ESTIMATING HEAD MOTION FROM MR-IMAGES

Clemens Pollak[1,*], David Kügler[1,*], and Martin Reuter[1,2,3]

[1]AI in Medical Imaging, German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany
[2]A.A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA, USA
[3]Department of Radiology, Harvard Medical School, Boston, MA, USA

ABSTRACT

Head motion is an omnipresent confounder of magnetic resonance image (MRI) analyses as it systematically affects morphometric measurements, even when visual quality control is performed. In order to estimate subtle head motion, that remains undetected by experts, we introduce a deep learning method to predict in-scanner head motion directly from T1-weighted (T1w), T2-weighted (T2w) and fluid-attenuated inversion recovery (FLAIR) images using motion estimates from an in-scanner depth camera as ground truth. Since we work with data from compliant healthy participants of the Rhineland Study, head motion and resulting imaging artifacts are less prevalent than in most clinical cohorts and more difficult to detect. Our method demonstrates improved performance compared to state-of-the-art motion estimation methods and can quantify drift and respiration movement independently. Finally, on unseen data, our predictions preserve the known, significant correlation with age.

Keywords Motion estimation · MRI quality · Deep learning · Motion tracking

Corresponding author: Martin Reuter (`martin.reuter[at]dzne.de`)

1 Introduction

Head motion is a ubiquitous challenge for magnetic resonance image (MRI) acquisition. It causes a range of image artifacts that introduce bias in downstream analysis [1–6], which persists despite expert quality control [1, 2]. While initially explored for clinical cohorts with increased motion levels [5–8] or induced motion [1, 9], less research focuses on motion in studies of healthy, compliant population cohorts [2, 10] such as the Rhineland Study [11, 12]. Critically, the lack of a sensitive and reliable motion estimation method to quantify subtle motion hinders the inclusion of motion estimates in statistical models to control motion-induced biases. For example, careful visual inspection of the Rhineland Study dataset, used in this paper, did not detect any cases with clearly visible motion artefacts that would warrant exclusion. Yet, even in the 75-participant subset reserved for testing, a statistically significant correlation of motion with age can be shown, underlining the need for sensitive estimation and control of head motion in MRI analyses.

In this paper, we propose a method to directly estimate head motion from the acquired MR image. We measure head motion during MRI acquisition via head tracking with a depth camera and establish a ground truth motion score per sequence. This is contrary to the currently established paradigm of predicting discrete motion severity levels established via an expert manual quality control process [8–10, 13–18]. Expert ratings are limited by their subjectivity to

[*]These authors contributed equally to this work

the specific task and human perception [18–20] hindering their general utility, specifically for compliant, low-motion cohorts. Camera-based motion measurements, on the contrary, are objective and sensitive even for low-motion cohorts, where motion-induced image artifacts are almost invisible.

Since the rise of deep learning, tools have been able to predict expert motion ratings with increasing accuracy [8–10, 13–18] sometimes addressing previous limitations, for example the subjectivity to the task [18]. Currently, the only alternative to the prediction of expert labels is to predict a perceptual image similarity metric between low motion "baseline" images and high-motion images, which are retrospectively simulated [21]. This approach replaces the human annotation task by a comparison of images with a perceptual similarity metric (SSIM), which may also suffer from similar limitations as expert labels. Moreover, the methods accuracy on real-world data relies on realistic, high-quality simulation, which also has to be adapted to each acquisition sequence. Meanwhile, our method can be directly retrained even on different modalities, without any changes. Recent work in the field of in-MRI motion tracking enabled highly accurate tracking of head-motion during acquisition with the MRI scanner [22, 23], optical cameras [24, 25] or other devices [26, 27]. Yet, until tracking devices and methods are deployed to all imaging sites, image-derived motion estimates may help reduce potential motion induced biases – even retrospectively.

Our contributions are threefold – we 1. introduce – for the first time – the estimation of an objective motion score from images of three MRI sequences, 2. present a deep-learning-based solution, which outperforms DenseNet and state-of-the-art quality/motion estimation methods on a dataset of compliant, low-motion participants, and 3. quantify motion from respiration and (relaxation) drift. Finally, our method detects the significant, known correlation between predicted motion and age. We will publish our code on GitHub*.

2 Materials & Methods

2.1 Data acquisition

The Rhineland study is an ongoing population study recruiting a representative cohort of healthy participants above the age of 30. Our dataset describes a subset (ages 30 to 95 years) of 500 participants (282 female) with T1w, T2w and FLAIR images (at 0.8, 0.8 and 1.0 mm isotropic voxel size, respectively) following a standardized acquisition protocol*. This dataset also includes expert quality-labels of T1w images (*PASS*, *WARN* or *FAIL*), where *WARN* indicates visible and *FAIL* strong artifacts (insufficient for downstream analysis). However, no images are rated as *FAIL* and only 9 images as *WARN*. The low number of *WARN* and *FAIL* cases is likely founded on high compliance and extensive efforts to reduce head motion during MRI acquisition including tight head padding, scheduled speaking breaks, careful participant instruction, and calming nature scenes shown during the scan.

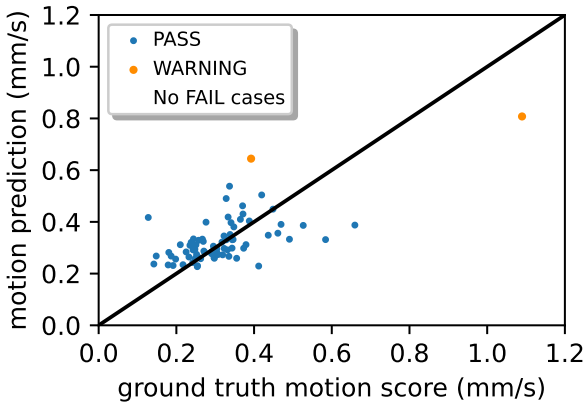

Figure 1: Plot of ground truth and predicted motion score. Blue *PASS* and orange *WARN* (visible artifacts) test images may be perfectly separated with a threshold of 0.6.

To quantify the participant's head motion, a video of depth images showing a portion of their face is collected concurrently to the scan [24]. Individual frames are aligned with a reference frame resulting in time series of rigid transformations. To obtain a per-sequence, scalar *motion score*, we 1. synchronize MRI and depth camera, 2. compute Jenkinson's transformation differences [29] per pair of transformations, and 3. extract and average values for the duration of individual sequences. The Jenkinson's transformation difference summarizes rigid transformations by averaging displacements within a spherical head model. The final *motion score* quantifies the average motion in millimeter per second. We randomly split the dataset into training, validation, and evaluation sets of 350, 75 and 75 participants, respectively.

2.2 Distinguishing motion patterns

In addition to the per-sequence motion score, we distinguish between three prominent types of in-scanner head motion

*https://github.com/Deep-MI/head-motion-from-MRI/

*For details on the acquisition of 3D T1w MPRAGE (scan length 6.5 min), 3D T2w (4.6 min) and FLAIR (4.5 min) images, see Lohner et al. [28].

corresponding to three frequency bands: i) head drift, ii) periodic motion due to breathing, and iii) "noisy motion", which we expect is hard to estimate. To determine global filter thresholds, we estimate upper and lower median respiratory frequency of participants during the T1w sequence from an independent respiration sensor as 0.1 Hz and 0.5 Hz. We apply symmetric Butterworth filters (low-pass, band-pass and high-pass, respectively) to the time series of Jenkinson's transformation differences. From the filtered signals, we aggregate the motion score as before resulting in three frequency-dependent targets.

2.3 Neural network & training

For the estimation of motion scores from 3D MR images, we adopt a fully convolutional neural network (CNN) from brain age estimation [30], an established regression task in medical imaging. The lightweight architecture permits training the 3D CNN on a single NVIDIA A100 GPU with batch size two. Instead of directly regressing the motion score, we follow Peng et al. [30] in their approach and 1. generously define the expected range of motion [0, 3.12] mm/s, 2. split it into 40 bins of 'prototypes', 3. for each prototype calculate the probability that the current motion score belongs into the prototype, and 4. train the CNN using a Kullback-Leibler loss (Adam optimizer for 500 epochs, approx. $10\,\mathrm{h}$ on one A100 GPU). To reconstruct the motion score from the predicted probability distribution, we sum the product of prototype centers and predictions. Many standard data augmentation strategies are not suitable for motion estimation, since the re-sampling of images affects the image noise, which is why we avoid interpolation of images completely. Helpful data augmentation, on the other hand, include intensity scaling in the range [0.9,1.1] and random flipping with 30% probability along all axes. In our ablation study (Section 3.3), we explore different pre-processing operations. A focus on the 8 Least Significant Bits (LSB8) is useful for this task, leaving only an integer representation of the fine image differences.

2.4 Evaluation & statistical methods

We evaluate the regression model with the coefficient of determination (R^2 score – a measure for the average error) and Spearman's rank correlation coefficient (Spearman's ρ – a measure for correct ranking). The R^2 score normalizes the mean squared error to a range of $[-\infty, 1]$, where score <0 indicates a prediction error worse than a constant prediction of the dataset mean and a score of 1 indicates perfect predictions. The Spearman's ρ, on the other hand, is defined in the range $[0, 1]$. It is not affected by large, absolute errors of outliers and more sensitive to prediction errors, where the sampling of values is denser (i.e. more sensitive to small errors on values close together). We also analyze the rank correlation between motion and age using this method. We use the R^2 score as the primary metric for ablation, and select parameters that have the highest R^2 score on the validation set in experiments.

	Method	R^2	Spr-ρ
re-trained	Ours	**0.433**	**0.584**
	DenseNet [31]	0.395	0.447
	SFCN [30]	0.275	0.454
	MIQA CNN [8]	-0.273	0.192
QCtools	MIQA MA / QS	-	0.243 / 0.240
	MIQA MA / QS[1]	-	-[2] / 0.338
	AES [7]	-	0.110

Table 1: Our methods outperforms SOTA approaches in both Spearman's ρ (Spr-ρ) and R^2 scores (only valid for predicted motion scores) when predicting motion scores from T1w images on the test set. MA: motion artefact score, QS: quality score, [1]images standardized, [2]no motion detected

3 Results

We visualize the performance of our method on the unseen test set in Figure 1, which illustrates good correlation between ground truth measured motion (horizontal axis) and predictions from images (vertical). Perfect predictions would lie on the black line. Our method perfectly separates *PASS* (no artifacts) and *WARN* (mild artifacts) cases. A horizontal separation line at $\approx 0.6\,\mathrm{mm/s}$ can be found but no vertical line for the ground truth motion score.

3.1 Comparison with state-of-the-art motion estimation

To the best of our knowledge, there is currently no competing method to predict measured, in-scanner head motion from MR images. Since quality estimation methods cannot be easily re-trained on our dataset, which has few *WARN* and no *FAIL* labels, we compare with the pre-trained MIQA quality estimator [8] and Average Edge Strength (AES) [7], a heuristic known to correlate with motion [7]. Additionally, we compare our method with three deep learning architectures re-trained on our dataset: i) DenseNet [31], ii) SFCN [30], a CNN for brain age prediction, and iii) the CNN used by MIQA [8]. Our method, which is an improvement of SFCN (e.g. initializer), achieves the best correlation

Input	Target	R^2	Spr-ρ
T1	motion score	0.433	**0.584**
T2	motion score	0.362	0.556
FLAIR	motion score	0.299	0.489
T1	drift	0.183	**0.637**
T1	breathing band	0.185	0.382
T1	noisy motion	0.050	0.337

Table 2: The evaluations show generalization of our method to predicting the motion score on T2w and FLAIR images and very promising Spr-ρ performance for detection of drift on the test set.

Method	Images	R^2	Spr-ρ
CNN MSE loss	T1 (LSB8)	0.117	0.376
CNN ours	T1 (unprocessed)	0.341	0.551
CNN ours	T1 (robust scaling)	0.322	0.489
CNN ours	T1 (remove head)	0.369	0.525
CNN ours	T1 (LSB8)	**0.393**	0.491

Table 3: Ablation of loss function and image pre-processing on the validation set. LSB8: 8 Least Significant Bits.

with ground truth motion scores on the unseen test set (Table 1). The negative R^2 of the re-trained MIQA CNN indicates failed generalization.

The pre-trained MIQA tool aggregates predictions of expert ratings for 9 artifact types including a 'motion artifacts' (MA) score into a continuous quality score (QS). We select their probabilities as potential correlates of our motion score. Since MIQA has been trained on the $1\,\mathrm{mm}$ T1w-PREDICT-HD dataset, we test, whether rescaling and resampling images with FastSurfer's [32] conform tool to $1\,\mathrm{mm}$ reduces domain shift. We measure a low but significant correlation between QS and motion score as well as between MA and motion score on the native $0.8\,\mathrm{mm}$ images. While previous work reported that probabilities of ratings, like those of MIQA, quantify subtle differences in motion, despite binary ground truth labels [16], MA probabilities are zero for all conformed images (consistent with the lack of manual *FAIL* labels). QS, on the other hand, significantly correlates with the motion score. In addition to the deep learning estimators, we evaluate the performance of AES [7], but do not find significant correlation between AES and the ground truth motion score.

3.2 Generalization to T2/FLAIR and motion types

We test the generalizability of our method to new tasks by predicting the motion score for T2w and FLAIR images, as well as predicting drift and respiratory motion on T1w images. For each task we re-train the network architecture and show the results in Table 2. While the R2 score is not directly comparable across different tasks, we find good performance by purely re-training the model for these modalities. The Spearman's-ρ in T2w and FLAIR experiments is similar to T1w experiments despite optimization on T1w only.

To quantify different motion types, we define two distinct aggregates of participant motion: i) slow relaxation-drift over time, and ii) periodic head motion due to breathing. We filter motion estimates as described in Section 2.2 and train our architecture to predict the aggregates (Table 2). Our method can predict both slow drift (low frequency) motion and respiratory (medium frequency) motion from the T1w images. The aggregate of high frequencies, which are not associated with known motion types (noisy motion), cannot be predicted.

3.3 Ablation study

To optimize the parameters for our method, we compare choices for loss and preprocessing on the validation set. A critical finding of this work is that direct prediction trained using an Mean Squared Error (MSE) loss only achieves mild correlation with the motion score, while using a multi-dimensional probability distribution together with a Kullback-Leibler divergence loss yields a large performance uplift. For this MSE-loss ablation in Table 3 (top), we reduce the last layer to a single output and remove the softmax.

Several measures of image quality have taken advantage of the background signal to determine a reduction in image quality [33, 34]. Consequently, we explore the effects of four image pre-processing operations on the prediction quality in Table 3: No pre-processing, FastSurfer's [32] robust scaling (removing 20% of high intesity voxels), removing the head with FreeSurfer's head segmentation tool [35, 36] (the result is just the background) and dropping the Most Significant Bits leaving us with the 8 Least Significant Bits (LSB8). The latter approach, which drops the information about the absolute size of values directly on the images integer representation, surprisingly outperforms other ways of adjusting image intensities. This finding was validated in multiple experiments and across additional hypotheses.

3.4 Correlation with age

In adult populations, increased head motion in the MR scanner is associated with the increased age of participants [4,

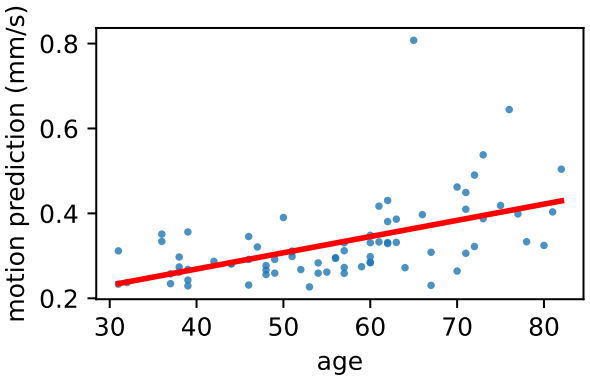

Figure 2: Age is significantly correlated with our motion predictions. This reflects the ground truth motion score labels (not shown) and is outlined by a linear fit, displayed as a red line.

37]. We can also measure this correlation within the ground truth motion score using Spearman's ρ and find a significant correlation with age on the whole dataset as well as only on the test set ($p < 0.001$). Our predictions also present this correlation with age on the test set ($p < 0.001$) as illustrated in Figure 2 with the linear fit.

4 Discussion & conclusion

We introduce a novel task of in-scanner head motion estimation directly from MR images. The presented deep learning method provides sensitive predictions of motion levels capturing subtle correlations with known confounders such as age – all in unseen images that pass visual inspection. Two factors contribute to this sensitivity: predicting the probability distribution of 'prototypes' together with the Kullback-Leibler divergence loss, and the input of least-significant-bit (LSB) images. Why LSB images are preferable to raw structural images or their background should be further investigated in future work.

A well-known limitation of deep learning methods is the limited generalizability to unseen datasets. Large differences in motion levels between cohorts and the chosen acquisition parameters greatly affect the appearance of motion artifacts, hence we expect dedicated training datasets will be required to for re-training and generalization to unseen MR imaging sequences. Additionally, future work should investigate, whether motion estimation itself is also affected by biases, like age and diseases, which are known to be an indicator of increased motion levels.

Our method ranks images by their motion level better than comparable, state-of-the-art methods for MRI motion and quality estimation from expert ratings. Therefore, it may aid quality control procedures in the identification and exclusion of cases with artifacts, which is also indicated by the clear separation between *PASS* and *WARN* labels in our experiments (Figure 1). However, confirmation on a dataset with more strongly motion-affected cases is required. Additionally, our method transfers well to the prediction of alternative targets for respiratory- and drift motion and from T2w and FLAIR images with comparable accuracy.

Finally, our method enables an analysis of other, perhaps unknown, correlates of motion as well as the integration of motion scores as a control variable in statistical models. This is particularly valuable in longitudinal cohort studies, like the Rhineland Study, to disentangle the bias of motion effects from other effects such as participant's age and diseases.

5 Compliance with ethical standards

The Rhineland Study is carried out in accordance with the recommendations of the ICH-GCP standards. Written informed consent was obtained from all participants in accordance with the Declaration of Helsinki. Approval was granted by the Ethics Committee of University Bonn.

6 Acknowledgments

This work was supported by DZNE institutional funds, by the Federal Ministry of Education and Research of Germany (031L0206), and the Helmholtz-AI project DeGen (ZT-I-PF-5-078). We thank the Rhineland Study group (PI Monique Breteler) for supporting the data acquisition and management. The authors do not have any conflict of interest.

References

[1] M. Reuter et al. "Head motion during MRI acquisition reduces gray matter volume and thickness estimates". In: *Neuroimage* 107 (2015), pp. 107–115.

[2] A. Alexander-Bloch et al. "Subtle in-scanner motion biases automated measurement of brain anatomy from in vivo MRI". In: *HBM* 37.7 (2016), pp. 2385–2397.

[3] A. D. Gilmore, N. J. Buser, and J. L. Hanson. "Variations in structural MRI quality significantly impact commonly used measures of brain anatomy". In: *Brain informatics* 8.1 (2021), pp. 1–15.

[4] N. K. Savalia et al. "Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion". In: *HBM* 38.1 (2017), pp. 472–492.

[5] H. R. Pardoe, R. K. Hiess, and R. Kuzniecky. "Motion and morphometry in clinical and nonclinical populations". In: *Neuroimage* 135 (2016), pp. 177–185.

[6] A. F. Rosen et al. "Quantitative assessment of structural image quality". In: *Neuroimage* 169 (2018), pp. 407–418.

[7] Domenico Zaca et al. "Method for retrospective estimation of natural head movement during structural MRI". In: *Journal of Magnetic Resonance Imaging* 48.4 (2018), pp. 927–937.

[8] D. Zukić et al. "Medical Image Quality Assurance using Deep Learning". In: *MIDL*. 2022.

[9] T. Küstner et al. "Automatic motion artifact detection for Whole-Body magnetic resonance imaging". In: *ICASSP*. IEEE. 2018, pp. 995–999.

[10] I. Fantini et al. "Automatic detection of motion artifacts on MRI using Deep CNN". In: *PRNI*. IEEE. 2018, pp. 1–4.

[11] M. Breteler et al. "Mri in the rhineland study: a novel protocol for population neuroimaging." In: *Alzheimer's Dement.* 10 (2014), p. 92.

[12] T. Stöcker. "Big data: the Rhineland study". In: *ISMRM*. 2016.

[13] A. Largent, K. Kapse, S. D. Barnett, et al. "Image quality assessment of fetal brain MRI using multi-instance deep learning methods". In: *JMRI* 54.3 (2021), pp. 818–829.

[14] S. J. Sujit et al. "Automated image quality evaluation of structural brain MRI using an ensemble of deep learning networks". In: *JMRI* 50.4 (2019), pp. 1260–1267.

[15] J. J. Ma et al. "Diagnostic image quality assessment and classification in medical imaging: Opportunities and challenges". In: *ISBI*. IEEE. 2020, pp. 337–340.

[16] T. Küstner et al. "Automated reference-free detection of motion artifacts in magnetic resonance images". In: *MAGMA* 31.2 (2018), pp. 243–256.

[17] I. Stpień et al. "Fusion of Deep Convolutional Neural Networks for No-Reference Magnetic Resonance Image Quality Assessment". In: *Sensors* 21.4 (2021), p. 1043.

[18] K. Lei et al. "Artifact-and content-specific quality assessment for MRI with image rulers". In: *Medical Image Analysis* 77 (2022), p. 102344.

[19] H. H. Barrett et al. "Task-based measures of image quality and their relation to radiation dose and patient risk". In: *Physics in Medicine & Biology* 60.2 (2015), R1.

[20] M. Pryde et al. "The Performance of Image Quality Metrics Depends on the Diagnostic Task: A Case Study in Stroke MRI". In: *ISMRM* (2022).

[21] A. Sciarra et al. "Reference-less SSIM Regression for Detection and Quantification of Motion Artefacts in Brain MRIs". In: *MIDL*. 2022.

[22] M. D. Tisdall et al. "Volumetric navigators for prospective motion correction and selective reacquisition in neuroanatomical MRI". In: *Magn. Reson. Med.* 68.2 (2012), pp. 389–399.

[23] O. C. Andronesi, P. K. Bhattacharyya, W. Bogner, et al. "Motion correction methods for MRS: experts' consensus recommendations". In: *NMR in Biomedicine* 34.5 (2021), e4364.

[24] J. M. Slipsager, A. H. Ellegaard, S. L. Glimberg, et al. "Markerless motion tracking and correction for PET, MRI, and simultaneous PET/MRI". In: *Plos one* 14.4 (2019), e0215524.

[25] H. R. Pardoe et al. "Estimation of in-scanner head pose changes during structural MRI using a convolutional neural network trained on eye tracker video". In: *Magnetic Resonance Imaging* 81 (2021), pp. 101–108.

[26] M. Musa, S. Sengupta, and Y. Chen. "MRI-Compatible Soft Robotic Sensing Pad for Head Motion Detection". In: *IEEE Robotics and Automation Letters* 7.2 (2022), pp. 3632–3639.

[27] M. Laustsen et al. "Tracking of rigid head motion during MRI using an EEG system". In: *Magn. Reson. Med.* (2022).

[28] V. Lohner et al. "Incidental findings on 3 T neuroimaging: cross-sectional observations from the population-based Rhineland Study". In: *Neuroradiology* 64.3 (2022), pp. 503–512.

[29] M. Jenkinson et al. "Measuring transformation error by RMS deviation". In: *Technical report* (1999).

[30] H. Peng et al. "Accurate brain age prediction with lightweight deep neural networks". In: *Medical image analysis* 68 (2021), p. 101871.

[31] G. Huang et al. "Densely connected convolutional networks". In: *CVPR*. 2017, pp. 4700–4708.

[32] L. Henschel et al. "FastSurfer". In: *Bildverarbeitung für die Medizin 2020*. Springer, 2020, pp. 208–208.

[33] O. Esteban et al. “MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites”. In: *PloS one* 12.9 (2017), e0184661.

[34] B. Mortamet et al. “Automatic quality assessment in structural brain magnetic resonance imaging”. In: *Magn. Reson. Med.* 62.2 (2009), pp. 365–372.

[35] B. Fischl. “FreeSurfer”. In: *Neuroimage* 62.2 (2012), pp. 774–781.

[36] F. Ségonne et al. “A hybrid approach to the skull stripping problem in MRI”. In: *Neuroimage* 22.3 (2004), pp. 1060–1075.

[37] C. R. Madan. “Age differences in head motion and estimates of cortical morphology”. In: *PeerJ* 6 (2018).