion are also plotted on this Figure to demonstrate the greater variance in precision/recall scores this introduces. The real-trained classifier achieved precision and recall of 97% and 98% for classification of corrupted volumes, results comparable to the state-of-the-art results reported in Kelly et al. (2016). Intra-rater agreement on the test set was 99%, showing this classifier approaches human level performance. The simulation-trained classifier achieved precision and recall of 95% and 93% for the most common F1-determined threshold (0.94, occurred in 3/7 subjects). If the lower range threshold was used (0.87) precision and recall was 83% and 97%, and for the upper threshold (0.97) these values were 96% and 85%. Fig. 5 shows results for both classifiers on some of the test data. Fig. 6 shows the mean classifier score for each volume in the test set, along with the classification for each volume.

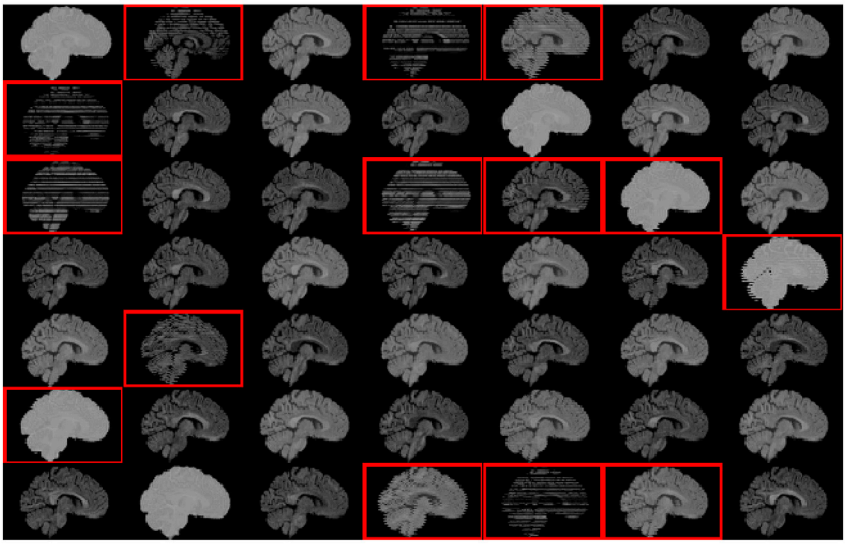
Discussion

In this work we compared the performance of a QC tool trained on simulated data to that of a tool trained on real data. The real-trained classifier achieved near-human performance, confirming the findings in Kelly et al. (2016). The simulation-trained classifier demonstrated performance approaching that of the real-trained classifier. It was able to detect the majority of the motion corrupted volumes in the test set, though it showed slightly reduced precision and recall compared to the real-trained classifier.

The classifier presented here is a modified version of the one presented by Kelly et al. (2016). It offers comparable performance when trained on real data whilst offering a number of improvements. Firstly, ours involves training a single neural network, compared to the 11 trained in Kelly et al. (2016). Our classifier only requires magnitude data, whilst the previous classifier uses both magnitude and phase data, and we don't need to distinguish between b-values for training. Our final decision is made by a simple thresholding of the classifier outputs, whilst Kelly et al. (2016) requires the additional training of a random forest classifier on the CNN outputs.



(a) Real data.



(b) Simulated data.

Fig. 3. Real and simulated data. Red bounding boxes indicate the volume was assigned a ground-truth label as containing intra-volume movement. Each image is taken from a different volume.

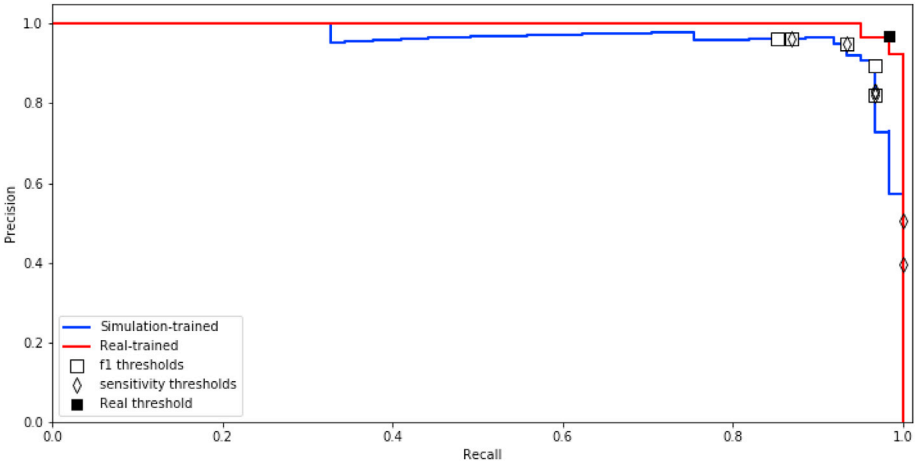


Fig. 4. Precision-recall curve for both classifiers in the test set, consisting of 516 volumes. The threshold for the real-trained classifier of 0.5 is plotted on the curve, as are the seven thresholds determined for both the F1- and sensitivity-based criterions for the simulation-trained classifier.

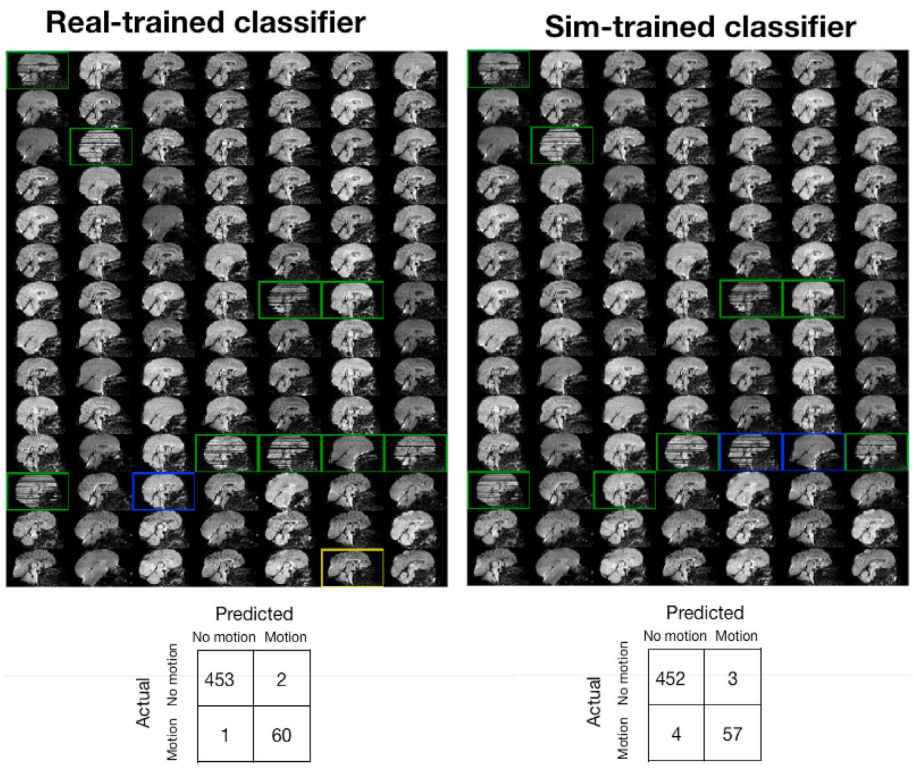


Fig. 5. Sample classifier results for images in the test-set. Green border indicates a correct classification as containing motion, blue borders indicate false-negatives and yellow borders are false-positives. Confusion matrices for classification on all 516 volumes in the test set are shown below. Threshold of 0.94 used for the sim-trained results.

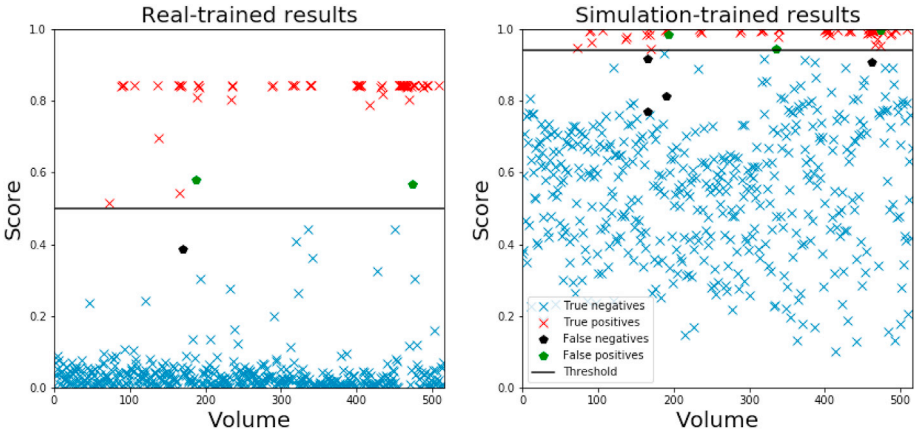


Fig. 6. Classifier scores for each volume in the test-set. The score is produced by averaging the classifier outputs for the 30 slices classified in each volume. Threshold of 0.94 used for the simulation-trained plot.

We only trained our classifier to detect movement artefacts, but there are many more artefacts that would ideally be identified by the QC process, and future work will look into extending the classifier. In theory, classifiers such as the one proposed in this work would be well suited to identifying any artefacts that a human is able to spot when inspecting slices through a volume, including ghosting, fat-artefacts, susceptibility and RF spikes. The classifier is currently not suitable for detecting artefacts that require visual comparison to other images (such as eddy-currents) or those best detected by examining model residual maps (Tournier et al., 2011). One advantage of training using simulated data is that the training set can be designed to include numerous examples of artefacts that might be very rare in practice (such as RF spikes). Training to identify these rarer artefacts on real data may require labelling a very large dataset in order to find sufficient training examples. The classifier is designed to be used on data before any post-processing (e.g. motion correction) is applied, in order to flag up problematic volumes, but they could potentially also be applied to automatically check if post-processing has successfully corrected visible problems in the data. We focused on DW-MR in this study, but other modalities (such as fMRI) contain both movement and other artefacts, and the approach demonstrated in this work can be adapted to produce classifiers for these modalities.

One potential source of error was the automatic threshold chosen to produce ground-truth labels for the simulated data. A volume with more than 2.5 mm translation or rotations greater than 2.5° was labelled as containing intra-volume movement. If volumes with this level of movement looked significantly different to volumes in the real data that were manually labelled as containing movement the simulation-trained

classifier's performance would be affected. We investigated this by performing manual QC on two subjects from the simulated dataset. Manual and automatic QC agreed for 95% of cases. When they disagreed it was because the automatic threshold picked up on slightly more subtle cases of movement. This could have led to a slightly more sensitive classifier than the real-trained one, but this does not seem to have been the case. Future work could investigate better ways of assigning labels to the simulated data. For example, the thresholding could take into account when the movement occurs in the acquisition of a volume as well as the type of movement occurring (e.g. translations, rotations).

There is room for more investigation. It would be interesting to understand how similar the real and simulated data need to be in order to obtain good performance. The resolution of simulated data was different to the dHCP data, and multiband wasn't simulated, meaning movement of slices was not correlated across a volume. Further work could determine whether these differences affect performance, as well as how performance depends on the amount of movement simulated, signal dropouts, image contrast and choice of *b*-values and directions. This ties into how well the trained classifiers will generalise to new, unseen datasets — for example a dataset acquired on adult subjects, or with a different protocol. It is worth nothing that the classifier would need to be re-trained for protocols where the artefacts manifest themselves differently, such as 3D imaging or non-EPI based methods. We could also look at how much training data is required for good performance; in this work we matched the amount of simulated and real data used for training, but we could test whether performance can be improved even further by using more simulated data for training. Moving to a GPU-accelerated simulator (Xanthis et al., 2014) would facilitate easily producing larger quantities or training data, and also enable higher-resolution data to be simulated. Whilst it is an advantage that the classifier only requires the images, it would be interesting to ascertain whether supplying further information (such as *b*-values) could assist classification.

There are some drawbacks to the approach described in this paper. These centre around the difficulties inherent in training on one domain (simulated data) and then classifying in a different domain (real data). This can be seen in Fig. 6. It shows that the simulation-trained classifier was good at spotting artefacts, and assigned high scores to volumes containing movement, but was much less sure for volumes that did not contain movement, producing scores with a very large spread. This contrasts with the real-trained classifier, which was able to assign high scores to volumes with movement and low scores to volume without. This meant the simulation-trained classifier had a smaller margin for error, which caused the occasional large mistake: in Fig. 5, it can be seen the simulation-trained classifier predicted a false-negative on a volume that quite clearly contained movement artefacts — the errors in the real-trained classifier tend to be more straightforward 'borderline cases' that a human might find difficult to classify. The shift in scores caused by the transfer between domains also meant that the simulation-trained classifier still requires some labelled, real-data for calibration. Whilst this is still a big reduction in the amount of labelled data required when compared to the amount needed to train the classifier, it would be ideal if none was required. Furthermore, the choice of subject used for the calibration introduces variability into the final performance of the classifier. One potential way to address these issues would be to work on increasing the realism of the simulations, for example by including some of artefacts not simulated such as susceptibility, fat artefacts, and the complex noise distributions caused by multiple receiver coils. However, we believe a more promising way to address these limitations is by including recent research on domain adaptation in machine learning, i.e. getting a classifier trained to perform well in domain A (e.g. simulated data) to perform well on domain B (e.g. real data) without requiring any labelled data from domain B. One of the simplest ways to do this is to include a small amount of real data in training. Fig. 7 shows some preliminary results using this approach, training on the simulated dataset and 20 volumes from the real set. The classifier is able to produce much lower scores for normal volumes, enabling the standard threshold of 0.5 to be used. It seems unlikely these results are driven fully by the real data as the real data only contained a single volume containing movement artefacts, so the simulated data was necessary to help the classifier learn what a motion-corrupted volume looks like.

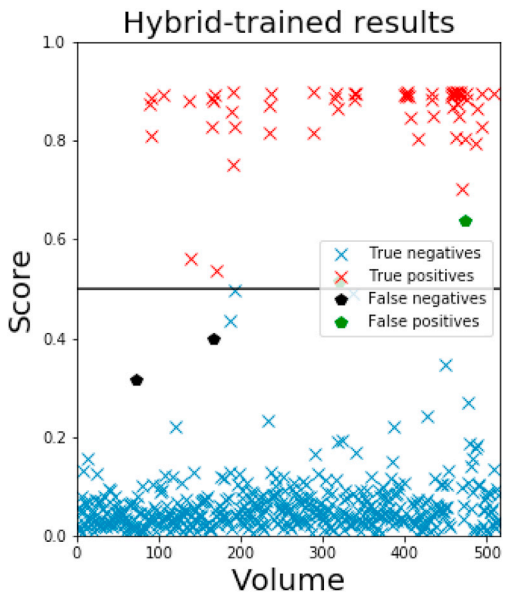


Fig. 7. Scores for a hybrid classifier trained on the full simulated dataset and 20 volumes from the real dataset. Results shown for a threshold of 0.5.

Whilst the hybrid result is promising, it is similar to the calibration approach in that it still requires some real data to be manually labelled. Ideally a classifier would be able to learn from the real data in a fully unsupervised manner, and these unsupervised approaches will be the focus of future developments. Examples of such approaches in the literature (outside of QC) include Kamnitsas et al. (2017) in which they encourage the classifier to learn features which are domain-invariant by pitting it adversarially against a discriminator which attempts to predict the domain the classifier is working in by examining the classifier's activations. In Bousmalis et al. (2017), a neural network is used to adapt simulated data to appear more like real data; training on this adapted data gives performance equivalent to training on real data.

Acknowledgements

MG is supported by the EPSRC (EP/L504889/1) and the EPSRC Centre for Doctoral Training (EP/L016478/1). HZ is supported by the EPSRC (EP/L022680/1), the MRC (MR/L011530/1) and the Royal Academy of Engineering Research Exchanges with China and India. ID is supported by the Leverhulme Trust. MG and HZ are additionally supported by the Royal Society International Exchange Scheme with China.

Data were provided [in part] by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

These results were obtained using data made available from the Developing Human Connectome Project funded by the European Research Council under the European Union's Seventh Framework Programme (FP/2007–2013) / ERC Grant Agreement no. 319456.

References

Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S., Jbabdi, S., Hernandez Fernandez, M., Vallee, E., Vidaurre, D., Webster, M., McCarthy, P.D., Rorden, C., Daducci, A., Alexander, D., Zhang, H., Dragonu, I., Matthews, P., Miller, K.L., Smith, S.M., 2017. Image

processing and quality control for the first 10,000 brain imaging datasets from UK biobank. bioRxiv. http://www.biorxiv.org/content/early/2017/04/24/130385.
Andersson, J.L., Graham, M.S., Drobnjak, I., Zhang, H., Filippini, N., Bastiani, M., 2017. Towards a comprehensive framework for movement and distortion correction of diffusion MR images: within volume movement. Neuroimage 152, 450–466. https://doi.org/10.1016/j.neuroimage.2017.02.085.
Andersson, J.L.R., Graham, M.S., Zsoldos, E., Sotiropoulos, S.N., 2016. Incorporating outlier detection and replacement into a non-parametric framework for movement and distortion correction of diffusion MR images. Neuroimage 141, 556–572.
Bousmalis, K., Irpan, A., Wohlhart, P., Bai, Y., Kelcey, M., Kalakrishnan, M., Downs, L., Ibarz, J., Pastor, P., Konolige, K., et al., 2017. Using Simulation and Domain Adaptation to Improve Efficiency of Deep Robotic Grasping arXiv preprint arXiv: 1709.07857.
Chollet, F., et al., 2015. Keras. https://github.com/fchollet/keras.
dHCP Consortium, 2017. dHCP data release documentation, 2017, version 1.1. https://data.developingconnectome.org/downloads/documentation/DataReleaseDocumentation.pdf.
Drobnjak, I., Gavaghan, D., Süli, E., Pitt-Francis, J., Jenkinson, M., Aug 2006. Development of a functional magnetic resonance imaging simulator for modeling realistic rigid-body motion artifacts. Magn. Reson. Med. 56 (2), 364–380. http://www.ncbi.nlm.nih.gov/pubmed/16841304.
Drobnjak, I., Pell, G.S., Jenkinson, M., sep 2010. Simulating the effects of time-varying magnetic fields with a realistic simulated scanner. Magn. Reson. Imag. 28 (7), 1014–1021. http://www.ncbi.nlm.nih.gov/pubmed/20418038.
Essen, D.C.V., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Penna, S.D., Feinberg, D., Glasser, M.F., Harel, N., Heath, A.C., Larson-prior, L., Marcus, D., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S.E., Prior, F., Schlaggar, B.L., Smith, S.M., Snyder, A.Z., Xu, J., Yacoub, E., Consortium, W.-m. H.C. P., Eeg, M.E.G., 2012. The Human Connectome Project: a data acquisition perspective. Neuroimage 62 (4), 2222–2231.
Esteban, O., Birman, D., Schaer, M., Koyejo, O.O., Poldrack, R.A., Gorgolewski, K.J., 2017. MRIQC: predicting quality in manual MRI assessment protocols using No-Reference image quality measures. bioRxiv 1–18. http://biorxiv.org/content/early/2017/02/24/111294.abstract.
Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 2016. Deep Learning, vol. 1. MIT press Cambridge.
Graham, M.S., Drobnjak, I., Zhang, H., 2016. Realistic simulation of artefacts in diffusion MRI for validating post-processing correction techniques. Neuroimage 125, 1079–1094. http://linkinghub.elsevier.com/retrieve/pii/S1053811915010289.
Hughes, E.J., Winchman, T., Padormo, F., Teixeira, R., Wurie, J., Sharma, M., Fox, M., Hutter, J., Cordero-Grande, L., Price, A.N., et al., 2016. A dedicated neonatal brain imaging system. Magn. Reson. Med.
Iglesias, J.E., Lerma-usabiaga, G., Garcia-Peraza-Herrera, L.C., Martinez, S., Paz-alonso, P.M., 2017. Retrospective head motion estimation in structural brain MRI with 3D CNNs. Medical image computing and computer-assisted intervention. In: MICCAI … International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 314–322.
Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: International Conference on Information Processing in Medical Imaging. Springer, pp. 597–609.
Kelly, C., Pietsch, M., Counsell, S., Tournier, J.-d, 2016. Transfer learning and convolutional neural net fusion for motion artefact detection. Proc. Intl. Soc. Mag. Reson. Med. 1–2.
Kingma, D., Ba, J., 2014. Adam: a Method for Stocha Stic Optimization arXiv preprint arXiv:1412.6980.
Leppert, I.R., Almli, C.R., McKinstry, R.C., Mulkern, R.V., Pierpaoli, C., Rivkin, M.J., Pike, G.B., 2009. T2 relaxometry of normal pediatric brain development. J. Magn. Reson. Imag. 29 (2), 258–267.
Liu, Z., Wang, Y., Gerig, G., Gouttard, S., Tao, R., Fletcher, T., Styner, M., 2010. Quality control of diffusion weighted images. In: Proceedings of SPIE–the International Society for Optical Engineering. In: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.844748.
Marami, B., Mohseni Salehi, S.S., Afacan, O., Scherrer, B., Rollins, C.K., Yang, E., Estroff, J.A., Warfield, S.K., Gholipour, A., 2017. Temporal slice registration and robust diffusion-tensor reconstruction for improved fetal brain structural connectivity analysis. Neuroimage 156 (April), 475–488.
Miller, K.L., Alfaro-Almagro, F., Bangerter, N.K., Thomas, D.L., Yacoub, E., Xu, J., Bartsch, A.J., Jbabdi, S., Sotiropoulos, S.N., Andersson, J.L., et al., 2016. Multimodal population brain imaging in the UK biobank prospective epidemiological study. Nat. Neurosci. 19 (11), 1523–1536.
Oguz, I., Farzinfar, M., Matsui, J., Budin, F., Liu, Z., Gerig, G., Johnson, H.J., Styner, M., Jan 2014. DTIPrep: quality control of diffusion-weighted images. Front. Neuroinf. 8 (January), 4 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3906573{&}tool=pmcentrez{&}rendertype=abstract.
Oubel, E., Koob, M., Studholme, C., Dietemann, J.L., Rousseau, F., 2012. Reconstruction of scattered data in fetal diffusion MRI. Med. Image Anal. 16 (1), 28–37. https://doi.org/10.1016/j.media.2011.04.004.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22 (10), 1345–1359.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2015. Rethinking the inception architecture for computer vision. arXiv preprint. http://arxiv.org/abs/1512.00567.
Tournier, J.-D., Mori, S., Leemans, A., 2011. Diffusion tensor imaging and beyond. Magn. Reson. Med. 65 (6), 1532–1556. http://doi.wiley.com/10.1002/mrm.22924.
Williams, L.-A., Gelman, N., Picot, P.A., Lee, D.S., Ewing, J.R., Han, V.K., Thompson, R.T., 2005. Neonatal brain: regional variability of in vivo mr imaging relaxation rates at 3.0 tinitial experience. Radiology 235 (2), 595–603.
Xanthis, C.G., Venetis, I.E., Chalkias, A., Aletras, A.H., 2014. Mrisimul: a gpu-based parallel approach to mri simulations. Medical Imaging. IEEE Transactions on 33 (3), 607–617.
Yendiki, A., Koldewyn, K., Kakunoori, S., Kanwisher, N., Fischl, B., 2014. Spurious group differences due to head motion in a diffusion mri study. Neuroimage 88, 79–90.