Published in final edited form as:

Magn Reson Imaging. 2013 June ; 31(5): 707–717. doi:10.1016/j.mri.2012.11.007.

A quality control method for detecting and suppressing uncorrected residual motion in fMRI studies

Anthony G. Christodoulou^{a,b,*}, Thomas E. Bauer^{a,e}, Kent A. Kiehl^{a,c}, Sarah W. Feldstein Ewing^{a,f,e}, Angela D. Bryan^{a,c,d}, and Vince D. Calhoun^{a,e}

^aThe Mind Research Network, Albuquerque, New Mexico

^bUniversity of Illinois at Urbana-Champaign, Department of Electrical and Computer Engineering, Urbana, Illinois

^cUniversity of New Mexico, Department of Psychology, Albuquerque, New Mexico

^dUniversity of New Mexico, Center on Alcoholism, Substance Abuse and Addiction, Albuquerque, New Mexico

^eUniversity of New Mexico, Department of Electrical and Computer Engineering, Albuquerque, New Mexico

^fUniversity of New Mexico, University Honors Program, Albuquerque, New Mexico

Abstract

Motion correction is an important step in the functional magnetic resonance imaging (fMRI) analysis pipeline. While many studies simply exclude subjects who are estimated to have moved beyond an arbitrary threshold, there exists no objective method for determining an appropriate threshold. Furthermore, any criterion based only upon motion estimation ignores the potential for proper realignment. The method proposed here uses unsupervised learning (specifically k-means clustering) on features derived from the mean square derivative (MSD) of the signal before and after realignment to identify problem data. These classifications are refined through analysis of correlation between subject activation maps and the mean activation map, as well as the relationship between tasking and motion as measured through regression of the canonical hemodynamic response functions to fit both estimated motion parameters and MSD. The MSD is further used to identify specific scans containing residual motion, data which is suppressed by adding nuisance regressors to the general linear model; this statistical suppression is performed for identified problem subjects, but has potential for use over all subjects. For problem subjects, our results show increased hemodynamic activity more consistent with group results; that is, the addition of nuisance regressors resulted in a doubling of the correlation between the activation map for the problem subjects and the activation map for all subjects. The proposed method should be useful in helping fMRI researchers make more efficient use of their data by reducing the need to exclude entire subjects from studies and thus collect new data to replace excluded subjects.

© 2012 Elsevier Inc. All rights reserved.

Correspondence to: Thomas Bauer, 7220 Peregrine Rd NE, Albuquerque, New Mexico 87113, Tel.: 505.550.3559, tbauer01@unm.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

motion correction; motion detection; realignment; regression; quality control; functional magnetic resonance imaging

1. Introduction

The need for correction of subject motion in functional magnetic resonance imaging (fMRI) studies is a universal and well-documented problem [1]. The severity of this problem in fMRI studies is aggravated by the anticipation of very small signal changes (typically between 1% and 5%) in the hemodynamic response, such that a very small amount of motion may appear to trigger false activations or deactivations [2,3]. While there are physical methods of reducing subject motion (e.g. bite bars, facial masks), these methods are often both inconvenient and distracting to subjects; such distractions may trigger spurious activations unrelated to the given tasking, leading to inaccurate results [4]. Furthermore, these methods are merely attempts at motion prevention, and substantial motion in need of correction may still occur. As such, postprocessing image registration techniques are universally used to perform spatial realignment [5]. Zitová and Flusser offer a comprehensive survey of general image registration algorithms available to researchers [6]. Even perfect realignment is not always enough to fully correct for the effects of motion, however, as the movement of spins into different slice-selection regions may affect the acquired MR signal in the form of spin-history artifacts which cannot be corrected with realignment alone [7].

While research into the overall effectiveness of these methods has been done in the past, there is limited research on subject-specific quality control of the data output by these realignment algorithms. A very common approach to motion correction quality control used in many fMRI studies solely involves analysis of estimated motion parameters as output by the realignment algorithm in charge of motion correction, specifically the exclusion of any subject whose estimated motion surpasses a certain arbitrary threshold, usually on the order of one voxel or less [8–11]. However, any approach based on estimated motion parameters necessarily assumes that the realignment algorithm has successfully estimated subject motion, and also assumes that said motion is not correctable beyond a certain amount of movement. We propose that what may be estimated as a small amount of motion could actually be the result of motion underestimation, implying the existence of uncorrected residual motion in the supposedly realigned data. Similarly, large values for estimated motion parameters may be the result of accurate motion estimation resulting in proper realignment (although not necessarily removal of spin-history artifacts). Furthermore, motion thresholds are not robust enough to account for any characterizations of motion beyond maximum displacement. For example, there remains the distinct possibility of excluding subjects who exhibit only one large, brief movement or a gradual, correctable drift associated with the heating of the gradient coils, while still including subjects who oscillate continuously and quickly at displacements below the chosen threshold. The ignored timings of these motions are not insignificant, as even small motions may have considerable negative effects on analysis if they are at all correlated with the stimuli [1].

Other, more sophisticated methods have been proposed by Diedrichsen and Shadmehr [12] and Lemieux et al. [13], both of which effectively reweight the influence of those scans assumed to have motion artifacts in respect to the general linear model (GLM). While the proposed method involves automatically adjusting the influence of motion-affected scans, it is more closely related to the Lemieux method in that it uses the same approach to reweighting (in the form of additional subject-specific regressors to the GLM design matrix)

and avoids the need for reliable noise variance estimates required by the former method. The method by Lemieux et al. involves the augmentation of the design matrix by scan-nulling regressors at or around any scan where the estimated motion exceeds the 0.2 mm threshold. While the shape of our additional regressors is the same as in the Lemieux method, our proposed method uses unsupervised learning techniques and multi-stage classification based on a number of metrics to automatically identify both datasets and timepoints in need of design matrix augmentation instead of relying on arbitrary motion thresholds.

We have developed a quality control method employing metrics intrinsic to the data before and after realignment as a classification tool for judging the effectiveness of realignment on each fMRI subject in any given study, allowing the researcher to identify and address the above issues. While this method does not directly incorporate the physics behind spin-history artifacts, the inclusion of pre-realignment data in the classification process addresses the potential for spin-history artifacts independent of the realignment process; inclusion of post-realignment data addresses the quality of realignment itself. The relationship between spin-history artifacts and post-realignment error variance is discussed more comprehensively by Andersson et al. [14]. Our proposed quality control method is intended for use in group studies, as it is meant to find natural boundaries between good and bad quality data (in the context of residual motion). For any subjects identified as having residual motion, our chosen metrics may also be used to identify improperly realigned scans to be suppressed and ignored during analysis through the introduction of additional regressors to the general linear model.

This study explored motion with an ongoing study of high-risk adolescents (PI: Bryan), a subject demographic found to evidence excessive motion [15]. Prior to the development of the proposed method, subjects had been excluded from fMRI analysis based on their maximum estimated translation from starting position parallel to either of the x -, y -, or z -axes. Exclusion of subjects based on thresholds of both 2 mm and 3 mm in any dimension was attempted, but due to the excessive motion exhibited by the adolescent subjects, these criteria were found to be prohibitively exclusive, necessitating a more precise method of identifying residual motion and excluding bad timepoints.

2. Methods

2.1 Criteria and Approach

The metric used in the first stage of our method, the mean square derivative (MSD), is based on the concept of mean square error (MSE), the most common difference metric in image processing [16]. The mean square derivative is the MSE between successive scans of an acquisition, or the spatial mean of the squared time derivative. The MSD does not measure the specific motion parameters required to realign data, but rather measures the existence and severity of motion. The MSD function $F(t)$ can be calculated from the 4-D BOLD signal $S(x, y, z, t)$ as

$$F(t) = \frac{\sum_{x=1}^X \sum_{y=1}^Y \sum_{z=1}^Z |S(x, y, z, t) - S(x, y, z, t-1)|^2}{XYZ}, \quad \text{for } t=2, 3, \dots, T \quad (1)$$

where X , Y , Z , and T are the number of voxels per scan in the x , y , and z dimensions and the number of scans, respectively.

An MSE-based metric is an ideal quality control measure for fMRI, as it is effective in detecting the existence of motion while being relatively insensitive to changes in BOLD

signal. We demonstrated this relative insensitivity by separately simulating motion and BOLD signal increases on sample acquisitions from the study in question and recording their effects on the MSE. For a simulation using voxels sized $3.75 \text{ mm} \times 3.75 \text{ mm} \times 4.55 \text{ mm}$, a very minor 1 mm shift in a single direction results in a mean square error equivalent to the result of an 8% change in BOLD activation over the entire brain. A full-voxel shift of 3.75 mm is equivalent to a 29% change in activation. As typical changes in BOLD signal due to hemodynamic response range between 1% and 5%, unusually large spikes in the mean square derivative should reliably detect the existence of residual motion without large influences from the task activity (refer to Fig. 1 for visual details). While the MSE metric is very useful in the detection of significant motion artifacts, it is not ideal for the computation of precise realignment parameters [17]. As such, most state-of-the-art image registration algorithms minimize other, more computationally demanding error metrics during the realignment process. A comparison of the relative performances of some of these error metrics with respect to fMRI realignment has been conducted by Freire et al. [18], with particular attention being given to the Geman-McClure (GM) robust estimator, the correlation ratio, and mutual information.

This stage of the proposed method is intended merely to *detect* improper realignment rather than to fine-tune realignment parameters. Thus, MSE may be considered more useful as an independent “second opinion” on the quality of realignment than more advanced metrics utilized by current image registration algorithms, as MSE has not already been directly minimized by the very realignment process we are attempting to evaluate. That is to say, it is necessary to use an independent metric such as MSE for quality control to avoid simply reaffirming the output of whichever realignment algorithm is chosen during analysis. Luo and Nichols have also used MSE-related metrics to identify problematic timepoints [19], although not within the framework of unsupervised learning as proposed here.

The proposed method includes three stages, an overview of which can be seen in Fig. 2. In the first stage of the method, three features are collected from the MSD signals both before and after realignment for each subject, for a total of six features per subject. The features are then grouped using k-means clustering to find natural boundaries between acceptable and unacceptable data.

The second stage of the method uses two more metrics to complement the first stage. The first of these second-stage metrics is the correlation between the subject activation map and the mean activation map for all subjects in the corresponding group (possible groups being either patients or controls). The second of these metrics represents the relationship between tasking and motion, where tasking is represented by the hemodynamic response functions (HRFs). Three multiple regressions are performed, using as regressors the canonical hemodynamic HRFs for each stimulus, their first derivatives, and the squares of each HRF and derivative: these functions are regressed to fit the estimated total displacement, the estimated total rotation, and the original MSD. The maximum R-value resulting from these three multiple regressions is used as a measure of the relationship between tasking and motion (the second metric). The two metrics from this stage, activation map correlation and tasking-motion relationship, are used to further identify possible motion-prone subjects.

A third stage employs modifications to the general linear model involving exclusion of the timepoints subject to the most uncorrected motion. Using the MSD to identify bad timepoints, we introduce impulse functions (single nonzero values at the corresponding timepoints) as extra regressors into the GLM to effectively negate the influence of questionable timepoints in the regression analysis (See Fig. 3). The modification of the GLM with scan-nulling impulse regressors has precedent in the work of Lemieux et al. with respect to reducing the effects of head motion in studies of patients with epilepsy [13].

2.2 Stage One

Using the MSDs of both the original scan data and realigned data, six features are extracted: the mean of the original MSD signal, the standard deviation of the original MSD, the percentage of original MSD greater than a chosen threshold, the mean of the realigned MSD, the standard deviation of the realigned MSD, and the percentage of realigned MSD greater than a chosen threshold. The first chosen threshold is one standard deviation from the mean of all MSD signals, and the second is one standard deviation from the mean of all realigned MSD signals. These two empirical thresholds are used for feature extraction purposes, so the use of one standard deviation versus other standard deviations does impact the methods classification of samples.

To find natural boundaries between acceptable and unacceptable data, we employ k-means clustering using the six MSD statistics as features. Because the distributions of the percentage-based features will likely differ from the remaining feature distributions by multiple orders of magnitude, all features should be normalized according to a distribution with a mean of zero and a standard deviation of one. Each feature will then have equal weighting prior to clustering, helping to objectively characterize the motion for each subject. We have chosen a majority-voting approach to classification, wherein samples are clustered and classified multiple times with different numbers of possible clusters. This reduces the dependency of classification on a single classification attempt or on a single decision boundary. In this paper, the samples are clustered five times (although any odd number of clustering attempts will work for majority voting), and are first divided into two classes, then into three classes, four classes, five classes, and six classes. For each clustering run, classes with above-average distributions of feature values should be aggregated and designated “high”, and classes with below-average distributions of feature values should be aggregated and designated “low”, resulting in a total of five classifications per subject. These classifications lead to an overall quality verdict through majority voting, with any sample with three or more “high” or “low” classifications being ultimately classified as “good” or “bad”, respectively. In the case of an ongoing study, these classified samples can serve as a training set for a nearest-neighbor classifier that can be easily used on future acquisitions in the study.

2.3 Stage Two

Once the preliminary classifications have been found, a second stage may be used to further identify motion-prone subjects. The two metrics used here are the correlation between each subject activation map and the mean activation map from the corresponding control group, and the strength of relationship between tasking and motion as computed by multiple regression of HRF basis functions to motion parameters. The reasoning behind use of the first measure is that regardless of estimated motion or MSD, a very low or negative correlation between activation for a specific subject and the mean activation implies disparate enough data, either due to motion or other unknown circumstances, to warrant the further scrutiny provided by the third stage of the method. Note that we are not suggesting complete removal of these subjects from the study based on this metric alone, as removal of subjects due to lack of correlation between activation maps may violate the assumption of statistical independence. This step of the method is only meant to further identify those subjects whose activation is particularly unusual when compared to the mean. It is also possible that unusual activation maps are a sign of a subject performing the task incorrectly, a sign of significant quality problems with scanner hardware, or any number of other problems. This metric was found to be quite useful when evaluating data from multiple scanners in the fBIRN and MCIC multi-site studies (both of which include hundreds of subjects and multiple scanners); Segall et al. discuss a similar metric to identify outliers in a structural study using the fBIRN and MCIC data [20].

The relationship between tasking and motion is computed by regressing the canonical HRFs for each stimulus, their first derivatives, and the squares of each HRF and derivative three times: once to fit the estimated total displacement from starting position, once to fit the estimated total rotation from starting position, and once to fit the MSD signal of the data collected before realignment. HRF derivatives and their squares are included as regressors to help correct for possible latency errors present in the canonical HRFs [21]. The estimated total displacement $D(t)$ is computed using the Euler distance formula on the motion parameters representing translation in the x , y , and z directions as output by the chosen realignment algorithm; the total rotation $R(t)$ is estimated as the arccosine of the product of the cosines of the pitch α , roll β , and yaw γ , also as output by the realignment algorithm.

$$D(t) = \sqrt{x(t)^2 + y(t)^2 + z(t)^2} \quad (2)$$

$$R(t) = \cos^{-1}[\cos(\alpha(t))\cos(\beta(t))\cos(\gamma(t))] \quad (3)$$

We then compute the R-values resulting from each multiple regression and its corresponding target function; the largest of these three values is taken to be a measure of the relationship between tasking and motion. A significant relationship between tasking and motion suggests a difficulty in being able to distinguish tasking responses from motion artifacts, leading to the possibility of motion undercorrection or even overcorrection [1,17]. The correlation between tasking and motion was observed in a study performed by Field et al, in their study they determined a significant correlation to have an R-value greater than 0.52 [26]. Although realignment algorithms based on mutual information have demonstrated an improved ability to distinguish between tasking responses and motion when compared to the realignment algorithms packaged with most fMRI analysis software, the observation of any significant relationship between tasking and motion still indicates possible trouble during realignment. Such cases should result in a “bad” classification for the subject in question, regardless of stage one result.

A sample will be further scrutinized by the third stage of the method unless both the first and second stages of the method resulted in good classifications; this ensures that all questionable datasets will be included in the method’s third stage.

2.4 Stage Three

As an alternative to simply excluding subjects from a particular analysis, it is possible to automatically identify problem timepoints and include extra “nuisance” regressors into the GLM with the intent of eliminating the statistical influence of these timepoints. In this sense, specific scans containing residual motion are mathematically ignored rather than subjected to further correction attempts. In order to negate the influence of a timepoint for any given subject during regression, a single impulse function can be included as a nuisance regressor to the GLM, present only in that subject’s portion of the design matrix [13]. Recall that regression minimizes the squared error between a linear combination of the regressors and the target data; if the nuisance regressor consists of all zeros except for a single nonzero element at the time corresponding to the residual motion artifact, the system will have the flexibility to ensure a perfect fit resulting in a squared error of zero for that timepoint [22]. In this way, the final least square error solution to the regression is completely unaffected by the timepoint in question, lessening the resulting residual error. As long as the original regressors which compose the initial design matrix have at least one nonzero value outside of the identified problem timepoints (a virtual certainty), they are assured linear independence of the proposed nuisance regressors. As such, the nuisance regressors may be

used alongside other motion-related regressors which have already been included to help reduce the impact of subject motion, as is described and evaluated by Lund et al. [23]. A similar effect may also be accomplished by reweighting the influence of specific timepoints in the GLM, resulting in a “soft” exclusion of timepoints, but the “hard” exclusion method of including scan-nulling regressors will be presented here to demonstrate the more extreme case of completely removing timepoints from analysis.

The nuisance regressors simply augment the existing design matrix to suppress the influence of the data at the chosen timepoints, allowing inclusion of problem subjects in analysis while greatly reducing the effects of their motion on the group statistics. While the construction of a subject-specific design matrix would initially appear to violate the assumption of statistical independence underlying the GLM, that assumption relies on the presence of a stationary noise process with time-homogeneous variance: physiological noise in the form of subject motion already violates this assumption [12]. Therefore, the addition of any subject-specific regressors specifically meant to model subject motion is in an attempt to compensate for the violation of the stationary noise assumption which has already occurred. Lemieux et al. further address the statistical implication of using scan-nulling impulse regressors to remove the effects of motion artifacts at specific timepoints [13].

Regarding the actual identification of problem timepoints, we have empirically determined problem timepoints to be those with realigned MSD values greater than one standard deviation above the MSD signal mean. Each of these problem timepoints then necessitates its own nuisance regressor in the form of an impulse function centered at the timepoint in question. Fig. 3 shows the selection of problem timepoints and their corresponding nuisance regressors.

2.5 Data Source

All data were captured from fMRI scans of adolescent subjects between the ages of 14 and 18 using a 3T Siemens Trio scanner. Thirty-three slices 3.5 mm thick were captured axially, with a separation of 4.55 mm between the centers of adjacent slices. Each slice measures 64×64 pixels; each pixel is sized $3.75 \text{ mm} \times 3.75 \text{ mm}$. Thus, each scan has total dimensions of $64 \times 64 \times 33$ with voxels of size $3.75 \text{ mm} \times 3.75 \text{ mm} \times 4.55 \text{ mm}$. Two tasks from the study in question, both cognitive tasks associated with decision making, were used for the purposes of this research; these tasks are ideal for a motion study, as there is a higher probability of task-correlated motion during the error trials in each task [24]. The first task has a length of 220 scans and the second has a length of 189 scans; for both tasks, each scan has a TE of 29 ms and a TR of 2 s.

3. Results

We performed the method on the previously described data, using acquisitions from two different tasks and employing the INRIAlign toolbox built into Statistical Parametric Mapping (SPM5) as our realignment algorithm [18]. Because INRIAlign employs the Geman-McClure robust estimator to identify subject motion, MSD functions well as a “second opinion”. Using data from 50 subjects for the first task and 45 subjects for the second task, we extracted the mean and standard deviation of the original MSD, the percentage of original MSD values above 2×10^5 , the mean and standard deviation of the realigned MSD, and the percentage of realigned MSD values above 1.5×10^5 . The first threshold is one standard deviation from the mean of all MSD signals and the second is one standard deviation from the mean of the realigned MSD signals. Clustering and majority voting for the first task yielded 46 subjects with good data and four subjects with bad data; the same process for the second task yielded 34 subjects with good data and 11 with bad data. The nearest-neighbor classifiers based on these results exhibited no errors during

leave-one-out cross validation. Leave-one-out cross validation is a common way of determining the effectiveness of a classification method on a particular training set: it consists of the removal of a single sample from the dataset and the subsequent comparison of its true classification to its classification as determined by a decision rule generated using the remaining samples as a training set, as repeated for all samples in the dataset. The presence of disparities between classifications for any sample implies ambiguity on the part of either the training set or the classification method, and in the context of this research helps to determine both the success of the original clustering method that created the training set and the potential for proper classification of future subjects.

For the second stage of the method, the activation correlations were distributed between values of -0.21 and 0.78 for the first task, and between values of -0.22 and 0.85 for the second task, with all high and low values corroborating the first stage results (we chose 0.20 as the low threshold for activation correlation), save for one exception. There was a single sample from the first task that had been cleared as “good” by the MSD clustering stage yet exhibited an activity correlation of -0.11 ; accordingly, this one sample was reclassified as a problem sample. For both tasks, all maximum R-values for tasking-motion regression were below 0.18 , implying that for all samples, tasking and motion were distinguishable and separable enough that changes to the first stage classifications were unnecessary.

According to the method’s third stage, problem timepoints were empirically determined as those with MSD values one standard deviation above the mean of the realigned MSD signal. Inclusion of the extra nuisance regressors for the five subjects with bad data from task one doubled the correlation between the group activation map for the five subjects and the mean activation map from 0.19 to 0.38 , even though only a single subject was reclassified due to the correlation measure of the second stage. Fig. 7 shows the increased activation seen from inclusion of the nuisance regressors.

4. Discussion

The goals of the method were both proper identification of uncorrected residual motion and the subsequent suppression of this residual motion through statistical techniques in the GLM. The ideal results should show proper separation between good and bad clusters from stage one according to the chosen metrics, agreement between the first and second stage results, and the increased power in the GLM output for the bad cluster after the addition of nuisance regressors. The presence of these three properties in the results would imply a measure of success in the proposed method as a step in the fMRI analysis pipeline between motion correction and the general linear model, helping to catch motion that will inevitably be missed by any motion correction algorithm.

The first stage results show a reasonable separation between good and bad samples. Despite the fact that clustering was performed in six dimensions, there are clear one-dimensional divisions between good and bad samples for nearly every feature; the only exceptions to this are the two features regarding percentage of MSD above the threshold (see Figs. 3 and 4), and this lack of one-dimensional separation is only present in the samples from the first task. All features for the second task show one-dimensional separation. The existence of one-dimensional separation in specific features as demonstrated by classification results is not a necessary property of a classifier to be declared reliable, but the existence of such separation here does serve to inspire confidence in the method. The fact that both nearest-neighbor classifiers created using the clustering output as training data exhibit no errors whatsoever in leave-one-out cross validation is an even stronger indication of clustering success, especially in consideration of the relatively small number of “bad” classifications in the first task. Even with only four “bad” samples from task one, the exclusion of any one of the “bad” samples

from the training set (25% of the class) still results in the same classification for the excluded sample.

The analysis in Fig. 6 helps to show the difference between classifications resulting from the first stage of the proposed method and classifications resulting from the original motion cutoff method initially used for the study in question. While the results from the MSD method do agree better with a cutoff of 3 mm maximum translation parallel to either of the x -, y -, or z -axes rather than a cutoff of 2 mm, neither cutoff agrees enough with the results to function as a simple substitute. It is further visible in Fig. 6 that for the first task, the sample with the highest maximum estimated translation still exhibited low enough MSD values to be classified as having an acceptable level of residual motion; it may be possible that the high maximum value for estimated motion is an effect of successful motion correction and that the slightly lower motion estimations for other samples are an effect of underestimation and incomplete correction.

The R-values representing the tasking-motion relationship for each subject in stage two were all below 0.18 which allowed us to conclude the correlation was not significant enough to suggest a difficulty in being able to distinguish tasking responses from motion artifacts so no classifications were changed. While these low values are an indication of motionless or correctable data, these measures are likely still liberal estimates for some subjects due to the likelihood of the R-value being driven by the motion found in scans which were simply regressed away in stage three. That is to say, if there are but a very few large motions, all of which coincide with the tasking, they will drive the resulting R-value to a high value regardless of how weakly correlated the tasking and motion are in the remaining scans. Yet these are the same scans that would end up being removed from the GLM through the use of nuisance regressors. These liberal estimations serve to be more vigilant against possible motion, even if the case of high R-values did not apply to these data.

Although there were very few changes made in the override stage, we still consider it to be a valuable step of the process, certainly in providing context to the success of the nuisance regression stage. That is to say, the change in subject activation map correlation resulting from the addition of nuisance regressors to the GLM helps to judge the effectiveness of the method in reducing the effects of motion. Judging the success of motion-related nuisance regression based on a comparison of activation maps before and after augmentation of the design matrix has precedent in the work of Johnstone et al. [25]. In the case of our own results, the introduction of nuisance regressors caused a dramatic increase in the consistency of activation between the problem subjects and the group (as evidenced by the doubling of the correlation between activation maps to well within normal range); because this change occurred solely due to the exclusion of a few high-MSD timepoints, it is very likely that the original inconsistency in activation was mainly the result of subject motion. Thus, the change observed in second stage metrics strongly suggests the proposed method's success in reducing the effects of motion on fMRI analysis.

5. Conclusion

Based on the systematic nature of the proposed method and the observed results, we believe that this method offers neuroimaging researchers a less arbitrary set of criteria for the identification of problem subjects than current criteria involving the arbitrary thresholding of estimated motion. Additionally, we expect that the option to suppress residual motion through changes to the general linear model instead of excluding entire subjects will allow researchers to make more efficient use of their data, potentially minimizing the time and expense required to scan additional subjects to replace those previously excluded. For those wishing to further save on time, it may be possible to bypass the first two steps of the

method and still achieve quality results by automatically generating and including nuisance regressors for all subjects in a study. Suppression of high-MSD scans for all subjects would not only provide increased vigilance in detecting residual motion, but may, from an objective experimental standpoint, be an attractive option to ensure the truly uniform treatment of all subjects during post-processing. While our designation of problem timepoints as those with realigned MSD values one standard deviation above the mean does ensure that only a fraction of timepoints are excluded, it may be preferable under these circumstances to set a common MSD threshold across all subjects to give increased control over the total number of suppressed scans. The effectiveness of this and other possible condensations of the method may be the subject of future exploration.

Acknowledgments

This research was supported by grants from the National Institutes of Health awarded to Vince D. Calhoun (2R01 EB000840) and Angela D. Bryan (R01 AA017390).

References

1. Hajnal JV, Myers R, Oatridge A, Schwieso JE, Young IR, Bydder GM. Artifacts due to stimulus correlated motion in functional imaging of the brain. *Magn Reson Med*. 1994; 31(3):283–291. [PubMed: 8057799]
2. Aguirre GK, Zarahn E, D’Esposito M. The Variability of Human, BOLD Hemodynamic Responses. *NeuroImage*. 1998 Nov; 8(4):360–369. [PubMed: 9811554]
3. Miezin FM, Maccotta L, Ollinger JM, Petersen SE, Buckner RL. Characterizing the Hemodynamic Response: Effects of Presentation Rate, Sampling Procedure, and the Possibility of Ordering Brain Activity Based on Relative Timing. *NeuroImage*. 2000 Jun; 11(6):735–759. [PubMed: 10860799]
4. Jiang A, Kennedy DN, Baker JR, Weisskoff RM, Tootell RBH, Woods RP, Benson RR, Kwong KK, Brady TJ, Rosen BR, et al. Motion detection and correction in functional MR imaging. *Hum Brain Mapp*. 1995; 3(3):224–235.
5. Friston KJ, Ashburner J, Frith CD, Poline J, Heather JD, Frackowiak RSJ. Spatial registration and normalization of images. *Hum Brain Mapp*. 1995; 3(3):165–189.
6. Zitová B, Flusser J. Image registration methods: a survey. *Image Vis Comput*. 2003 Oct; 21(11):977–1000.
7. Friston KJ, Williams S, Howard R, Frackowiak RS, Turner R. Movement-related effects in fMRI time-series. *Magn Reson Med*. 1996 Mar; 35(3):346–355. [PubMed: 8699946]
8. Breiter HC, Aharon I, Kahneman D, Dale A, Shizgal P. Functional Imaging of Neural Responses to Expectancy and Experience of Monetary Gains and Losses. *Neuron*. 2001; 5;30(2):619–639.
9. Durston S, Mulder M, Casey B, Ziermans T, Vanengeland H. Activation in Ventral Prefrontal Cortex is Sensitive to Genetic Vulnerability for Attention-Deficit Hyperactivity Disorder. *Biol Psychiatry*. 2006; 11;60(10):1062–1070.
10. MacIntosh BJ, McIlroy WE, Mraz R, Staines WR, Black SE, Graham SJ. Electrodermal recording and fMRI to inform sensorimotor recovery in stroke patients. *Neurorehabil Neural Repair*. 2008 Dec; 22(6):728–736. [PubMed: 18784267]
11. Moeller FG, Steinberg JL, Dougherty DM, Narayana PA, Kramer LA, Renshaw PF. Functional MRI study of working memory in MDMA users. *Psychopharmacology (Berl)*. 2004 Dec 10; 177(1):185–194. [PubMed: 15221201]
12. Diedrichsen J, Shadmehr R. Detecting and adjusting for artifacts in fMRI time series data. *NeuroImage*. 2005 Sep; 27(3):624–634. [PubMed: 15975828]
13. Lemieux L, Salekhaddadi A, Lund T, Laufs H, Carmichael D. Modelling large motion events in fMRI studies of patients with epilepsy. *Magn Reson Imaging*. 2007; 7;25(6):894–901.
14. Andersson JLR, Hutton C, Ashburner J, Turner R, Friston K. Modeling Geometric Deformations in EPI Time Series. *NeuroImage*. 2001 May; 13(5):903–919. [PubMed: 11304086]

15. Feldstein, Ewing SW.; Filbey, FM.; Claus, ED.; Nyalakanti, P.; Kiehl, KA.; Hutchison, KE.; Bryan, AD. Neurological Correlates of Alcohol Abuse Behavior with High Risk, Juvenile Justice-Involved Youth. *Alcohol Clin Exp Res.* 33(s1):296a.
16. Pratt, WK. Digital image processing: PIKS Scientific inside. 4. Hoboken, NJ: Wiley-Interscience; 2007.
17. Freire L, Mangin J-F. Motion Correction Algorithms May Create Spurious Brain Activations in the Absence of Subject Motion. *NeuroImage.* 2001 Sep; 14(3):709–722. [PubMed: 11506543]
18. Freire L, Roche A, Mangin J-F. What is the best similarity measure for motion correction in fMRI time series? *IEEE Trans Med Imaging.* 2002; 21(5):470–484. [PubMed: 12071618]
19. Luo W, Nichols TE. Diagnosis and exploration of massively univariate neuroimaging models. *NeuroImage.* 2003 Jul; 19(3):1014–1032. [PubMed: 12880829]
20. Segall JM, Turner JA, van Erp TGM, White T, Bockholt HJ, Gollub RL, Ho BC, Magnotta V, Jung RE, McCarley RW, Schulz SC, Lauriello J, Clark VP, Voyvodic JT, Diaz MT, Calhoun VD. Voxel-based morphometric multisite collaborative study on schizophrenia. *Schizophr Bull.* 2009; 35(1):82–95. [PubMed: 18997157]
21. Calhoun VD, Stevens MC, Pearlson GD, Kiehl KA. fMRI analysis with the general linear model: removal of latency-induced amplitude bias by incorporation of hemodynamic derivative terms. *NeuroImage.* 2004 May; 22(1):252–257. [PubMed: 15110015]
22. Fox, J.; Monette, G. An R and S-Plus Companion to Applied Regression. Thousand Oaks, CA: Sage Publications, Inc; 2002.
23. Lund TE, Nørgaard MD, Rostrup E, Rowe JB, Paulson OB. Motion or activity: their role in intra- and inter-subject variation in fMRI. *NeuroImage.* 2005 Jul 1; 26(3):960–964. [PubMed: 15955506]
24. Kiehl KA, Liddle PF, Hopfinger JB. Error processing and the rostral anterior cingulate: An event-related fMRI study. *Psychophysiology.* 2000; 37(2):216–223. [PubMed: 10731771]
25. Johnstone T, Walsh KSO, Greischar LL, Alexander AL, Fox AS, Davidson RJ, Oakes TR. Motion correction and the use of motion covariates in multiple-subject fMRI analysis. *Hum Brain Mapp.* 2006; 27(10):779–788. [PubMed: 16456818]
26. Field AS, Yen YF, Burdette JH, Elster AD. False cerebral activation on BOLD functional MR images: study of low-amplitude motion weakly correlated to stimulus. *Am J Neuroradiol.* 2000; 21:1388–1396. [PubMed: 11003269]

Fig. 1.

Comparison between sensitivities of mean square error to motion and to BOLD signal change. In terms of effect on the MSE, a small amount of motion is equivalent to a very large change in BOLD signal. The MSE for the first and second set of images are very similar but there are some white distortions in the images where the MSE resulted from subject motion and not an increase in the BOLD signal.

Fig. 2.
Block diagram overview of the proposed method.

Fig. 3. Demonstration of selection of problem timepoints above the empirically determined threshold one standard deviation from the mean and the corresponding nuisance regressors used to suppress these timepoints.

Fig. 4. Distributions of features derived from the original MSD signal, shown with first-stage classifications. There are clear one-dimensional separations between good and bad samples of nearly every feature; the only exceptions being the two features regarding percentage of MSD above the thresholds.

Fig. 5. Distributions of features derived from the realigned MSD signal, shown with first-stage classifications. Samples still display one-dimensional separation inspiring confidence in the purposed method.

Fig. 6.

Distributions of maximum estimated translations in any of the x , y , or z directions, shown with first-stage classifications. The dotted and dashed lines represent the motion thresholds of 2 mm and 3 mm, respectively, that were previously used as criteria for the study (and which represent the industry standard). Samples to the right of any given threshold line would originally have been excluded from analysis, with only those samples to the left of the threshold line being utilized; however, it can be seen that the proposed method detected both “good” and “bad” datasets on either side of each threshold line. In fact, there is no motion threshold that would perfectly agree with our first-stage results.

Effects of nuisance regressors on group analysis of problem subjects

Fig. 7.

The effects of addition of nuisance regressors to the general linear model for the five problem subjects from task one. Note the increased activation observed after inclusion of the nuisance regressors, particularly the addition of activity present in the parietal lobe.