Retrospective head motion estimation in structural brain MRI with 3D CNNs

Juan Eugenio Iglesias[1,2], Garikoitz Lerma-Usabiaga[2], Luis C. Garcia-Peraza-Herrera[1], Sara Martinez[2], and Pedro M. Paz-Alonso[2]

[1] University College London, United Kingdom
[2] Basque Center on Cognition, Brain and Language (BCBL), Spain

Abstract. Head motion is one of the most important nuisance variables in neuroimaging, particularly in studies of clinical or special populations, such as children. However, the possibility of estimating motion in structural MRI is limited to a few specialized sites using advanced MRI acquisition techniques. Here we propose a supervised learning method to retrospectively estimate motion from plain MRI. Using sparsely labeled training data, we trained a 3D convolutional neural network to assess if voxels are corrupted by motion or not. The output of the network is a motion probability map, which we integrate across a region of interest (ROI) to obtain a scalar motion score. Using cross-validation on a dataset of $n = 48$ healthy children scanned at our center, and the cerebral cortex as ROI, we show that the proposed measure of motion explains away 37% of the variation in cortical thickness. We also show that the motion score is highly correlated with the results from human quality control of the scans. The proposed technique can not only be applied to current studies, but also opens up the possibility of reanalyzing large amounts of legacy datasets with motion into consideration: we applied the classifier trained on data from our center to the ABIDE dataset (autism), and managed to recover group differences that were confounded by motion.

1 Introduction

The negative impact of head motion on measurements derived from brain MRI has recently been a subject of study in the neuroimaging literature. In the context of functional connectivity studies, it has been shown that head motion has substantial, systematic effects on the timecourses of fMRI data, leading to variations in correlation estimates and functional coupling [1, 2]. In diffusion MRI, motion typically produces increased radial diffusivity estimates, while decreasing axial diffusivity and fractional anisotropy measures [3]. In morphometric studies with structural MRI, it has recently been shown that head motion decreases the estimates of cortical thickness and gray matter volumes [4]. Therefore, head motion is an important confounding factor that can undermine the conclusions of MRI-based neuroimaging studies. While motion certainly affects studies with a single healthy group, it is a particularly important factor in group studies involving clinical or special populations, such that one group is more prone to moving in the scanner than the other (e.g., Parkinson's).

To mitigate these problems, one would ideally use motion correction methods at acquisition. These techniques can be prospective or retrospective. The former attempt to dynamically keep the measurement coordinate system fixed with respect to the subject during acquisition. Head motion can be tracked with an external system (e.g., camera and markers [5]) or with image-based navigators [6, 7]. Retrospective methods attempt to correct for motion *after* the acquisition. Some retrospective algorithms exploit information from external trackers as well [8], while others use the raw k-space data [9]. Unfortunately, neither prospective motion correction nor external trackers are widely available yet. Moreover, there are immense amounts of legacy MRI data for which the raw k-space data are not available (since only reconstructed images are normally stored in the PACS), which limits the applicability of retrospective k-space techniques.

A simpler, more extended alternative to reconstructing motion-free images is to estimate a measure of motion, manually or automatically. The former is typically in the form of a quality control (QC) step, in which a human rater disregards scans that display motion artifacts. Despite its simplicity, manual QC is neither continuous nor reproducible, and can introduce bias in subsequent analyses. This problem can be ameliorated with automated techniques, which generate continuous, reproducible motion scores that can be used in two different ways: as automated QC and as nuisance factors. In automated QC, subjects with scores over a threshold are left out in a systematic and reproducible manner. When used as nuisance factors, scores are regressed out from the target variable to reduce the impact of motion on the analysis [3], so no subjects are discarded.

In functional and diffusion MRI, head motion can be estimated from the parameters of the transforms that co-register the different frames. In structural MRI, however, the absence of temporal information makes extracting measures of motion more difficult. Here we present a machine learning approach to retrospectively quantify motion from structural brain MRI. To the best of our knowledge, this is the first motion estimation method that relies solely on image intensities. Motion detection is cast as a supervised classification problem, which is solved with a convolutional neural network (CNN). We use a 3D network architecture (similar to 3D U-net [10]) with a nonlinear data augmentation scheme that enables learning with sparsely annotated MRI scans. This is a key feature in our application, since image regions corrupted by motion artifacts (e.g., ghosting, blurring) have ill-defined boundaries, and are difficult to manually delineate with precision – especially in 3D. We also model uncertainty in the CNN with dropout at testing [11], and a scalar motion score is produced by averaging the probability map estimated by the CNN across an application-dependent ROI.

Our technique requires no specialized equipment, and can be used to analyze both prospective and legacy MRI data. We evaluated the method with two datasets involving motion-prone populations (children and autism). Using an ROI including the cortical ribbon and an underlying layer of white matter, we show that our motion score is closely connected with cortical thickness (which is known to be sensitive to motion [4]), accurately predicts the results of human QC, and recovers group differences confounded by motion in a group study.

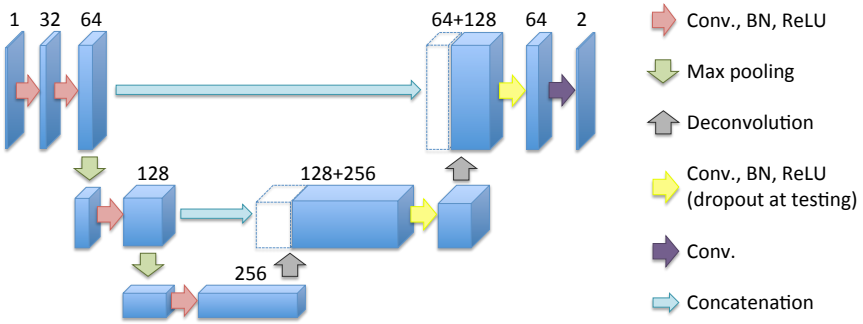

Fig. 1: CNN architecture. Conv. stands for convolution, BN for batch normalization, and ReLU for rectified linear unit. The number of feature maps is displayed above each layer.

2 Methods

2.1 Voxel classifier

The core of our method is a classifier that produces, for each voxel, an estimate of the probability that its intensity is corrupted by motion artifacts. As classifier, we use a 3D CNN based on the 3D U-net architecture [10], which is robust against sparsely labeled training data. Our architecture is shown in Figure 1. The network is leaner than in [10], since we do not need a large receptive field to detect motion artifacts, and also for faster training and inference.

The network has an analysis and synthesis stage with three levels of resolution. The input is a 64^3 voxel cube. At the analysis stage, the convolution layers have kernels of size 3×3×3 (stride 1), and are followed by rectified linear units (ReLU), batch normalization [12] and max pooling (2×2×2, stride 2). At the synthesis stage, deconvolutions (2×2×2, stride 2) are followed by a 3×3×3 convolutional layer and a ReLU. In testing, we also implement random dropout at these ReLUs, in order to obtain different samples of the approximate posterior distribution of the output [11]. Shortcut connections link layers of matching resolution at the analysis and synthesis stages, providing the latter with information at increasingly higher resolution at each level. In the last layer, a 1×1×1 convolution reduces the number of outputs to two, corresponding to motion and no motion. We used weighted cross-entropy as loss function, which makes it straightforward to train on sparsely labeled data, by setting the weight of unlabeled voxels to zero. The output is a 42^3 voxel tile, with a receptive field of size 22^3 voxels.

2.2 Computation of the measure of head motion

Following [13], we use an average probability within an ROI as global score:

$$M = \frac{1}{|\Omega_{ROI}|} \sum_{\boldsymbol{x} \in \Omega_{ROI}} p_m(\boldsymbol{x}) = \frac{1}{|\Omega_{ROI}|} \sum_{\boldsymbol{x} \in \Omega_{ROI}} \frac{\exp[m(\boldsymbol{x})]}{\exp[n(\boldsymbol{x})] + \exp[m(\boldsymbol{x})]}, \quad (1)$$

where M is our global motion score, Ω_{ROI} is the ROI domain, $\boldsymbol{x}$ is a voxel location, and $p_m(\boldsymbol{x})$ is the probability that the voxel at location $\boldsymbol{x}$ is motion corrupted. Such probability is computed as the softmax of $n(\boldsymbol{x})$ and $m(\boldsymbol{x})$, which are the strengths of the activations of the no-motion and motion units at the final layer of the CNN, respectively. As much as a single $p_m(\boldsymbol{x})$ is a weak measure of head motion, its average across the ROI provides a robust estimate [13].

3 Experiments and Results

3.1 MRI data and manual annotations

We used two different datasets in this study. The first dataset (henceforth the "in-house" dataset) consists of brain MRI scans from $n = 48$ healthy children aged 7.1-11.5 years, acquired with a 3T Siemens scanner using an MP-RAGE sequence at 1 mm isotropic resolution. Two separate sets of ground truth annotations were created for this dataset: at the scan level (for testing automatic QC) and at the voxel level (for training the CNN). At the scan level, we made two sets of QC annotations: one by a trained RA (SM), which we used as ground truth ($n_{\text{pass}} = 34$, $n_{\text{fail}} = 14$), and a second by JEI, with inter-rater variability purposes.

At the voxel level, creating dense segmentations is time consuming and hard to reproduce due to the difficulty of placing accurate boundaries around regions with motion artifacts, particularly in 3D. Instead, we made sparse annotations as follows. First, the RA went over the QC-passed scans, and identified slices in different orientations (axial / sagittal / coronal, approximately 30 per scan) that displayed no motion artifacts. The voxels inside the brain in these slices were all labeled as "no motion", whereas all other voxels in the scan were not used in training. Then, the RA went over the QC-failed scans, and drew brushstrokes on regions inside the brain that clearly showed motion artifacts, making sure that the annotations were highly specific. These voxels were labeled as "motion", whereas the remaining voxels were not used to train the classifier. The process took approximately 10-15 minutes per scan.

In order to test our classifier in a practical scenario and assess its generalization ability, we used a second dataset: the Autism Brain Imaging Data Exchange (ABIDE [14]). Even though effect sizes are notoriously small in autism spectrum disorder (ASD), ABIDE is a representative example of the type of application for which our method can be useful, since children with ASD might be more prone to moving in the scanner. We used a subset of ABIDE consisting of the $n = 111$ subjects (68 controls, 47 ASD) younger than 12 years (range: $10 - 12$). This choice was motivated by: 1. staying in the age range in which children with ASD still have increased cortical thickness [15, 16]; and 2. matching the population with that of the in-house dataset. This subset of ABIDE was acquired on nine different scanners across different sites, mostly with MP-RAGE sequences at 1 mm resolution (see [14]).

In both datasets, image intensities were coarsely normalized by dividing them by their robust maximum, computed as the 98^{th} percentile of their intensity distribution. Cortical thickness measures were obtained with FreeSurfer [17].

3.2 Experimental setup

The motion metric from Equation 1 was computed for the scans from both datasets as follows. For the in-house dataset, we used cross-validation with just two pseudorandom folds (since training the CNN is computationally expensive), ensuring that the number of QC-fails was the same in both. For ABIDE, rather than retraining the CCN on the whole in-house dataset, we processed the scans with the two CNNs that were already trained and averaged their outputs.

The 3D CNNs were trained end-to-end from scratch using a modified version of their publicly available implementation, which is based on the Caffe framework [18]. Data augmentation included: translations; linear mapping of image intensities (slope between 0.8 and 1.2); rotations (up to 15 degrees around each axis); and elastic deformations based on random shifts of controls point and B-spline interpolation (control points 16 voxels apart, random shifts with standard deviation of 2 voxels). Stochastic gradient descent was used to minimize the weighted cross-entropy. We used different (constant) weights for the positive and negative samples to balance their contributions to the loss function. We trained until the cross-entropy flattened for the training data (i.e., no validation set), which happened at 60,000 iterations (approximately 10 hours on a Nvidia Titan X GPU). In testing, we used a 50% overlap of the input tiles to mitigate boundary artifacts. Further smoothness was achieved by the dropout at testing scheme [11] (probability: 0.5), which also increased the richness in the distribution of output probabilities. The final probability of motion for each voxel was computed as the average of the available estimates at each spatial location.

We evaluated our proposed approach both directly and indirectly. For direct validation, we assessed the ability of the motion score to predict the output of human QC of the in-house dataset. For the indirect validation, we examined the relationship between our motion score and average cortical thickness, as well as the ability of the score to enhance group differences when regressed out. To compute the motion score, we used an ROI (Ω_{ROI}) comprising the cortical ribbon (as estimated by FreeSurfer) and an underlying 3 mm layer of cerebral white matter, computed by inwards dilation with a spherical kernel.

3.3 Results

Qualitative results: Figure 2 shows sagittal slices of four sample MRI scans with increasingly severe artifacts, along with the corresponding outputs from the CNN: (a) is crisp and motion-free, and few voxels produce high probability of motion; (b) shows minimal ringing on the superior and frontal regions; (c) shows moderate motion; and (d) displays severe blurring and ringing due to motion, such that the CNN produces high probabilities around most of the ROI.

Quantitative results on in-house dataset: Figure 3(a) shows the distributions of the motion scores for the two QC groups, which are far apart: a non-parametric test (Wilcoxon signed-rank) yields $p = 5 \times 10^{-8}$. Therefore, a classifier based on thresholding the score can closely mimic human QC, reaching 0.916 accuracy and 0.941 area under the receiver operating characteristic (ROC)

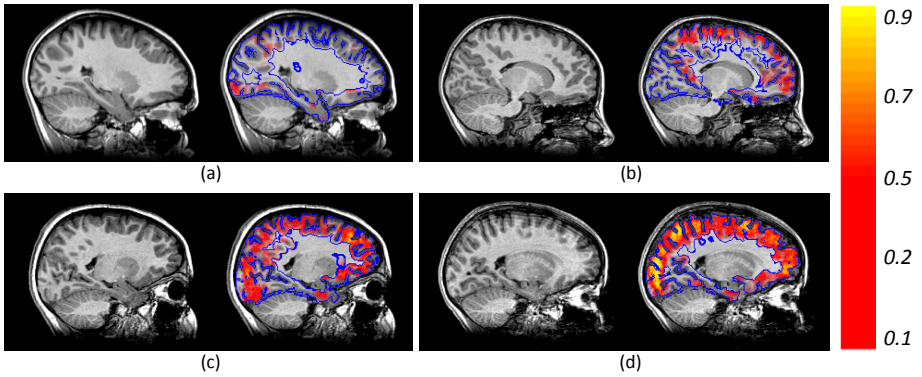

Fig. 2: Sagittal slices of four cases and corresponding probability maps (masked by the ROI, outlined in blue). (a) $M = 0.12$ (lowest in dataset). (b) $M = 0.19$. (c) $M = 0.25$. (d) $M = 0.32$ (failed QC).

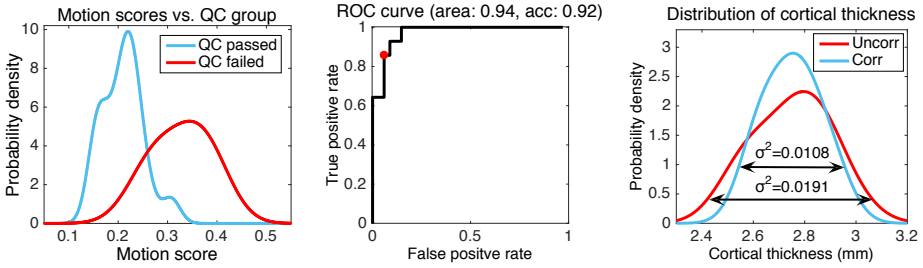

Fig. 3: (a) Distribution of motion scores for the two QC groups. (b) ROC for automatic QC based on score thresholding; the dot marks the operating point: 91.6% accuracy. c) Distribution of cortical thickness with and without correction.

curve; see Figure 3(b). This performance is close to the inter-rater variability, which was 0.958. We also found a strong negative correlation between our score and mean cortical thickness: $\rho = 0.66$ (95% C.I. [-0.79,-0.46], $p = 3 \times 10^{-7}$). When correcting for motion, the variance of the cortical thickness decreased from 0.0191 mm^2 to 0.0108 mm^2, i.e., by 37% ($R^2_{adj} = 0.42$); see Figure 3(c).

Results on ABIDE dataset: Using a Wilcoxon signed-rank test, we found differences in motion scores between the two groups ($p = 0.03$), a circumstance that can undermine the conclusion of cortical thickness comparisons. We built a general linear model for the left-right averaged mean thickness of each FreeSurfer cortical region, with the following covariates: age, gender, group, site of acquisition and, optionally, our motion score. Introducing motion as a covariate in the model changed the results considerably, as shown by the significance maps in Figure 4, which are overlaid on an inflated, reference surface space ("fsaverage").

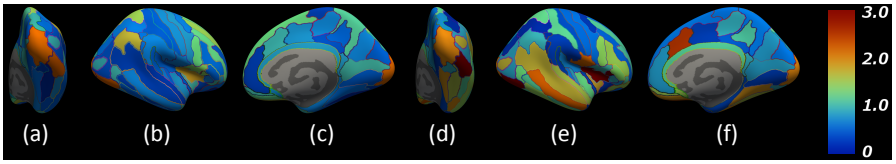

Fig. 4: Region-wise significance map for differences in cortical thickness between ASD and control group (left-right averaged). The color map represents $-\log_{10} p$. (a) Inferior-posterior view, model without motion. (b) Lateral view, model without motion. (c) Medial view, model without motion. (d-f) Model with motion.

Figure 4(a,d) shows an inferior-posterior view exposing the occipital lobe and lingual gyrus, areas in which increased cortical thickness has been reported in children with ASD [16]. The motion-corrected model increases the effect size in the occipital lobe (particularly the inferior region) and detects differences in the lingual gyrus that were missed by the model without motion – possibly because the effect of motion was very strong in this region ($p = 5 \times 10^{-7}$ for its slope).

Figure 4(b,e) shows a lateral view, in which correction by motion reveals effects in the temporal lobe and the insula, which would have been otherwise missed. The thicknesses of both of these regions showed a strong association with our motion score: $p = 5 \times 10^{-9}$ and $p = 2 \times 10^{-8}$, respectively. Finally, the model with motion also detected missed differences in the mid-anterior cingulate cortex, as shown in the medial view in Figure 4(c,f) (effect of motion: $p = 3 \times 10^{-8}$).

4 Discussion

This work constitutes a relevant first step to retrospectively estimate in-scanner motion from structural MRI scans, without requiring external trackers or raw k-space data. The technique not only enables sites without means for specialized MRI acquisition to consider motion, but also makes it possible to reanalyze legacy datasets correcting for motion, which can considerably change the results – as we have shown on ABIDE, without even fine-tuning our CNN to this dataset.

Our method is specific to population and MRI contrast. However, once a CNN has been trained, accurate motion estimates can be automatically obtained with the method for all subsequent scans within a center, with some generalization ability to other datasets. Training datasets for other MRI contrasts can be created with limited effort (ca. 10 hours), since training relies on sparely labeled data. Moreover, manual labeling effort could in principle be saved by fine-tuning our CNN to a new dataset, using only a handful of (sparsely) annotated scans.

Future work will follow three directions: 1. Fine-tuning the CNN to other datasets; 2. Testing the method on other morphometric measures and ROIs (e.g., hippocampal volume); and 3. Extension to motion *correction*, by training on a (possibly synthetic) set of matched motion-free and motion-corrupted scans.

Acknowledgement: This research was supported by the European Research Council (Starting Grant 677697, project BUNGEE-TOOLS).

References

1. Van Dijk, K.R., Sabuncu, M.R., Buckner, R.L.: The influence of head motion on intrinsic functional connectivity MRI. Neuroimage **59**(1) (2012) 431–438
2. Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., Petersen, S.E.: Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage **59**(3) (2012) 2142–2154
3. Yendiki, A., Koldewyn, K., Kakunoori, S., Kanwisher, N., Fischl, B.: Spurious group differences due to head motion in a diffusion MRI study. Neuroimage **88** (2014) 79–90
4. Reuter, M., Tisdall, M.D., Qureshi, A., Buckner, R.L., van der Kouwe, A.J., Fischl, B.: Head motion during MRI acquisition reduces gray matter volume and thickness estimates. Neuroimage **107** (2015) 107–115
5. Maclaren, J., Armstrong, B.S., Barrows, R.T., Danishad, K., Ernst, T., Foster, C.L., Gumus, K., et al.: Measurement and correction of microscopic head motion during magnetic resonance imaging of the brain. PLOS one **7**(11) (2012) e48088
6. White, N., Roddey, C., Shankaranarayanan, A., Han, E., Rettmann, D., Santos, J., Kuperman, J., Dale, A.: PROMO: Real-time prospective motion correction in MRI using image-based tracking. Magnetic Resonance in Medicine **63** (2010) 91
7. Tisdall, D., Hess, A., Reuter, M., Meintjes, E., Fischl, B., van der Kouwe, A.: Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI. Magnetic resonance in medicine **68**(2) (2012) 389–399
8. Glover, G.H., Li, T.Q., Ress, D.: Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magnetic resonance in medicine **44**(1) (2000) 162–167
9. Batchelor, P., Atkinson, D., Irarrazaval, P., Hill, D., Hajnal, J., Larkman, D.: Matrix description of general motion correction applied to multishot images. Magnetic resonance in medicine **54**(5) (2005) 1273–1280
10. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-net: learning dense volumetric segmentation from sparse annotation. In: Lecture Notes in Computer Science. Volume 9901. (2016) 424–432
11. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv preprint:1506.02142 (2015)
12. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint:1502.03167 (2015)
13. Coupé, P., Eskildsen, S.F., Manjón, J.V., Fonov, V.S., Collins, D.L.: Simultaneous segmentation and grading of anatomical structures for patient's classification: application to Alzheimer's disease. NeuroImage **59**(4) (2012) 3736–3747
14. Di Martino, A., Yan, C.G., Li, Q., Denio, E., Castellanos, F.X., Alaerts, K., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry **19**(6) (2014) 659–667
15. Wallace, G.L., Dankner, N., Kenworthy, L., Giedd, J.N., Martin, A.: Age-related temporal and parietal cortical thinning in autism spectrum disorders. Brain (2010) 3745–3754
16. Zielinski, B.A., Prigge, M.B., Nielsen, J.A., Froehlich, A.L., Abildskov, T.J., Anderson, J.S., Fletcher, P.T., Zygmunt, K.M., et al.: Longitudinal changes in cortical thickness in autism and typical development. Brain **137**(6) (2014) 1799–1812
17. Fischl, B.: Freesurfer. Neuroimage **62**(2) (2012) 774–781
18. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: 22$^{\text{nd}}$ ACM international conference on Multimedia. (2014) 675–678