NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage xx (2005) xxx – xxx

Comparison of fMRI motion correction software tools

T.R. Oakes,* T. Johnstone, K.S. Ores Walsh, L.L. Greischar, A.L. Alexander, A.S. Fox, and R.J. Davidson

Waisman Laboratory for Brain Imaging, University of Wisconsin-Madison, WI 53705, USA

Received 17 July 2004; revised 1 April 2005; accepted 5 May 2005

Motion correction of fMRI data is a widely used step prior to data analysis. In this study, a comparison of the motion correction tools provided by several leading fMRI analysis software packages was performed, including AFNI, AIR, BrainVoyager, FSL, and SPM2. Comparisons were performed using data from typical human studies as well as phantom data. The identical reconstruction, preprocessing, and analysis steps were used on every data set, except that motion correction was performed using various configurations from each software package. Each package was studied using default parameters, as well as parameters optimized for speed and accuracy. Forty subjects performed a Go/No-go task (an event-related design that investigates inhibitory motor response) and an N-back task (a block-design paradigm investigating working memory). The human data were analyzed by extracting a set of general linear model (GLM)-derived activation results and comparing the effect of motion correction on thresholded activation cluster size and maximum *t* value. In addition, a series of simulated phantom data sets were created with known activation locations, magnitudes, and realistic motion.

Results from the phantom data indicate that AFNI and SPM2 yield the most accurate motion estimation parameters, while AFNI's interpolation algorithm introduces the least smoothing. AFNI is also the fastest of the packages tested. However, these advantages did not produce noticeably better activation results in motion-corrected data from typical human fMRI experiments. Although differences in performance between packages were apparent in the human data, no single software package produced dramatically better results than the others. The "accurate" parameters showed virtually no improvement in cluster *t* values compared to the standard parameters. While the "fast" parameters did not result in a substantial increase in speed, they did not degrade the cluster results very much either.

The phantom and human data indicate that motion correction can be a valuable step in the data processing chain, yielding improvements of up to 20% in the magnitude and up to 100% in the cluster size of detected activations, but the choice of software package does not substantially affect this improvement.

Introduction

Head motion can have a profound effect on the activation signal from fMRI studies (Hajnal et al., 1994; Thacker et al., 1999). Even if motion is small compared to the fMRI voxel size, it can still corrupt the raw BOLD images, invalidating the assumption that the variation of intensity between image frames is due primarily to changes in cerebral physiology (Friston et al., 1996). The need to remove this confound from the data led to efforts to measure and correct for head motion, usually with post hoc methods such as coregistration of each fMRI volume to a reference volume (Thacker et al., 1999; Jenkinson and Smith, 2001). The term "motion correction" usually refers to small (~mm) intrasubject corrections, which typically only correct for translation and rotation either within or across scan runs of the same subject. Most motion correction algorithms use a rigid-body fit which assumes the head shape is constant between frames. An implicit assumption is that motion occurring during the acquisition of a given frame can be treated as if it occurred all at once, e.g., in the small amount of time between frames. While this is usually not the case, it is nevertheless a practical assumption which permits motion correction to proceed with reasonable accuracy.

The development of automated image coregistration algorithms enabled large data sets to be corrected for motion. Roger Woods' automated image registration (AIR) method (Woods et al., 1992, 1993) used information taken from both the object and target images to create a cost function, which quantified the amount of overlapping information. An early comparison of the leading coregistration methods of the day (Strother et al., 1994) determined that a coregistration optimization scheme based on the measurement of the similarity of the spatial distributions of voxel values, as embodied by the AIR algorithm, was superior to other techniques. This technique was subsequently applied to motion correction of fMRI images (Jiang et al., 1995). A high-quality but time-consuming interpolation such as the 3D sync-interpolation with Hanning window was shown to be desirable for fMRI data (Ostuni et al., 1997). A variety of software packages and algorithms are currently available to correct fMRI time series for motion (Friston et al., 1994, 1995; Cox, 1996; Cox and Jesmanowicz, 1999; Biswal and Hyde, 1997; Studholme et al., 1997; Woods et al., 1998a,b; Singh et al., 1998; Kim et al., 1999; Ciulla and Deek, 2002).

* Corresponding author.
E-mail address: troakes@wisc.edu (T.R. Oakes).

Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2005.05.058

There is a relative paucity of comparisons of fMRI analysis packages in general, although some comparisons of the coregistration components of various software tools have been undertaken. West et al. (1997) compared twelve (12) brain coregistration techniques, but fMRI was not one of the modalities examined. Similarly, a more recent article (Hellier et al., 2003) compared 6 different types of coregistration algorithms, but used high-resolution MRI images instead of typically lower-resolution functional BOLD images. Koole et al. (1999) examined eight different algorithms for coregistration of SPECT functional images; the AIR package was the only one examined that is also commonly used for fMRI, and AIR was determined to be overall the fastest and most accurate. Since none of these works explicitly addressed registration of typical fMRI-BOLD images, their conclusions provide only modest guidance in selecting an fMRI motion correction tool.

An examination of several fMRI analysis packages was performed using what are now considered older versions of currently available software. Gold et al. (1998) tested features including motion correction from five packages (AFNI 2.01, SPM96, Stimulate 5.0, MEDIMAX 2.01, and FIT) and described three others (FIASCO, Yale, and MEDx 2.0). The Gold et al. (1998) article is admittedly more descriptive than analytic and does not contain ratings for accuracy of motion correction, although the comments regarding usability are informative.

Morgan et al. (2001) examined the efficacy of motion correction of three packages (SPM99b, AFNI98, and AIR3.08). This work featured an innovative computer-generated phantom based on actual EPI data to create a known amount of movement with simulated noise and local activations. The motion correction was examined both in light of the accuracy compared to the introduced movement, and also with respect to various indices of activation detection as determined by yet another package, Stimulate (Strupp, 1996). The results demonstrate that correcting for motion between frames increases the specificity of activation findings; furthermore, the contemporary versions of each software package examined produced similar results both with respect to the accuracy of motion correction as well as the ability to correctly detect an activation signal in a phantom. Three different motion correction approaches employed in a recent effective-connectivity project (Gavrilescu et al., 2004) also yielded no substantial difference in results.

A study examining fMRI tools (Ardekani et al., 2001) using simulated data determined that the motion correction implemented by SPM99 was slightly more accurate than AFNI, but that AFNI was several times faster, and AFNI was also more robust in the presence of noise. However, these authors concluded that AIR was the least accurate of the four packages studied, which contrasts with the results of Koole et al. This discrepancy emphasizes the variability in algorithm performance depending on such factors as the characteristics of the image data and the software parameters selected, and it also underscores the difficulty in generalizing evaluations of algorithms and analysis tools beyond the data sets they examined. This is particularly true for simulated phantom data, which may not capture the many subtle components present in biologically derived data. An important question addressed in the current work is how well various measures of algorithm accuracy (usually obtained from phantom data) can predict the relative performance of a given algorithm used with actual human data.

In the present study, we compare five (5) software tools which are in common use for motion correction of fMRI-BOLD data: AFNI (Cox, 1996), AIR (Woods et al., 1992, 1993, 1998a,b), BrainVoyager (Brain Innovations, Maastricht, The Netherlands), FSL (Smith et al., 2001), and SPM2 (Friston et al., 1994, 1995). (A sixth tool, VoxBo (Aguirre et al., 1997, 1998; Zarahn et al., 1997), was also examined, but results are not included since VoxBo's motion correction algorithm is a direct port of an older version of SPM96b and is about to be replaced.) Descriptions of each of the packages used in this work are given in Table 1 and are confined to issues relevant to motion correction. The information on file format compatibility may become irrelevant if the software authors adopt the proposed NIfTI file format standard (see http://nifti.nimh.nih.gov/).

Experimental paradigms

Most fMRI paradigms can be categorized as either a block design, where a single experimental condition is presented in a block of time occupying several TRs followed by a similar period of a control or alternate condition; or as an event-related design (Dale and Buckner, 1997), where single events of duration typically less than the TR are presented. Because of their differing time scales and the influence of activation-correlated motion, it is possible that motion correction will vary in efficacy for block versus event-related experiments. Both types of experimental paradigms were used in this study.

Data sources

Data from 40 human subjects for both a block- and event-related design were taken from the Wisconsin Imaging Tools Evaluation Resource (WINTER), a data set acquired for the express purpose of investigating methodological aspects of data analysis. A simulation (phantom) study was also undertaken to inform the results from the human study as well as to obtain absolute measures of registration and interpolation accuracy. In particular, the phantom data were used to obtain a data set with a known interplay between motion and the activation signal (see e.g., Freire and Mangin, 2001). The phantom and human data agree on several points, allowing us to bridge results from previous authors who used either phantom or human data but not both. The data sets used in this work (human and phantom) are being characterized in detail and will be made available in the future to researchers for testing and benchmark purposes.

Project goals

Our primary goal was to determine whether one motion correction tool was better than others with regard to the detected activations from a typical analysis using a general linear model (GLM) approach, and the extent to which quantifiable measures of algorithm performance influenced the relative GLM results. A secondary goal was to determine the relative contribution of motion correction toward improving activation detection for a block design and for an event-related design. The scope of this article is to compare motion correction through the criteria of the effect on the subsequent GLM-derived *t* test maps, to quantify the accuracy of the motion estimation and interpolation steps,

Table 1
Comparison of characteristics for each software tool

Software	Interface	OS supported	Source code available	File format	Cost function	Optimization technique	Interpolation	Additional features
AFNI	GUI, command line	Unix, Linux, MacOSX, Windows (cygwin)	Yes	Proprietary, ANALYZE-7.5, NIfTI	Weighted least squares	Iterative gradient descent	Fourier, also trilinear, *n*-degree polynomial	4-way 3D shear matrix factorization of rotation matrix for speed
AIR	Command line	Unix, Linux, Windows, most others	Yes	ANALYZE-7.5 (3D only)	Least squares with intensity rescaling (and other options)	Powell variant	Scanline chirp-z (Fourier variant), several others	
Brain Voyager	GUI, some scripting	Unix, Linux, Windows, MacOSX	No	Proprietary, ANALYZE-7.5	Least squares	Levenberg–Marquardt or (if that fails) gradient descent	Trilinear and sinc	
FSL McFLIRT	GUI, command line	Unix, Linux, MacOSX, Windows (cygwin)	Yes	ANALYZE-7.5, NIfTI	Normalized correlation (with several other options)	Multistart coarse search followed by two finer searches.	Trilinear and sinc.	Edge smoothing to eliminate small cost function discontinuities.
SPM2	GUI, Matlab command line, Matlab scripting	Unix, Linux, MacOSX, Windows; requires Matlab	Yes	ANALYZE-7.5, NIfTI	Least squares approach, Taylor expansion to parameterize image	Gauss–Newton	4th degree B-spline (approximates a windowed sinc function)	Optimized non-Matlab programs (*.mex files) only exist for some OS

and to relate the individual software accuracy to GLM analysis results.

Methods

Human studies

Forty subjects were recruited through local newspaper and pinup advertisements. All subjects provided informed consent, and all studies were performed in accordance with the policies of the UW-Madison's Human Subjects IRB. The goal was to recruit a range of subjects who would be representative of the "normal" types of subjects recruited as controls from the population at large.

Participants were screened by phone with an MRI-Compatibility Form, the Edinburgh Handedness Survey, and a Structured Clinical Interview for DSM-IV Axis I Disorders (SCID). Prior to the actual scanning session, subjects underwent a simulated scan in a mock scanner to accustom them to the MRI scanner environment and to train them in the performance of the different experimental tasks. Data were acquired from February through July of 2003. Subjects ranged in age from 18 to 50, with number and gender balanced within each decade: 18–29: M(8), F(6); 30–39: M(6), F(6); 40–50: M(8), F(6).

A working memory N-back task (Casey et al., 1998; Cohen et al., 1997; Smith et al., 1996) (with 0, 1, and 2 back trials, 51 s per block with 10 s of rest, 3 blocks of each condition) was used as a multi-condition block-design experiment. A Go/No-go task (Garavan et al., 1999; Liddle et al., 2001) (random ITI between 1.5 and 3.5 s, 120 Go trials, 30 No-go trials) was used to provide an event-related design. Both tasks required a response to stimuli through a button-box, so motion that is correlated with the activation paradigm is a distinct possibility.

All functional and anatomical brain images were acquired on a GE-SIGNA 3.0-T MRI scanner with a high-speed EPI gradient. Anatomical scans consisted of a high-resolution 3D T1-weighted inversion recovery fast gradient echo image (inversion time = 600 ms, 256 × 256 in-plane resolution, 240 mm FOV, 124 × 1.1 mm axial slices), a T1-weighted spin echo coplanar image with the same slice position and orientation as the functional images (256 × 256 in-plane resolution, 240 mm FOV, 30 × 4 mm sagittal slices with a 1-mm gap), and a T2-weighted fast spin echo image (256 × 256 in-plane resolution, 240 mm FOV, 81 × 2 mm sagittal slices). Functional scans were acquired using a gradient echo EPI sequence (64 × 64 in-plane resolution, 240 mm FOV, TR/TE/Flip = 2000 ms/30 ms/90°, 30 × 4 mm interleaved sagittal slices with a 1-mm interslice gap; 252 whole brain images per scan run for the N-back task, 203 whole brain images per scan run for the Go/No-go task). The signal-to-noise ratio (SNR) of the image data averaged over all subjects and over the entire brain volume was 130 ± 50, with a range of 80–250. This value includes regions of the brain with substantial inhomogeneity (dropout) artifact so there is substantial spatial variation.

Phantom simulations

A series of simulated fMRI data series (also known as phantoms) were created in order to mimic activations with known locations and magnitudes. To create the phantoms, AIR was used to determine the motion parameters for a rigid-body (6-parameter) motion correction for a sequence of 202 images (TR = 2 s) from one of the human subjects, and the transformation parameters were extracted from the resulting "*.air" files. A single EPI-based 3D image volume from the middle of the sequence was selected as the base image. A scriptable version of the AIR program "manualreslice" was used to create transform matrices for each of the 202 known motions, and the AIR program "reslice" was used with AIR's "scanline chirp-z" interpolation method (a Fourier variant) to create a series of translated and rotated images with realistic motion. The original motion parameters and the activation models are shown in Fig. 1.

Before the addition of motion to each phantom series, activations were added at specific locations in the original (pre-motion) image. Seven (7) different activation clusters were added; all activations were modeled as a 3-dimensional Gaussian distribution with a specific magnitude and FWHM. The activation properties are summarized in Table 2.

The activation magnitudes were modulated by a specified activation function. Two different activation functions were used in this study: (i) a block-design activation with 9 evenly spaced blocks (10 TRs on, 10 TRs off) and no activation at the beginning or end; and (ii) an event-related design with a spike-shaped activation every 8 TRs. Both activation functions were convolved with a synthetic hemodynamic response function.

Noise was added to each image volume after addition of the activations. The noise was modeled as normally distributed about the value for each voxel. The FWHM of the normal noise distribution could be specified in order to model different image SNR. In this study, two different SNR levels were modeled: (i) SNR = 200 and (ii) SNR = 80. These capture the range of SNR encountered in previous research using different scanners (1.5 T and 3.0 T) in brain regions with both good and poor signal coverage.

Several potentially influential sources of noise were not modeled in this set of phantoms, including physiological noise, non-linear warping that occurs when the internally distorted magnetic field is altered by rotation, the interaction of motion and susceptibility artifacts (Wu et al., 1997), multiple intrascan movements, and spin history effects (Friston et al., 1996). Intraframe motion was modeled for an interleaved acquisition scheme in some early phantoms but was found not to influence the results, so this aspect was not pursued.

Table 2
Phantom activation properties

Name	Location	FWHM (mm)	Magnitude
Left visual cortex 1	12, 12, 23	8.0	0.01000
Left visual cortex 2	12, 12, 26	5.0	0.01000
Left PFLC	9, 43, 41	6.0	0.00500
Right PFLC	20, 43, 42	6.0	0.00500
Left hippocampus	8, 33, 28	6.0	0.00250
Left motor cortex	9, 33, 40	3.0	0.00375
Right motor cortex	19, 33, 41	3.0	0.00500

"Location" refers to the 0-based location within the 3D array (dimensions = [64, 64, 32] voxels) for the base EPI image oriented in the acquisition direction (sagittal) with the superior part of the brain at the left and the anterior part of the brain at the top of the displayed image volume. "FWHM" refers to the width of the Gaussian-shaped activation in each of the three cardinal directions. "Magnitude" indicates the scaling factor for each activation relative to the regional average value at each respective location, e.g., 0.0100 indicates a cluster with a maximal pre-noise value 1% larger than the base image.

A set of 12 phantom time series were constructed with differing parameters as summarized in Table 3. Three different levels of motion were used: no motion (M0); standard motion as measured from a human subject scan (M1); and the M1 translation and rotation parameters multiplied by 3 (M3). These phantoms are referred to as the "quantitative phantoms". Each package was used to motion-correct the phantoms, and the resulting estimated motion parameters were compared to the known motion values. (BrainVoyager was not included in quantitative comparisons of accuracy

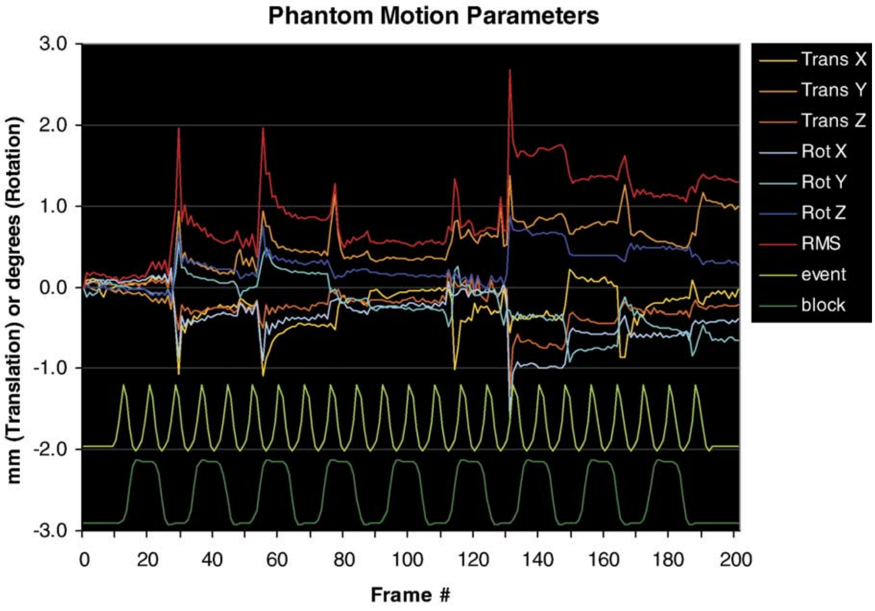

Fig. 1. Phantom motion parameters. Parameters were taken from the motion correction of a representative human subject using AIR. The translation parameters are in mm and the rotation parameters are in degrees. The root mean square (RMS) is the square root of the sum of the square of all 6 parameters. The event-related and block activation models are shown in green at the bottom (arbitrary units).

Table 3
Summary of phantom properties

Activation	SNR	Motion
Block	200	0
Block	200	1
Block	200	3
Block	80	0
Block	80	1
Block	80	3
Event	200	0
Event	200	1
Event	200	3
Event	80	0
Event	80	1
Event	80	3

"Activation" refers to whether the activation paradigm was modeled as a series of blocks or as a well-spaced event-related model. "SNR" refers to the signal to noise ratio in the resulting image, which incorporates the basal signal level and the normally distributed noise. "Motion" refers to a multiplication factor used to increase or decrease the magnitude of the motion parameters: "0" indicates no motion was added (these are the control data); "1" indicates the motion parameters were used without further modification; and "3" indicates that all motion parameters were multiplied by 3, in order to simulate a large amount of motion.

due to difficulty in converting motion parameters to a common coordinate system.) The transforms required to translate the parameters between packages are shown in detail in the online Supplementary Data.

The interpolation accuracy of the software packages was compared using a mostly zero-valued 3D image volume containing nine (9) well-separated single points with values of 1000 (the "interpolation phantom"). One point was at the volume center and the other eight were placed near each corner and 10 pixels in from the volume edges. Four of the software packages (AFNI, AIR, FSL, SPM2) were used to reslice the volume using x-, y-, and z-translations of 2 mm, and rotations of 2° about the x-, y-, and z-axis. Two types of interpolation algorithms were tested for each package: a trilinear type and a sinc-like or Fourier-class algorithm. For SPM2, only a sinc-like algorithm was tested.

Computer facilities

A set of 6 matched computers was used for all analysis (HP model xw6000, 2.66 GHz Pentium IV dual processors, 3.0 Gbyte RAM, 2 network cards, CD/RW, dual boot Windows2000 or RedHat Linux 8.0+ patches). For time-sensitive tests, data were copied to the local disk of a single computer.

Processing of human data prior to motion correction

MRI EPI-BOLD data were reconstructed using "epirecon", a program made available by GE Medical Systems to certain research facilities which use its MRI scanners. After reconstruction, functional data were converted to AFNI format and slice-time corrected using the "to3d" and "3dTshift" programs of AFNI. Image distortion was corrected using estimated B_0 field maps to shift image pixels along the phase encoding direction in the spatial domain (Jezzard and Balaban, 1995; Hutton et al., 2002). Duplicate copies of the corrected functional data were made in ANALYZE-7.5 format for motion correction with packages other than AFNI. The anatomical scans (T1, T2, coplanar) were reconstructed directly on the MRI scanner.

Standard and other motion correction parameters

Each software package was used to motion-correct each of the data sets using parameters that were determined to be the "standard" ones normally used for each package. The standard parameters were taken as those advocated in an instruction manual, an introductory course, or online discussion groups as default parameters, if available; or, lacking these resources, through our own experience with a reasonable balance between accuracy and speed. The authors for each of the packages were then consulted, and they verified that our standard parameters were either in fact the default parameters or were quite similar. If default values were not specified for one or more parameters for a particular package, we attempted to use parameters similar to the other packages; for instance, a sinc-type interpolation was used for all standard reslice approaches. The authors of each of the packages were further consulted for advice on motion correction parameters which were optimized for (i) speed and (ii) accuracy. In some cases these were similar or identical to the standard parameters. The parameters used for each software package and for each of the three conditions (standard, speed, and accuracy) are listed in Table 4. Unless specified, all parameters in Table 4 were default parameters, or the same as the standard configuration. Initial analysis of the phantom data indicated that the interpolation schemes suggested by the authors for the accurate parameters for AFNI (heptic, or seventh-order polynomial interpolation) and AIR (scanline chirp within plane, linear across planes, '−n 11') yielded notably worse results for cluster size and t values than the standard parameters. The accurate parameters reported here are thus the same as the standard parameters for these two packages. All combinations used a 6-parameter (rigid body) model.

GLM analysis

The program "fmristat" (Worsley et al., 2002) was selected for the general linear model (GLM) analysis because it has no motion correction capabilities itself, and so was not one of the packages considered in this work. The motion-corrected time series from all software packages were treated identically after motion correction was complete, and no further processing (such as spatial or temporal filtering) was performed.

Using a two-stage GLM approach, analysis with fmristat yielded a statistical parametric map of t values for each package. In the first stage GLM analysis, data from each human subject were individually modeled. Each condition was modeled with a regressor formed by convolving a boxcar or spike function representing stimulus "on" periods with an ideal hemodynamic response function. Contrast maps were generated by subtracting the parameter estimates for pairs of conditions. The contrast maps were then registered to an MNI template (Montreal Neurological Institute; Evans et al., 1992, 1994) with FLIRT software, using a two-stage registration procedure. First, the coplanar T1 volume was registered to the high-resolution T1 volume, using a 6-degree of freedom affine (rigid body) transform. The high-resolution T1 volume was registered to the MNI template using a 12-degree of freedom affine transform. These two transforms were then combined and applied to the contrast maps to bring them into MNI space.

Table 4
Standard, fast, and accurate parameters for each software package

	Standard	Optimized for speed	Optimized for accuracy
AFNI	interpolation = Fourier; max. number of iterations for convergence = 19; x_thresh = 0.02; rot_thresh = 0.03; delta = 0.7; single pass registration	interpolation = cubic	same as standard
AIR	interpolation = scanline chirp; cost function = least squares with intensity rescaling; thresholds = 6000; set flags "−j", "−q"	interpolation = trilinear; cost function = least squares; convergence = "−c 10"; sampling density = "−s 81 9 3"	same as standard
BrainVoyager	interpolation = trilinear; every other voxel in each dimension (i.e., 12.5% of voxels), maximum of 100 iterations	same as standard	interpolation = sinc; every voxel, threshold = 100 (of 255)
FSL	cost = normcorr; bins = 256; dof = 6; refvol = 0; scaling = 6.0; smooth = 1.0; stages = 3	same as standard	refvol = N/2 + 1; stages = 4; final interpolation = sinc
SPM2	quality =.75; fwhm = 5; sep = 4; interp = 4	quality = 0.5; fwhm = 7; sep = 6; interp = 1	quality = 1.0; fwhm = 2; sep = 2; interp = 5

Parameters not mentioned in fast and accurate categories are the same as standard parameters. For AFNI and AIR, the accurate parameters originally suggested by the software authors were less accurate than the standard parameters, which were used instead throughout this work.

For the human data, a second stage GLM analysis was performed with fmristat to produce group t statistic maps for each contrast. In preliminary analyses, it was noticed that the relative ranking of results from the different packages changed slightly depending on group size. This effect is most likely due to differing accuracies between packages for certain individuals, so the overall rankings changed depending on the constituent subjects. Rather than rely on the results for a single group of individuals, a series of subsets of the full data set were examined to avoid biasing the results in favor of the motion correction program which happened to yield the best t statistics for a particular group. From the original full group of 40 subjects, 7 subjects with excessive motion (>4 mm or 4°) were discarded. The remaining 33 subjects were evenly divided into low- or medium-motion categories based on a median split of the average root mean square (RMS) of the percent signal difference between each frame and the first frame. Averaged over each category, the RMS difference was as follows: low/event: 1.01%; medium/event: 1.31%; low/block: 1.34%; medium/block: 1.52%. Two distinct series of nested subgroups were then created containing 8, 12, 16, 20, 24, and 33 subjects to avoid idiosyncrasies in a given group biasing the results. Each subgroup contained all of the subjects from the immediately smaller group along with several additional subjects. For each subgroup, the number of low- and medium-motion subjects was balanced, in order to maintain continuity between group sizes with respect to motion magnitude. For every package and set of parameters, a GLM analysis was performed for each group size for both nested series. The results from each group were averaged across the two series.

Criteria for assessing accuracy of MC

Humans

For the human subject data, results were compiled for several clusters for each paradigm. The clusters were selected by first running the GLM analysis for the entire group of $N = 33$ subjects using the non-motion-corrected data, and thresholding the resulting t maps at $t = 2.0$ (corresponding to $P < 0.054$ uncorrected). The non-motion-corrected data were used as the basis for comparison in order to see the relative improvements of motion correction for each software package, as well as to include both actual and artifactual activation clusters in the comparison. Several clusters were selected after visual inspection; 6 clusters were selected for the event-related task and 5 were selected for the block design task (see Table 5). Variety in the clusters was sought with respect to location, cluster size and shape, superficiality, magnitude of maximal t value, and likeliness of the cluster being a motion artifact (Fig. 2 shows representative examples). For the event-related task, clusters were identified bilaterally in temporoparietal cortex, inferior frontal gyrus, anterior cingulate, and left motor cortex. These brain regions have previously been found to make up part of a network involved in manual response selection and inhibition (e.g., Liddle et al., 2001). For the block design task, clusters were identified bilaterally in the middle and inferior frontal gyri, as well as bilateral middle and superior temporal gyri, consistent with previous research on working memory tasks similar to this one (Cohen et al., 1997). In addition, a left prefrontal cluster

Table 5
Details of the six clusters examined in the event-related (Go/No-go) and five clusters in the block design (N-back) data sets

Cluster location	Design	Cluster volume (mm^3)	Cluster maximum t
Anterior cingulate	event	8200	5.28
Left parietal	event	4000	4.81
Right temporoparietal	event	4800	4.41
Left inferior frontal	event	4400	5.06
Right inferior frontal	event	4000	6.06
Motor cortex	event	5900	4.56
Left middle frontal	block	3900	4.91
Left temporal	block	4900	4.85
Right temporal	block	3400	5.28
Right inferior frontal	block	300	3.78
Left frontal (motion artifact)[a]	block	3600	5.66

[a] Details for the motion artifact cluster are based on non-motion-corrected data.

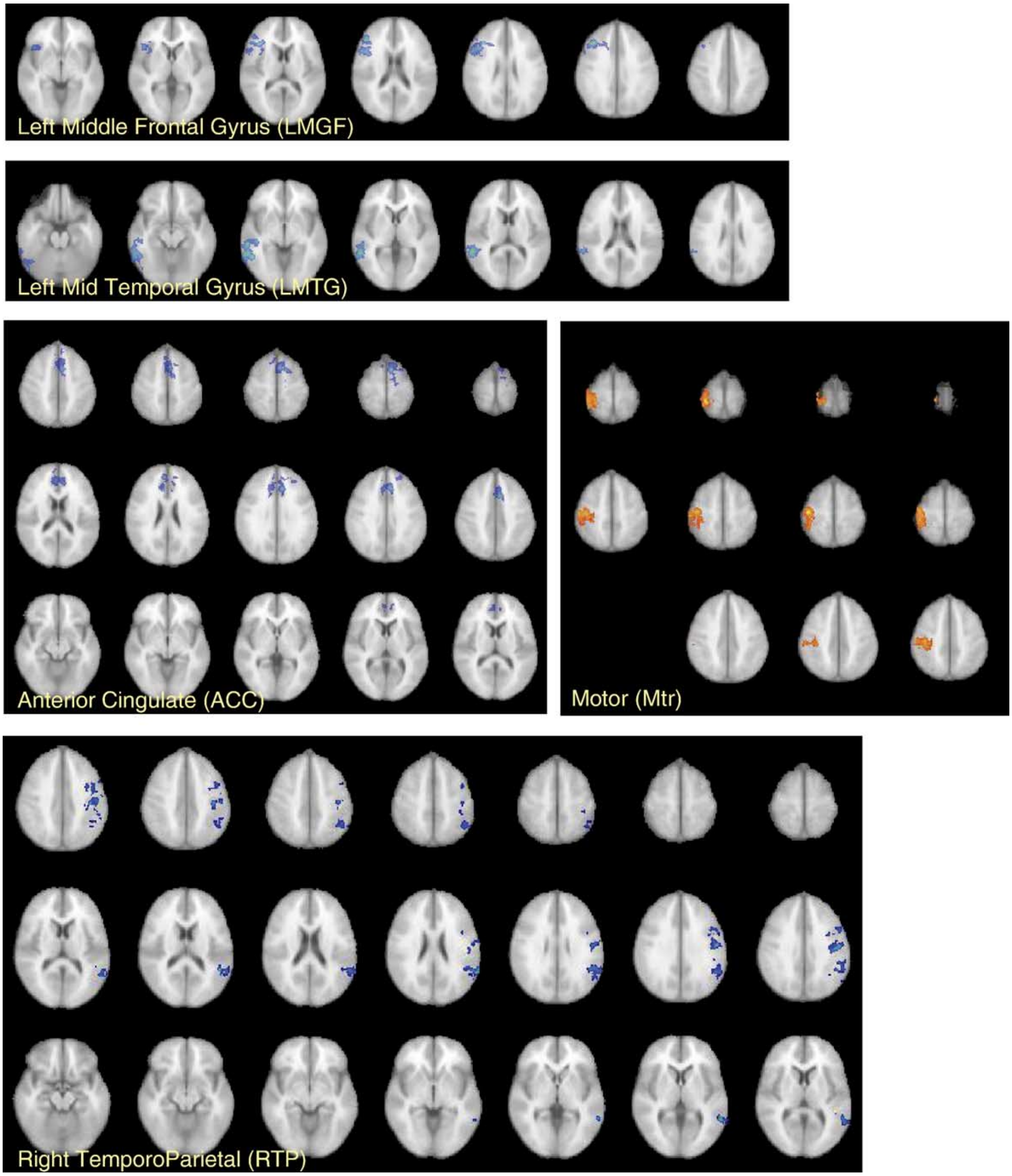
Left Middle Frontal Gyrus (LMGF)
Left Mid Temporal Gyrus (LMTG)
Anterior Cingulate (ACC)
Motor (Mtr)
Right TemporoParietal (RTP)

Fig. 2. Representative significant activation clusters for the block design (LMFG, LMTG) and the event-related design (ACC, Mtr, RTP). The clusters were selected in order to obtain a variety of shapes, volumes, and locations.

was identified as an obvious motion artifact, running in a narrow strip around the left anterior cortical surface (see Fig. 4). A binary mask was made for each selected cluster for $N = 33$ subjects.

Summary statistics for the motion-corrected clusters included cluster location (centroid and maximum value position) in MNI coordinates, mean, maximum, and standard deviation of t values within each cluster, and the number of voxels within each cluster for which the t value exceeded a significance of $P = 0.01$. The parameters which were found to be most unique and informative were the change in maximum t value across the whole cluster relative to non-motion-corrected data, and the change in the number of voxels exceeding $P = 0.01$ relative to non-motion-corrected data.

Phantoms

The phantom data set was used to examine the effect of motion correction on subsequent maximal t values for GLM-derived activation images. The maximal t values were tabulated for each of the seven (7) activation clusters within each phantom. A recovery coefficient was calculated by comparing each software tool's clusters to a reference data set with the same SNR but which had no added motion, and the same activation as the cluster being compared to it (Eq. (1)).

$$R = (\max\_t_{\text{cluster}} - \max\_t_{\text{NoMC}}) / (\max\_t_{ref} - \max\_t_{\text{NoMC}}) \qquad (1)$$

where $\max\_t_{\text{cluster}}$ is the maximum t value from an individual cluster, $\max\_t_{\text{NoMC}}$ is the maximum t value for that cluster from the corresponding non-motion-corrected phantom, and max_t_{ref} is the maximum t value from that cluster from the unmoved phantom with the same amount of noise. The recovery coefficient indicates the improvement of the maximal cluster value for motion correction relative to no motion correction, as a fraction of the total drop in t value attributed to motion. $R = 1$ indicates full recovery of t values, and 0 indicates no benefit from motion correction.

To evaluate the absolute motion estimation accuracy using the quantitative phantoms, the transform matrix output by each software package was transformed to a standard coordinate system (see Supplementary Data for details). For practical reasons the FLIRT coordinate format was chosen, but any other could equally have been chosen. The FLIRT program *rmsdiff* calculates the RMS difference between two transforms applied to an 80-mm sphere located at the center of the image (see Jenkinson et al., 2002). This program was used to compare the estimated motion transform for each time point as given by each software package with the known motion transform.

To assess interpolation accuracy, the average of the maximal values for each of the nine isolated points in the resliced interpolation phantom was used as an indicator, with values closest to the original value indicating a more accurate interpolation algorithm.

Criteria for assessing other aspects of software

To assess the speed of motion correction, 5 representative subjects were selected, and motion correction was performed on this group using each package. A single computer running Linux (RedHat-8) was used for all computations except BrainVoyager, for which the same computer was booted using the Windows2000 operating system. All data were first copied to that computer's local hard drive. For BrainVoyager only, the time to read the data from disk was not included in the timing total. The times required to motion-correct each subject were recorded for the two event-related scan runs and the block-design run. For some packages, separate times were obtained for alignment and reslicing, but since these times were not available for all packages, only the aggregate times are reported. An average time per single frame was calculated, and these were then averaged over the block and event-related data sets and all 5 subjects for each package.

To assess the ease of writing scripts for batch processing, the principle users of each of the software tools within this work compared each tool they were familiar with to other familiar tools, and other users both within and outside of our laboratory were also

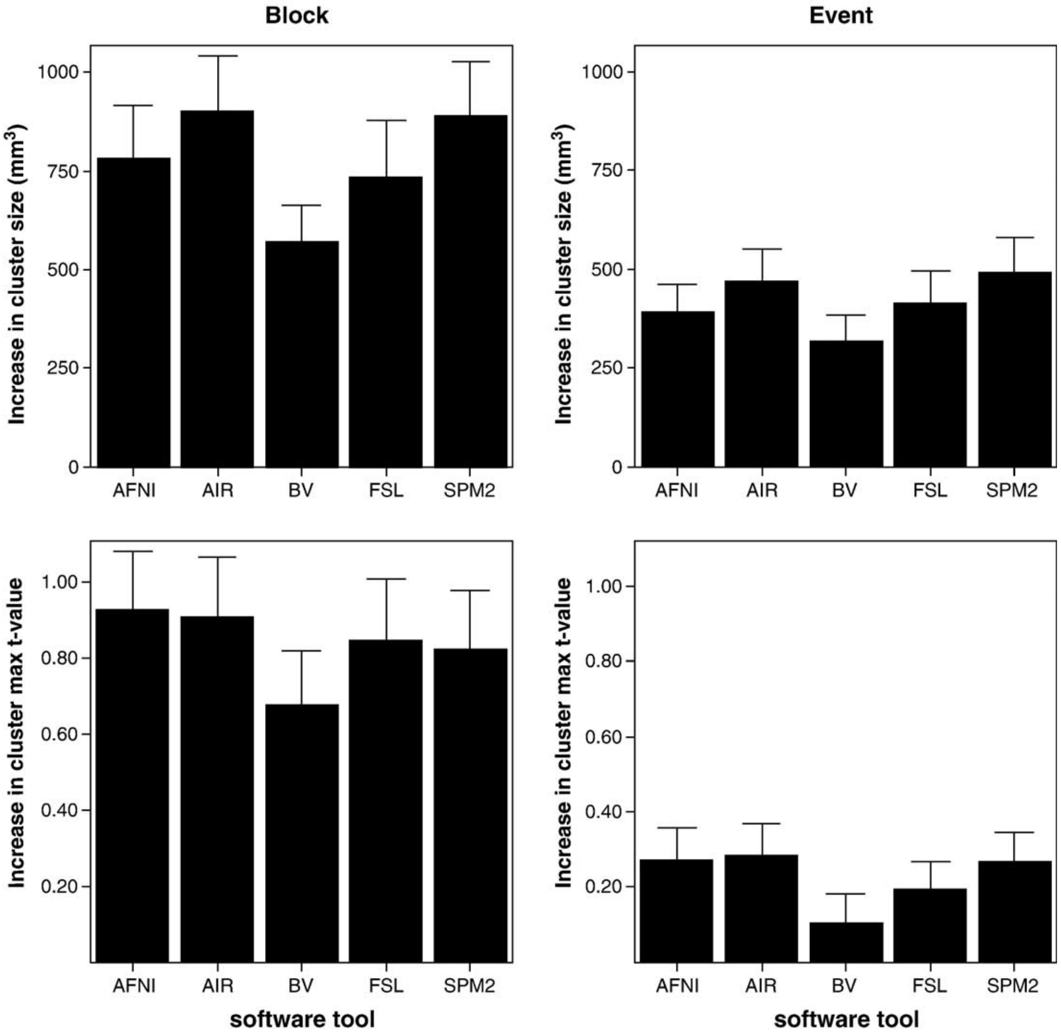

Fig. 3. Increase in cluster size (top) and maximum t statistic (bottom) for motion-corrected data from the block design task (left) and event-related task (right). Data are averaged across clusters and group sizes. Error bars represent 95% confidence intervals. Non-motion-corrected block design data had a mean cluster size of 632 mm^3 and max t statistic of 4.03. Non-motion-corrected event-related data had a mean cluster size of 1640 mm^3 and a max t statistic of 4.92.

consulted. All contributors to this somewhat subjective assessment had experience with at least two of the software tools, and none of the contributors had any particularly close association with any of the tools (i.e., the authors of the tools were not consulted for this aspect). Although in some cases many comments were obtained, we have attempted to restrict the observations to those that are pertinent to the motion correction features of each package.

Results

Human data

Comparisons of the motion-corrected *t* values and cluster sizes for each software tool relative to the corresponding non-motion-corrected values are shown for the block-design and event-related experiments (Fig. 3). Although the thresholded cluster size and maximum *t* statistic results depended upon group size for both the block and event-related designs, the pattern of results across software tools and motion correction settings (standard, fast, accurate) was the same for all group sizes. Statistical *F* tests of the group × tool and group × settings interactions were not significant (*F* tests <1.0 for these interactions). Motion correction settings had no effect on results, either as a main effect or in the interaction with software tool (all *F*'s <1). Fig. 3 thus presents results averaged across group size and software settings. For the block-design task, there was a significant difference across software tools in thresholded cluster size ($F(4,90) = 2.52$, $P = 0.046$) and in cluster maximum *t* statistic ($F(4,90) = 15.81$, $P < 0.0001$).

All motion correction tools provided an increase in the cluster maximum *t* statistic of about 0.8 in the block design experiment, which represents an increase of 20%. All tools performed similarly, although BrainVoyager increased *t* statistics by slightly less than the other tools. For changes in thresholded cluster size, all tools provided a mean increase in cluster size of about 100 voxels, which represents more than a 100% increase in cluster size over non-motion-corrected data. Once again, BrainVoyager provided a slightly smaller gain, and AIR and SPM2 a slightly higher gain in cluster size than the other tools.

Advantages of motion correction for the event-related design were more modest than for the block design. On average, the different tools provided an increase in the cluster maximum *t* statistic of 0.2, representing a 5% increase over non-motion-corrected data. Increases in thresholded cluster size were of the order of 50 voxels, an increase of 25% relative to non-motion-corrected data. There was a significant difference across software tools in cluster maximum *t* statistic ($F(4,90) = 3.52$, $P = 0.007$) and thresholded cluster size ($F(4,90) = 3.75$, $P = 0.005$). As with the block-design task, BrainVoyager increased *t* statistics by slightly less than the other tools. SPM2 and AIR provided slightly higher gains in cluster size than the other tools.

The levels of change in cluster maximum *t* statistic due to motion correction relative to the original maximal *t* values were substantial for the block design, but not the event-related design. Observed changes in cluster size were substantial for both tasks, and large enough to have an impact on the statistical analysis, since cluster size is an important component in most group level statistical thresholding techniques. It is worth noting that the FWHM of spatial correlation in motion-corrected data sets (a measure of spatial smoothness) varied by less than 0.1 mm across tools, and by less than 0.3 mm within each tool and across settings (e.g., standard vs. fast); this implies that differences in maximal *t* values were not due to differences in spatial smoothing.

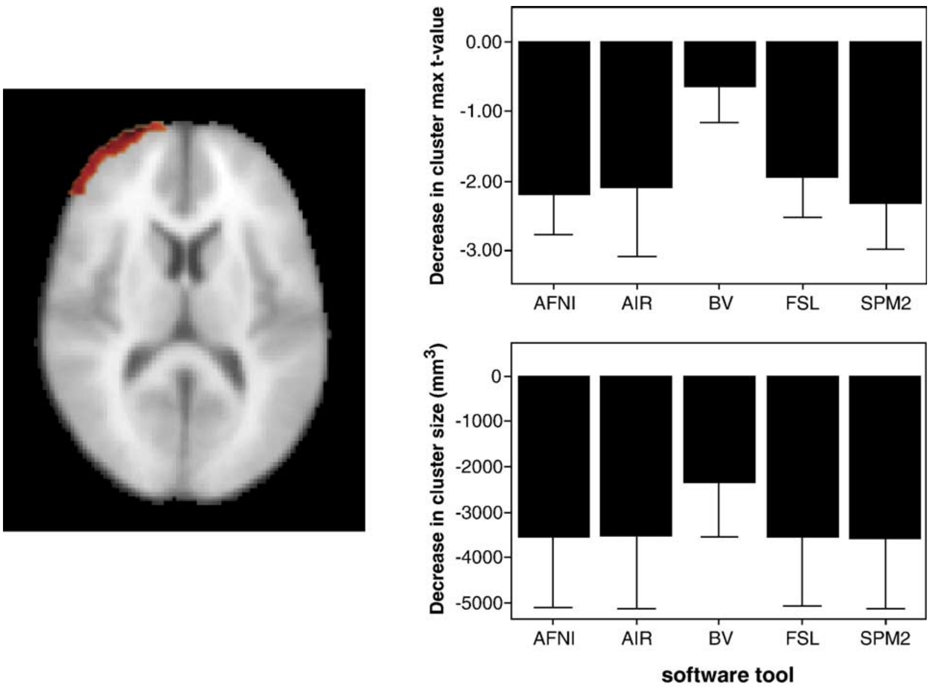

Fig. 4. Decrease in maximum *t* statistic (top right) and cluster size (bottom right) for the motion artifact-related activation cluster in the block design task (left). Data are averaged across clusters and group sizes. Error bars represent 95% confidence intervals. This artifactual cluster had a size of 3672 mm^3 and max *t* statistic of 5.66 in non-motion-corrected data.

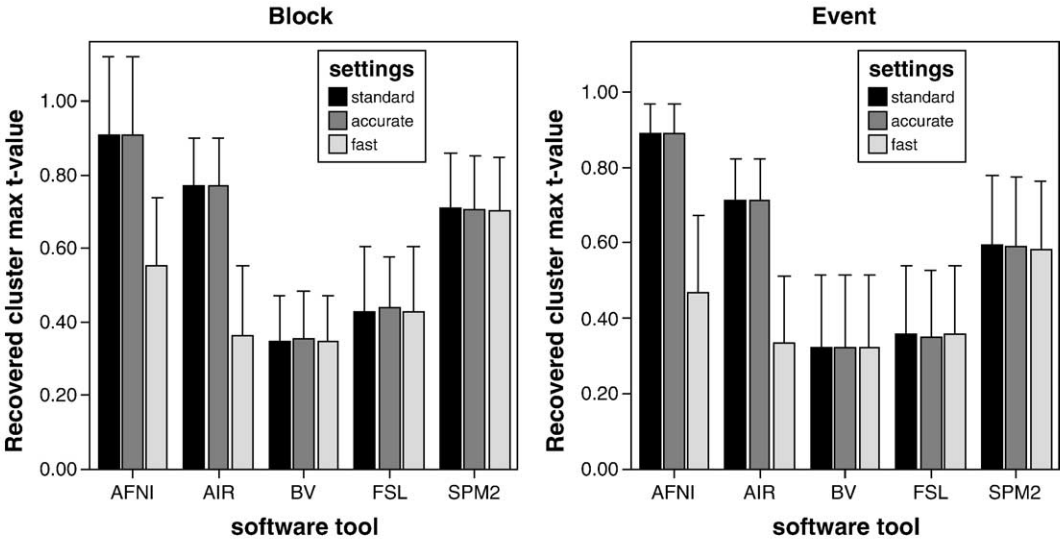

Fig. 5. Recovery in maximum *t* statistic for motion-corrected data from the block design phantom (left) and the event-related design phantom (right). Data are averaged across clusters. Error bars represent 95% confidence intervals.

A false activation attributed to a motion artifact was analyzed as above. The cluster was located in the left prefrontal region (see Fig. 4). All packages reduced this artifactual effect, although with the BrainVoyager motion correction some residual artifact remained at the suprathreshold level.

Phantom GLM data

For the simulated phantom data, more distinct differences than in the human data were found between motion correction tools. The ability of each package to recover cluster maximum *t* statistics by motion correction is demonstrated in Fig. 5, which shows the proportion of the drop in maximum *t* statistic due to motion that is recovered after motion correction (see Eq. 1). No significant interaction was found between software tool and effect of noise level or motion level, so the results in Fig. 5 are averaged over these factors. For the block design, all tools led to improvement over non-motion-corrected data. A significant effect of software tool on recovered maximum *t* statistic ($F(4,360) = 16.83$, $P < 0.0001$) was due to greater recovery for AFNI, and lower recovery for BrainVoyager and FSL. Results for event-related phantom data were similar, with AFNI showing the greatest *t* statistic recovery, and FSL and BrainVoyager the least (test for main effect of software tool: $F(4,360) = 15.51$, $P < 0.0001$).

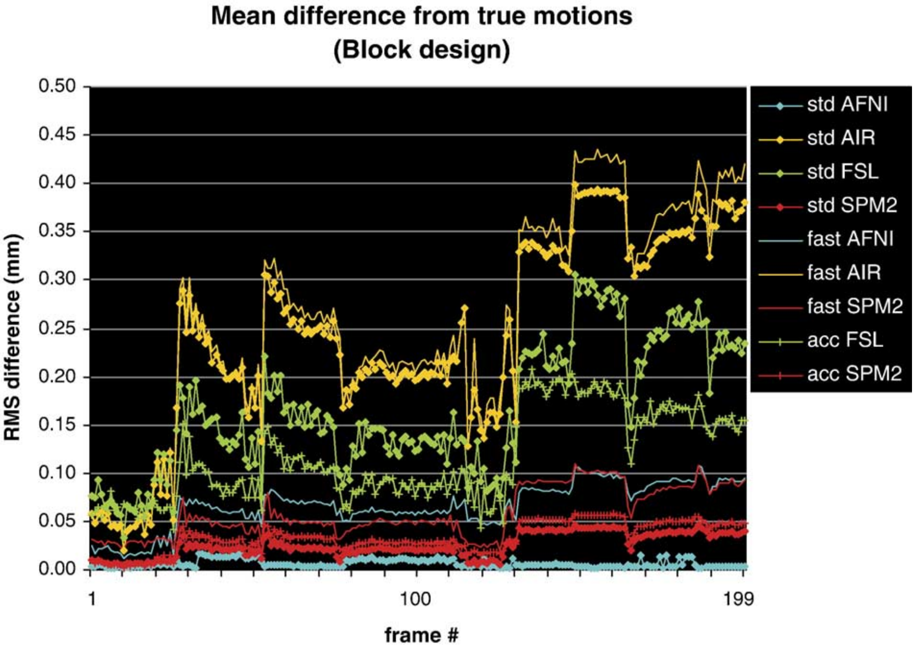

Fig. 6. Motion estimation accuracy. The absolute value of the mean difference of the calculated motion correction parameters (see Methods section G.2) is plotted for each time frame. The FSL "fast" parameters and the AIR and AFNI "accurate" parameters are not shown since they are identical to the corresponding standard ones.

There was an interaction of software tool and motion correction settings for both the block design ($F(8,360) = 2.31$, $P = 0.02$) and event-related design ($F(8,360) = 2.59$, $P = 0.009$) phantoms. This effect was due to maximum *t* statistic recovery being worse with the fast settings for AIR and AFNI. Each of the software packages yielded nearly identical GLM results between the standard and accurate parameters.

Estimation accuracy

The motion estimation accuracy for the phantom data is shown in Fig. 6. The most accurate results were achieved by AFNI, followed by SPM2. The accuracy of AIR, while still acceptable (less than 10% of a voxel dimension), was the poorest. The accuracy of AIR was affected very little by using either the fast or accurate parameters, while for FSL, the accurate parameters did indeed improve the accuracy by approximately 30%. Interestingly, for both AFNI and SPM2, the standard parameters were the most accurate.

Interpolation accuracy

The interpolation results indicate that the sinc-like interpolation algorithms are significantly more accurate and introduce less smoothing than trilinear-type algorithms (Fig. 7), since the sinc-like algorithms yield maximal values closer to the original starting value. AIR, FSL, and SPM2 are all quite similar within each algorithm type, while AFNI performed noticeably better for both types.

Speed comparison

Each of the software packages was operated using standard (default) parameters, and then using parameters optimized for speed and accuracy. For the standard parameters, the average times (in seconds) per 3D image were as follows: AFNI: 0.11; AIR: 0.36; BrainVoyager: 0.26; FSL: 1.93; SPM2: 1.35 (see Fig. 8). AFNI was at least twice as fast as any other package, with AIR and BrainVoyager ranked next. For the accurate parameters FSL and SPM2 yielded extremely long times (off of the chart in Fig. 8). However, the times for the accurate parameters are not very important, since little or no improvement in measured accuracy or cluster *t* value was actually derived from using these parameters.

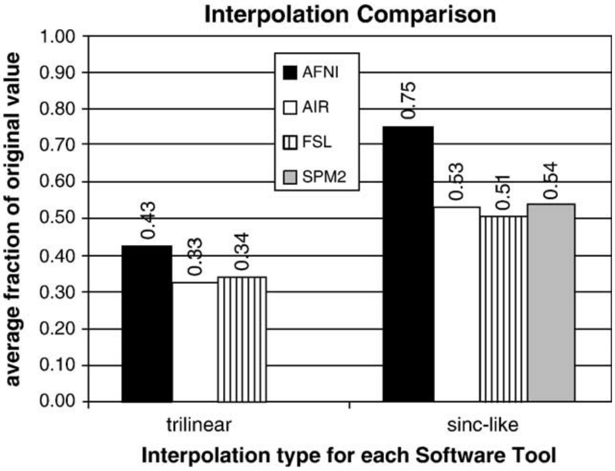

Fig. 7. Interpolation comparison using a resliced volume with isolated points. The average of the maximal value of each of the nine points was calculated. The starting value (1000.0) of the points was scaled to 1.0 in this plot. Values closer to 1.0 indicate less smoothing by the interpolation algorithm. A trilinear interpolation was not performed for SPM2.

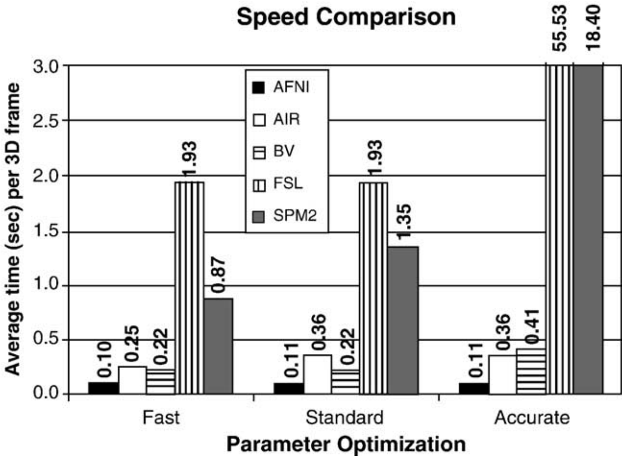

Fig. 8. Comparison of speed of motion correction. The average time (in seconds) per 3D frame is shown for each software package for the fast, standard, and accurate parameters. The numbers above the bars indicate the average times. The values for FSL and SPM2 for the accurate parameters exceed the range of the plot.

Usability

During the use of each of the packages in this work, comments were recorded about specific usability features such as the learning curve, file formats, ease of writing scripts for batch mode, and data display features for checking the motion parameters. Although the learning curve may be steep for several of the packages, there are very helpful courses available from several of the software authors for their packages, as well as a variety of excellent tutorials and manuals. As mentioned previously, most of the comments about file format issues will be obviated if each of the software tools supports the proposed NIfTI file format (see http://nifti.nimh.nih.gov/) as either a primary or secondary format. Unless specifically mentioned, each tool was able to provide an adequate motion correction for all data sets.

AFNI

AFNI has both command line programs and an extensive interactive GUI with which to analyze and examine data. All programs are easily integrated with any scripting language or environment. Many of the command line programs (including the motion correction program, "3dvolreg") are also included as AFNI plugins, with a graphic user interface. "3dvolreg" has options to save the motion estimates in text format, making it easy to graph motion estimates against time, or to include motion correction estimates as regressors in subsequent first-level GLM analysis. The learning curve for AFNI can be steep, since there is a large array of options and programs, and many of the handy shortcuts are not immediately apparent. Relatively few processing steps are hidden from the user or invoked automatically, but conversely the user must take care to specifically include all of the desired analysis components.

AIR

AIR has only a rudimentary GUI interface and display capability and was designed from the beginning to be incorporated into scripts, which is easy to do. It is not trivial to obtain the motion parameters, which are stored in a series of non-text motion parameter files (so-called "*.air" files). A program to sequentially extract and convert the transform matrix from each file must be created independently by the user. Other items unaddressed by AIR include the following: (i) the lack of integrated display programs; (ii) the necessity to match the compiled version of AIR with the data type and range; (iii) lack of support for floating-point data; and (iv) no support for 4D file format. Nevertheless, learning to use AIR is relatively easy, and it has proven to be a robust tool.

BrainVoyager (BV)

BrainVoyager is relatively easy to use on its own, although issues related to file format and a rigid directory structure make it more difficult to incorporate into a processing pipeline. Copies of the original data files are required to be located in the directory where it will create its "fmr project" file. Thus, a script must first copy the input files to the BV working directory, perform all the BV work, and finally delete these files. In the latest version, BV introduces the ability to write data into the ANALYZE-7.5 format for utilization by other software. There is a difference in the availability of parameters between the interactive menus and a scripted analysis, so not all motion correction parameters can be scripted. For instance, even though motion correction can be referenced to any volume when run interactively, scripting only allows the default correction to the first volume. The voxel intensity threshold for motion correction is also not scriptable. However, the BV programmers are generally quite accommodating to requests for adding features such as these. Motion corrections on a few of the data sets (5 of 120) could not be completed due to errors in the algorithm, specifically problems with the covariance matrix. No combination of parameters could be found which permitted motion correction for these five data sets.

FSL

FSL has both command line programs and GUI interfaces for most analysis operations, although not all command options are available through the GUI. When running a command from the GUI, the equivalent command line is output to the terminal, providing a basis for subsequent command line use. When carried out within the FSL fMRI analysis tool FEAT, motion correction estimates are output in easy to read HTML format. Text files of motion estimates and transform matrices are also easily generated. FSL includes scripts to facilitate the use of AFNI's display capabilities with FSL-generated output.

SPM2

The choice of Matlab as the environment for SPM2 has good and bad points, and several of the usability features stem from this. It is difficult to incorporate SPM2 procedures into a larger non-Matlab-based script, but once inside Matlab there are a variety of scripting options. The level of skill required to write scripts for motion correction is comparable to most other software tools. Although SPM2 can read a 4D ANALYZE-7.5 file, SPM2 must be launched from a directory that contains a 4D file, and the ability to read a 4D file must be explicitly requested. The motion correction parameters are saved in a file in the output directory and can be easily incorporated as regressors in a subsequent GLM analysis.

Discussion

The human as well as the phantom data indicate a substantial benefit to a GLM analysis from motion correction. The human data examined realistic changes in cluster size and t statistics and the ability to remove a motion-induced artifactual activation. All packages did an acceptable job and no single package emerged as the clear-cut leader, although BrainVoyager did not perform quite as well as the others. For the phantom data, BrainVoyager and FSL did not achieve the same level of registration accuracy as the others, as indicated by increased t statistic values. However, overall differences between packages are relatively small.

For the human data, a larger benefit of motion correction was seen for the block-design data compared to the event-related design. This difference may be due to the greater temporal match of brain hemodynamics related to activation between a block design and subject motion. Task-correlated motion is likely to be more of a problem in block designs than rapid event-related designs, since in the latter case task-correlated motion unfolds at a more rapid pace than the relatively sluggish hemodynamic response (Birn et al., 1999; Field et al., 2000). Thus, in block designs motion correction is likely to be particularly important.

The results from human data and phantom data were very similar, although there were greater performance differences between software packages for the phantom data than for the human data. The results from the human data can be regarded as more representative of realistic data analysis scenarios, in part because a large number of different motion sequences were present in the human data, and also because several known sources of noise were not modeled by the phantom. The phantom's model for noise is likely inadequate to capture the complexity of actual human data, since the phantom model did not include physiological noise (breathing, blood pulsatile motion, etc.), spin history effects, of the effect of motion on magnetic field distortion. Future work using this or a similar phantom incorporating additional noise elements into the model is likely to change the magnitude of the t value recovery and registration accuracy results, and perhaps also the relative performance ranking of the various packages.

Somewhat disappointingly, there was virtually no improvement in the t values or cluster sizes when the motion correction parameters were optimized for accuracy (Fig. 5), although the accurate parameters were more time consuming for all software tools (Fig. 8), with a 10-fold or more increase in processing time for some tools. A variety of parameters were changed in an attempt to increase accuracy, with most involving an increased number of iterations and/or a more stringent convergence criteria required of the cost function (see Table 4). BrainVoyager and FSL also attempted a more sophisticated interpolation scheme. Freire et al. (2002) point out that some types of cost functions may be more robust than others. However, changing the available parameters for the cost functions had relatively little effect. All of the tools examined in this work use a variant of optimizing an intensity-based cost function at several sampling density levels. As discussed by Jenkinson et al. (2002), this optimization strategy is prone to being trapped in a local minimum, so the implementation is arguably one of the most important factors in motion estimate accuracy. The Powell variant used by AIR is insensitive to changes in available parameters and seems unable to avoid the same local minima, as AIR arrived at the same result regardless of the parameters. FSL has perhaps the most elaborate optimization approach but this did not translate into better accuracy. In general, increased estimation accuracy in other tools did not lead to improved *t* values or cluster size in the human data.

Given the range of results for motion estimation accuracy (Fig. 6) and interpolation accuracy (Fig. 7) from the phantom data, it was somewhat surprising that a larger difference between packages did not emerge in the human data. For example, even though the reported estimation accuracy for AIR is a factor of 10 or more less accurate than AFNI or SPM2, the actual performance of AIR for the phantom and human GLM results compared quite favorably. The limited resolution of fMRI data may constrain the gains that can be made by sub-voxel improvements in motion estimation. Another limitation is the assumption by all packages of rigid body motion: once the first-order effect of average interframe motion is detected and corrected, the remaining intraframe motion and other non-modeled effects may be similar for all packages and contribute a baseline variance to the GLM analysis that limits further gains. A more important factor is that the variance introduced into a group-level GLM analysis by intersubject differences in brain anatomy and function may overwhelm any small gains made by a more accurate motion correction algorithm. In this study, there was no spatial smoothing of individual or group data. In the majority of studies where data are smoothed, subtle differences in aspects of motion correction such as estimation accuracy and interpolation fidelity would have an even smaller impact. Once an acceptable motion correction is achieved, further minor methodological tweaks do little to improve the results in groupwise statistical analysis.

It should be reiterated that the human subjects involved in this work were considered to be “normal”, i.e., the population from which they were drawn typically contributes the control sample for many experiments. Different types of subject groups can be expected to have different intrascan motion properties (Seto et al., 2001), which could in turn lead to different results for motion correction comparisons. For example, it is possible that a motion correction technique that is optimal for typical small intrasession movements might be less robust to larger movements or intersession head position differences than a more generally applicable registration technique. For large intersession differences, some of the motion correction tools may be inappropriate due to a limited search space.

Parameters optimized for speed did not appreciably speed up the operation, and this approach yielded a decline in the *t* values and cluster sizes for AIR and AFNI. Given this, it would be imprudent to use the fast parameters, since they are likely to be less robust over a wider variety of subject populations and for larger movements.

AFNI, which was clearly the fastest of the motion correction tools, owes its fast performance time to a 3D implementation based on the work of Paeth (1986) that provides the basis for obtaining equivalent rotations using 1D shear movements, which are much more computationally efficient than 2D or 3D rotations. In essence, motion correction algorithms iteratively reslice an image volume using a series of translations/rotations and compare each result to a target file. Since the vast majority of the algorithm is typically spent on the reslice operation, improvements to this algorithm can be expected to have the greatest impact on speed.

Based on the results of this paper, general recommendations for motion correction parameters include (i) an optimization algorithm that avoids local minima and performs enough iterations to assure a good convergence; and (ii) a higher-quality interpolation (e.g., sinc-like) for motion estimation as well as final reslicing. These may seem obvious but bear mention, since in the two packages where fast parameters yielded decreased *t* values (AFNI and AIR), a trilinear interpolation was used and the convergence criteria were weakened. It is worth noting that AFNI, which was found to be the most accurate, uses the least distorting interpolation scheme and furthermore uses this scheme for the entire optimization sequence. It is also worth pointing out that the increased spatial smoothing associated with trilinear interpolation did not improve the GLM results; in fact, interpolation algorithms with less smoothing (e.g., sinc-like) yielded higher *t* values and larger cluster sizes than trilinear interpolations. Conversely the minimal smoothing characteristics of AFNI’s interpolation did not translate to correspondingly higher *t* values compared to other sinc-like interpolations. This perhaps reflects a practical limit on the impact of improved interpolation algorithms, which can depend on the SNR and spatial smoothness of the data.

In terms of ease of use, motion correction was straightforward to use for all packages. Writing batch scripts was also relatively easy for all packages, although the use of proprietary software environments or file formats can be a hurdle to combining several programs into an analysis pipeline. Important factors which facilitate use include clear instructions, stand-alone command line programs callable from any scripting or programming language, and a flexible and open file format.

Conclusion

This work was inspired by the hypothesis that the choice of motion correction software should be based on identifiable differences in performance, and not on what software the person at the neighboring desk was using when a researcher started learning to perform motion correction. However, the data presented in this work indicate that quantifiable differences in algorithm performance (e.g., registration accuracy, interpolation fidelity) do not necessarily predict differences in GLM analysis results. The effect of motion correction on GLM results is quite similar between software tools, so this criterion is relatively unimportant in choosing a software tool for motion correction. Other factors, such as processing speed, ease of use, and integration with other fMRI processing software, can be used to guide the selection. In the end, selecting a software package that is well supported locally may be the most compelling reason for choosing a particular motion correction software tool.

Acknowledgments

This work was funded by NIMH R01-MH067167, and by NIMH grant P50-MH069315 to RJD. Thanks to Micheal Anderle and Ron Fisher for assistance with data acquisition.

Appendix A

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2005.05.058.

References

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1997. Empirical analyses of BOLD fMRI statistics: II. Spatially smoothed data collected under null-hypothesis and experimental conditions. NeuroImage 5, 199–212.

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The variability of human, BOLD hemodynamic responses. NeuroImage 8, 360–369.

Ardekani, B.A., Bachman, A.H., Helpern, J.A., 2001. A quantitative comparison of motion detection algorithms in fMRI. Magn. Reson. Imaging 19, 959–963.

Birn, R.M., Bandettini, P.A., Cod, R.W., Shaker, R., 1999. Event-related fMRI of tasks involving brief motion. Hum. Brain Mapp. 7, 106–114.

Biswal, B.B., Hyde, J.S., 1997. Contour-based registration technique to differentiate between task-activated and head motion-induced signal variations in fMRI. Magn. Reson. Med. 38, 470–476.

Casey, B.J., Cohen, J.D., O'Craven, K., Davidson, R.J., Irwin, W., Nelson, C.A., Noll, D.C., Hu, X., Lowe, M.J., Rosen, B.R., Truwitt, C.L., Turski, P.A., 1998. Reproducibility of fMRI results across four institutions using a spatial working memory task. NeuroImage 8, 249–261.

Ciulla, C., Deek, F.P., 2002. Performance assessment of an algorithm for the alignment of fMRI time series. Brain Topogr. 14 (4), 313–332.

Cohen, J.D., Perlstein, W.M., Braver, T.S., Nystrom, L.E., Noll, D.C., Jonides, J., Smith, E.E., 1997. Temporal dynamics of brain activation during a working memory task. Nature 386, 604–608.

Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29, 162–173.

Cox, R.W., Jesmanowicz, A., 1999. Real-time 3D image registration for functional MRI. Magn. Reson. Med. 42, 1014–1018.

Dale, A.M., Buckner, R.L., 1997. Selective averaging of rapidly presented individual trials using fMRI. Hum. Brain Mapp. 5, 329–340.

Evans, A.C., Collins, D.L., Milner, B., 1992. An MRI-based stereotactic brain atlas from 300 young normal subjects. Proceedings of the 22nd Symposium of the Society for Neuroscience, Anaheim, CA, p. 408.

Evans, A.C., Collins, D.L., Neelin, P., MacDonald, D., Kamber, M., Marrett, T.S., 1994. Three-dimensional correlative imaging: applications in human brain mapping. In: Thatcher, R.W., Hallett, M., Zeffiro, T., John, E.R., Huerta, M. (Eds.), Functional Neuroimaging: Technical Foundations. Academic Press, San Diego, pp. 145–162.

Field, A.S., Yen, Y.-F., Burdette, J.H., Elster, A.D., 2000. False cerebral activation on BOLD functional MR images: study of low-amplitude motion weakly correlated to stimulus. Am. J. Neuroradiol. 21, 1388–1396.

Freire, L., Mangin, J.F., 2001. Motion correction algorithms may create spurious brain activations in the absence of subject motion. NeuroImage 14 (3), 709–722.

Freire, L., Roche, A., Mangin, J.F., 2002. What is the best similarity measure for motion correction in fMRI time series? IEEE Trans. Med. Imaging 21 (5), 470–484.

Friston, K.J., Jezzard, P., Turner, R., 1994. The analysis of functional MRI time-series. Hum. Brain Mapp. 1, 153–171.

Friston, K.J., Ashburner, J., Frith, C.D., Poline, J.-B., Heather, J.D., 1995. Spatial normalization and registration of images. Hum. Brain Mapp. 3, 165–189.

Friston, K.J., Williams, S., Howard, R., Frackowiak, R.S.J., Turner, R., 1996. Movement-related effects in fMRI time series. Magn. Reson. Med. 35, 346–355.

Garavan, H., Ross, T.J., Stein, E.A., 1999. Right hemispheric dominance of inhibitory control: an event-related functional MRI study. Proc. Natl. Acad. Sci. 96, 8301–8306.

Gavrilescu, M., Stuart, G.W., Waites, A., Jackson, G., Svalbe, I.D., Egan, G.F., 2004. Changes in effective connectivity models in the presence of task-correlated motion: an fMRI study. Hum. Brain Mapp. 21 (2), 49–63.

Gold, S., Christian, B., Arndt, S., Zeien, G., Cizadlo, T., Johnson, D.L., Flaum, M., Andreasen, N.C., 1998. Functional MRI statistical software packages: a comparative analysis. Hum. Brain Mapp. 6, 73–84.

Hajnal, J.V., Myers, R., Oatridge, A., Schwieso, J.E., Young, I.R., Bydder, G.M., 1994. Artifacts due to stimulus correlated motion in functional imaging of the brain. Magn. Reson. Med. 31, 283–291.

Hellier, P., Barillot, C., Corouge, I., Gibaud, B., Le Goualher, G., Collins, D.L., Evans, A., Malandaln, G., Ayache, N., Christensen, G.E., Johnson, H.J., 2003. Retrospective evaluation of intersubject brain registration. IEEE Trans. Med. Imaging 22 (9), 1120–1130.

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., Turner, R., 2002. Image distortion correction in fMRI: a quantitative evaluation. NeuroImage 16, 217–240.

Jenkinson, M., Smith, S., 2001. The role of registration in functional magnetic resonance imaging. In: Hajnal, J.V., Hill, D.L.G., Hawkes, D.J. (Eds.), Medical Image Registration. CRC Press, New York, pp. 183–198.

Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17 (2), 825–841.

Jezzard, P., Balaban, R.S., 1995. Correction for geometric distortions in echoplanar images from B_0 field variations. Magn. Reson. Med. 34, 65–73.

Jiang, A., Kenedy, D.N., Baker, J.R., Weisskoff, R.M., Tootell, R.B.H., Woods, R.P., Benson, R.R., Kwong, K.K., Brady, T.J., Rosen, B.R., 1995. Motion detection and correction in functional MRI imaging. Hum. Brain Mapp. 3, 224–235.

Kim, B., Boes, J.L., Bland, P.H., Chenevert, T.L., Meyer, C.R., 1999. Motion correction in fMRI via registration of individual slices into an anatomical volume. Magn. Reson. Med. 41 (5), 964–972.

Koole, M., D'Asseler, Y., Van Laere, K., Van de Walle, R., Van de Wiele, C., Lemahieu, I., Dierckx, R.A., 1999. MRI-SPET and SPET-SPET brain coregistration: evaluation of the performance of eight different algorithms. Nucl. Med. Commun. 20, 659–669.

Liddle, P.F., Kiehl, K.A., Smith, A.M., 2001. Event-related fMRI study of response inhibition. Hum. Brain Mapp. 12, 100–109.

Morgan, V.L., Pickens, D.R., Hartmann, Price, R.R., 2001. Comparison of functional MRI image realignment tools using a computer-generated phantom. Magn. Reson. Med. 46, 510–514.

Ostuni, J.L., Santha, A.K.S., Mattay, V.S., Weinberger, D.R., Levin, R.L., Frank, J.A., 1997. Analysis of interpolation effects in the reslicing of functional MR images. J. Comput. Assist. Tomogr. 21 (5), 803–810.

Paeth, A.W., 1986. A fast algorithm for general raster rotation. Proc. Graphics Interface '86. Canadian Information Processing Society, Vancouver, Canada, pp. 77–81.

Seto, E., Sela, G., McIlroy, W.E., Black, S.E., Staines, W.R., Bronskill, M.J., McIntosh, A.R., Graham, S.J., 2001. Quantifying head motion associated with motor tasks used in fMRI. NeuroImage 14 (2), 284–297.

Singh, M., Al-Dayeh, L., Patel, P., Kim, T., 1998. Correction for head movements in multi-slice EPI functional MRI. IEEE Trans. Nucl. Sci. 45, 2162–2167.

Smith, E.E., Jonides, J., Koeppe, R.A., 1996. Dissociating verbal and spatial working memory using PET. Cereb. Cortex 6, 11–20.

Smith, S., Bannister, P., Beckmann, C., Brady, M., Clare, S., Flitney, D., Hansen, P., Jenkinson, M., Leibovici, D., Ripley, B., Woolrich, M., Zhang, Y., 2001. FSL: new tools for functional and structural brain image analysis. Seventh Int. Conf. on Functional Mapping of the Human Brain.

Strother, S.C., Anderson, J.R., Xu, X.L., Liow, J.S., Bonar, D.C., Rottenberg, D.A., 1994. Quantitative comparisons of image registration techniques based on high-resolution MRI of the brain. J. Comput. Assist. Tomogr. 18 (6), 954–962.

Strupp, J.P., 1996. Stimulate: a GUI based fMRI analysis software package. NeuroImage 3, S607.

Studholme, C., Hawkes, D.J., Hill, D.L.G., 1997. Robust fully automated 3-

D registration of MRI and PET images of the brain using multi-resolution voxel similarity measures. Med. Phys. 24 (1), 25–35.

Thacker, N.A., Burton, E., Lacey, A.J., Jackson, A., 1999. The effects of motion on parametric fMRI analysis techniques. Physiol. Meas. 20, 251–263.

West, J., Fitzpatrick, M., Wang, M.Y., Dawant, B.D., Maurer, C.R., Kessler, R.M., Maciunas, R.J., Barillot, C., Lemoine, D., Collignon, A., Maes, F., Suetens, P., Vandermeulen, D., van den Elsen, P.A., Napel, S., Sumanaweera, T.S., Harkness, B., Hemler, P.F., Hill, D.L.G., Hakes, D.J., Studholme, C., Maintz, J.B.A., Viergever, M.A., Malandain, G., Pennec, X., Noz, M.E., Maguire, G.Q., Pollack, M., Pelizzari, C.A., Robb, R.A., Hanson, D., Woods, R.P., 1997. Comparison and evaluation of retrospective intermodality brain image registration techniques. J. Comput. Assist. Tomogr. 21 (4), 554–566.

Woods, R.P., Cherry, S.R., Mazziotta, J.C., 1992. Rapid automated algorithm for aligning and reslicing PET images. J. Comput. Assist. Tomogr. 16 (4), 620–633.

Woods, R.P., Mazziotta, J.C., Cherry, S.R., 1993. MRI-PET human brain mapping registration with automated algorithm. J. Comput. Assist. Tomogr. 17 (4), 536–546.

Woods, R.P., Grafton, S.T., Holmes, C.J., Cherry, S.R., 1998a. Automated image registration: I. General methods and intrasubject, intramodality validation. J. Comput. Assist. Tomogr. 22 (1), 139–152.

Woods, R.P., Grafton, S.T., Watson, J.D.G., Sicotte, N.L., Mazziotta, J.C., 1998b. Automated image registration: II. Intersubject validation of linear and nonlinear models. J. Comput. Assist. Tomogr. 22 (1), 153–165.

Worsley, K.J., Liao, C., Aston, J., Petre, V., Duncan, G.H., Morales, F., Evans, A.C., 2002. A general statistical analysis for fMRI data. NeuroImage 15, 1–15.

Wu, D.H., Lewin, J.S., Duerk, J.L., 1997. Inadequacy of motion correction algorithms in functional MRI: role of susceptibility-induced artifacts. J. Magn. Res. Image 7, 365–370.

Zarahn, E., Aguirre, G.K., D’Esposito, M., 1997. Empirical analyses of BOLD fMRI statistics: I. Spatially unsmoothed data collected under null-hypothesis conditions. NeuroImage 5, 179–197.