

# What is the Best Similarity Measure for Motion Correction in fMRI Time Series?

L. Freire\*, A. Roche, and J.-F. Mangin

**Abstract**—It has been shown that the difference of squares cost function used by standard realignment packages (SPM and AIR) can lead to the detection of spurious activations, because the motion parameter estimations are biased by the activated areas. Therefore, this paper describes several experiments aiming at selecting a better similarity measure to drive functional magnetic resonance image registration. The behaviors of the Geman–McClure (GM) estimator, of the correlation ratio, and of the mutual information (MI) relative to activated areas are studied using simulated time series and actual data stemming from a 3T magnet. It is shown that these methods are more robust than the usual difference of squares measure. The results suggest also that the measures built from robust metrics like the GM estimator may be the best choice, while MI is also an interesting solution. Some more work, however, is required to compare the various robust metrics proposed in the literature.

**Index Terms**—Artifact, fMRI, motion correction, robust registration, spurious activation.

## I. INTRODUCTION

MOTION correction in functional magnetic resonance imaging (fMRI) time series is usually performed through the retrospective estimation of the subject's motion during the experiment. This motion is often modeled by a time series of three-dimensional (3-D) rigid body transformations, each transformation aligning one volume of the time series with the reference volume. The retrospective approach amounts then to the estimation of these transformations via the maximization of a similarity measure. A number of different similarity measures have been proposed in the literature in order to perform retrospective registration of 3-D data sets. Hence, a usual issue for the design of a registration procedure is the choice of the best similarity measure considering the characteristics of the problem being addressed [1]. This choice, indeed, may highly influence the procedure robustness and accuracy. This paper is dedicated to various experiments aiming at selecting the best similarity measure for motion correction in fMRI time series.

Realignment of fMRI time-series is today considered as a required preprocessing step before analysis of functional activa-

tion studies. When the subject movement is correlated with the task, indeed, the changes in signal intensity, which arise from head motion, can be confused with signal changes due to brain activity [2]. Nevertheless, standard realignment procedures are often not sufficient to correct for all signal changes due to motion. For instance, a nonideal interpolation scheme used to resample realigned images leads to motion-correlated residual intensity errors [3]. Other motion-correlated residuals may stem from “the spin history effect,” which occurs when the spin excitation schedule is changed by the subject motion [4], [5]. Finally, other motion-related artifacts can confound fMRI time series, such as intrascan motion and the interaction between motion and susceptibility artifacts [6], [7].

It has been reported that a number of residual motion-related artifacts after realignment are reduced by covarying out signal correlated with functions of the motion estimates [3], [4]. It has to be noted, however, that when the motion estimates are highly correlated with the task, this regression-based approach is bound to erase some actual activations. While this cost may appear as the price to pay in order to obtain a good protection against false positives, using this approach raises the issue of the motion estimate reliability. Indeed, if ever signal changes induced by the cognitive task slightly bias motion estimates in a systematic task-correlated way, the price of this correction may be very high. Without the correction, however, realignment from task-correlated motion estimates could induce some spurious activations. Hence, task-correlated motion estimates would be the worst artifact that can be imagined for a realignment method.

In a recent paper [8], it has been shown that this task-correlated motion estimate's artifact can occur with the realignment methods using the simple similarity measure made up by the sum of the squared differences between both images. This least squares measure, indeed, does not take into account potential outlier voxels related to functional activations. This observation was considered especially alarming because this least squares measure is underlying the two standard motion correction packages used by the brain mapping community (SPM [9] and AIR [10], [11]). Hence, we have decided to test a set of other similarity measures supposed to be more robust relative to outliers.

It has to be understood that the task-correlated bias problem, whose amplitude is usually small relative to voxel size [8], is not really related to the accuracy of the motion estimates. Reaching a high subvoxel accuracy for the motion estimates in actual time series, indeed, requires better models of the motion induced signal changes. For instance, spatial distortions related to echo-planar imaging (EPI) depend on the subject position in the scanner, which may confound motion estimation [12], [13]. Furthermore, the motion correction problem is very different from

Manuscript received November 9, 2001; revised February 25, 2002. Asterisk indicates corresponding author.

\*L. Freire is with the Service Hospitalier Frédéric Joliot, CEA, 91401 Orsay, France, Instituto de Biofísica e Engenharia Biomédica, FCUL, 1749-016 Lisboa, Portugal, and the Instituto de Medicina Nuclear, FML, 1649-028 Lisboa, Portugal (e-mail: lmfreire@fc.ul.pt).

A. Roche is with the Epidaure project, INRIA, Sophia Antipolis, France and also with Medical Vision Laboratory, University of Oxford, OX2 7BZ Oxford, U.K.

J.-F. Mangin is with the Service Hospitalier Frédéric Joliot, CEA, 91401 Orsay, France.

Publisher Item Identifier S 0278-0062(02)05532-5.

standard registration problems because the rigid-body transformation to be estimated is small (time series including large motions are often discarded). Hence, the motion correction algorithm is not plagued by the similarity measure large attraction basins located far away from the basin that includes the global optimum. Consequently, most of the similarity measures usually give relatively “good estimates” of the motion parameters. Therefore, our experiments mainly focus on the potential task-correlated bias observed in motion estimates whatever the estimate actual accuracy.

fMRI usually relies on fast acquisition schemes like EPI that yields low spatial resolution images (usually 3 mm). Therefore, while feature-based similarity measures relying for instance on edge maps [14] may appear as good candidates to overcome outlier influence, their accuracy is uncertain. With this in mind, we have chosen to limit our study to intensity-based measures. Considering the huge number of different intensity-based similarity measures proposed in the literature [15], [16], we have decided to select only one measure among each different family of the taxonomy recently proposed by Roche [1]. This taxonomy is related to the kind of dependence between the image intensities supposed to be verified when both images are matched. The similarity measure is intended to quantify how well this dependence is verified given a transformation between the images. In each family, one of the more standard measures has been selected, but other related measures more adapted to motion correction in fMRI time series could certainly be derived in the future.

Seven different motion estimation procedures are compared throughout the paper, which are enumerated here according to an increasing number of degrees of freedom (DOF) for the assumed dependence between intensities (this order is also related to the measure computational complexity).

- *Intensity conservation*: two different implementations of the difference of squares measure (SPM [9] and AIR [10]) and the Geman–McClure (GM) robust estimator [17], which takes into account the existence of potential outliers;
- *Affine dependence*: the ratio image uniformity function [10];
- *Functional dependence*: two symmetrical implementations of the correlation ratio [18];
- *Statistical dependence*: the mutual information (MI) [19]–[23].

The behavior of each procedure is studied first relative to a motion free time series including simulated activations. A second experiment is dedicated to the influence of the amplitude and extent of the simulated activated areas on the accuracy of the estimations of simulated motions. Finally, the different procedures are confronted with an actual time series obtained from a 3T magnet.

## II. MATERIAL AND METHODS

### A. fMRI Acquisitions

All fMRI studies were performed on a Bruker scanner operating at 3T using a 30-contiguous-slice two-dimensional EPI sequence (slice array of  $64 \times 64$  voxels). This sequence had

in-plane resolution of 3.75 mm and slice thickness of 4 mm. The potential bias induced by activations in realignment algorithms was evaluated in a human study using a design of two alternating visual stimuli. The subject’s head was cushioned inside the Bruker proprietary head radio-frequency coil assembly and two adjustable pads exerted light pressure to either side of the head.

### B. Similarity Measures

To bring two images into spatial alignment, a rigid-body transformation is applied to one of the images. The purpose of a similarity measure is to return a value indicating how well two images match given a certain transformation. Ideally, by maximizing the similarity measure one should find the transformation that registers the images. The optimal rigid-body transformation, however, usually depends on the chosen similarity measure and on the implementation of its optimization. The goal of this paper is to assess to what extent the optimal transformations given by various realignment methods are biased by activations. Hence, we do not address the complex problem of choosing the best optimization scheme to obtain this optimal transformation [24].

Seven different realignment methods based on five different similarity measures are used in our experiments. Each underlying implementation depends on a few parameters, which may slightly modify the realignment results. A number of works have been dedicated to evaluation of registration methods accuracy [11], [25]–[28]. While this is clearly a key point to compare similarity measures, such work requires the study of each parameter influence, which is far beyond the scope of this paper. Since our main goal is to highlight the potential bias induced by activations, we have chosen to set each parameter either to the best choice leading to acceptable computation time, or to the value commonly used by standard users. Apart from the standard SPM and AIR packages, all the procedures rely on a custom implementation including a cubic spline-based interpolation method available on the World Wide Web (<http://bigwww.epfl.ch/algorithms.html>, [29], [30]) and a Powell like optimization method. Custom versions of LS and RIU measures optimizations have been used in a previous paper to discard some potential confound related to our experiment schemes [8]. These custom implementations have presented the same qualitative behavior as SPM and AIR implementations. In particular, it was important to verify that the fact that the simulated time series stem from a cubic spline interpolation while SPM and AIR use other interpolation schemes (truncated *sinc* and linear interpolation) did not result in different behaviors.

- *LS-SPM*: the standard realignment algorithm in SPM96 (<http://www.fil.ion.ucl.ac.uk/spm>, [9]). The underlying similarity measure is simply a least square, namely the sum of the squared discrepancies between both images. One specificity of SPM implementation is the use of a first order Taylor series approximation of the rigid-body transformation effects. While this choice allows rapid minimization of the measure iteratively using singular-value decomposition (SVD), it may explain some differences with other implementations of

the same similarity measure [15]. A 3-D smoothing is applied before realignment to assure a good behavior of the Taylor expansion. We set the Gaussian kernel full-width at half-maximum (FWHM) to 8 mm. The number of iterations was set to 16. The whole set of slices was included in the SVD. We have checked that the realignment algorithm in SPM99, while slightly different (points outside the head are removed from the SVD), presents the same qualitative behavior relative to the bias induced by activations.

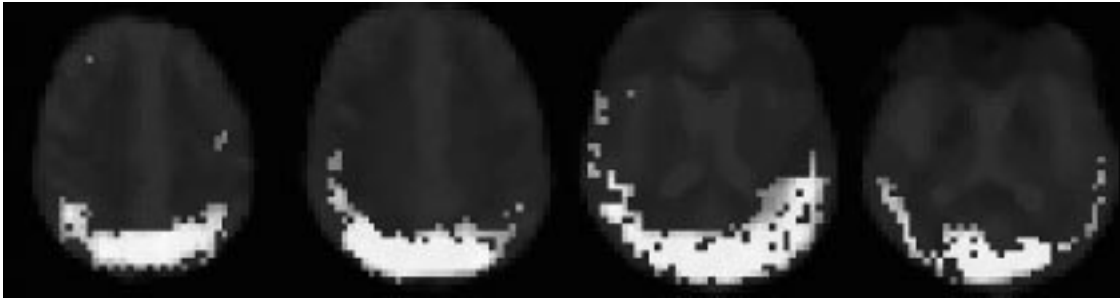
- *LS-AIR*: a second implementation of the least-square approach proposed in AIR 2.0 [10]. The minimization is done according to a Powell like unidimensional algorithm. A 3-D Gaussian smoothing is applied to the data to get more robust minimization (FWHM = 4 mm). A threshold was used to discard voxels with low signal (mainly located outside the head).
- *GM*: the GM robust estimator is an M-estimator supposed to reduce the influence of large residuals biasing the L2 metrics used by the difference of squares measure [17]. The contribution of each residual to the global measure is  $x^2/(1 + (x/C)^2)$ , where  $x$  denotes intensity difference and  $C$  is a scale parameter used to tune the cut power of this estimator. For comparison,  $C$  has been fixed to 0.2% of the mean brain intensity according to a preliminary experiment described in the following. Time-series realignment based on GM (and other M-estimators) have been implemented as a SPM toolbox called INRIAalign and is available on the World Wide Web at: <http://www-sop.inria.fr/epidaure/IRMF/INRIAalign.html>. This implementation is based upon a Newton scheme similar to the original SPM realignment technique. However, for the experiments, we have used the custom framework subjacent to CR, Crsym, and MI.
- *RIU-AIR*: the ratio image uniformity similarity function of AIR 2.0 [10]. This function is simply the standard deviation of a ratio image computed on a voxel-by-voxel basis. Minimization of this cost function increases the uniformity of the ratio image, which is independent of global scaling of the original images and improves registration. Preprocessing and minimization are performed like for the previous LS-AIR measure. The measures used in AIR 3.0 are the same but the minimization implementation has been refined.
- *CR*: the correlation ratio is an asymmetrical similarity measure assuming a functional dependence between both image intensities [18]. This measure, which has some relationship with the inter-modality measure proposed in [31], requires that the reference image be partitioned into a number of intensity isosets, namely broken up into areas of similar intensity. These areas are placed over the transformed image. Then the variance within each area is calculated and the similarity measure is defined from a weighted sum of the variances. Our local implementation relies on a partition in 64 isosets corresponding to intensity ranges with equal length. The correlation ratio assumes that each intensity isoset in the reference image should correspond to a low dispersion intensity range in

the image to be aligned. More sophisticated versions of the correlation ratio principle, which have not been tested in this paper, may include robust estimators to reduce outlier influence [1].

- *CRsym*: this similarity measure is the same as the previous one, but the roles of the two volumes are swapped. Hence, the iso-intensity-based partition is applied on the volume of the time series to be aligned.
- *MI*: mutual information [19]–[23]. MI is a measure originating from information theory, which assumes the least about intensity dependence. The underlying concept is entropy. The entropy of an image can be thought of as a measure of dispersion in the distribution of the image gray values. Given two images  $A$  and  $B$ , the definition of the mutual information  $MI(A, B)$  of these images is:  $MI(A, B) = E(A) + E(B) - E(A, B)$  with  $E(A)$  and  $E(B)$  the entropies of the images  $A$  and  $B$ , respectively, and  $E(A, B)$  their joint entropy. The joint entropy  $E(A, B)$  measures the dispersion of the joint probability distribution  $p(a, b)$ : the probability of the occurrence of gray value  $a$  in image  $A$  and gray value  $b$  in image  $B$  (at the same position), for all  $a$  and  $b$  in the overlapping part of  $A$  and  $B$ . The joint probability distribution should have fewer and sharper peaks when the images are matched than for any case of misalignment. Therefore, maximization of MI should correspond to the best registration. An implementation of MI can be found in SPM99. For the experiments described in this paper, however, our own implementation was used. The joint histogram computation includes a rebinning of each image gray level set to 64 values, where each voxel gray level contributes to the two closest values proportionally to the two underlying intervals.

### C. Simulations

Evaluation of the putative biasing effect due to activations was first achieved using artificial time-series. Each volume in the time series was created applying an artificial rigid-body motion  $T_{sim}$  to a reference image using a cubic spline-based interpolation method, available on the World Wide Web (<http://bigwww.epfl.ch/algorithms.html>, [29], [30]). This method embeds the volume in a surrounding space filled with null value. The reference image ( $64 \times 64 \times 30$ ,  $3.75 \times 3.75 \times 4$  mm) was one of the EPI BOLD image of the study mentioned above denoised with a standard  $3 \times 3 \times 3$  median filter. Gaussian noise was added to the reference image and to all frames of the time series in order to simulate the effects of thermal noise in fMRI scans (standard deviation: 2.5% of mean cerebral voxel value). Various artificial activations were then added either to the reference image or to the rest of the time series according to the simulation requirement. Three different activation patterns were manually drawn in the occipital lobe in order to mimic some visual activations observed during the underlying neuroscience study. These patterns were first filled with a random noise, then spatially filtered with a Gaussian (standard deviation: 2 mm). The resulting image was then masked according to the initial pattern. Some features of the final patterns are summarized in Fig. 1. A few slices presented



Pattern	A1	A2	A3
Size (%)	12.4	6.2	3.2
Mean (%)	1.26	1.19	1.18
Max (%)	2.04	2.03	2.03

Fig. 1. Summary of activation pattern features. Size is given as a percentage of the total number of brain voxels. Mean and Max denote mean and maximum signal increase for the activated voxels.

in Fig. 1 give an idea on the activation spatial profile. A range of activation amplitudes was studied using multiplicative factors.

Each frame of the artificial time series is aligned to the reference image using one of the registration methods, which yields an estimated rigid-body transformation  $T_{\text{est}}$ . Hence, the alignment error is given by the residual rigid-body transformation  $T_{\text{res}} = T_{\text{sim}} \times T_{\text{est}}^{-1}$ , where each transformation is represented by a standard homogeneous matrix. The origin for rotations is located in the center of the volume. The translation ( $E_t$ ) and rotation ( $E_r$ ) alignment errors are given by  $E_t = \text{sqrt}(T(1,4)^2 + T(2,4)^2 + T(3,4)^2)$  (in mm) and  $E_r = \cos^{-1}[(T(1,1) + T(2,2) + T(3,3) - 1)/2]$  (in degree). When required, the six motion parameters of a transformation  $T$  are given by:  $t_x = T(1,4)$ ,  $t_y = T(2,4)$ ,  $t_z = T(3,4)$ ,  $r_y = \sin^{-1}(T(1,3))$ ,  $r_x = \sin^{-1}(T(2,3)/\cos(r_y))$ , and  $r_z = \sin^{-1}(T(1,2)/\cos(r_y))$ .

### III. EXPERIMENTS

#### A. Simulated Activations Without Motion

The first experiment investigates whether some realignment method may lead to spurious task-related motion estimates in the absence of any initial misalignment in the time series. It has been shown in a previous paper that this artifact does exist with difference of square measure (LS), with an amplitude sufficient to induce additional spurious activations along high-contrast edges after resampling of the time series [8]. The different steps of this experiment can be summarized as follows:

- Generate an artificial time-series by duplicating the reference image 40 times;
- Include in each frame the activation pattern A1 (see Fig. 1) multiplied by an intensity which varies throughout the time series according to the time course given in Fig. 2 (two square stimuli convolved with a Gaussian; the maximal mean activation is 2.52%);
- Run the seven registration methods;
- Evaluate the six transformation parameters of  $T_{\text{est}}$  for each package;
- Compute cross correlation between each parameter and A1 time course.

Since a previous paper has shown that the difference of square measure leads to task-correlated motion estimations [8], a preliminary experiment is done with the GM estimator with a range of different values for the scale parameter. This experiment aims at evaluating whether this robust measure can overcome the influence of the large residuals related to the simulated activations. The estimations of the motion parameters presenting the highest bias with LS measure (see Fig. 4) are presented in Fig. 2 for eight different values of  $C$  (given as a percentage of the mean brain intensity). The charts refer to  $T_y$  and  $R_x$  (pitch). As expected, the charts show that when the scale parameter is too high, the GM estimator leads to task-correlated estimations like LS, which is shown by the high correlation coefficient values between the parameters and A1 time course. When this scale parameter is sufficiently low, however, the correlation with the simulated activation profile almost disappears, which tends to prove the efficiency of the robust metrics. From this experiment, we decided to set the value of the scale parameter to 0.2% of mean brain value.

Our previous paper has also shown that the bias amplitude using LS- and RIU-based measures increases with the width of the Gaussian kernel used to spatially smooth the data before registration in order to reduce the number of local minima [8]. Therefore, a second preliminary experiment was performed to assess the effect of a preliminary smoothing on the behavior of the new measures studied in this paper (GM, CR, and CRsym). The results for two of the motion parameters are presented in Fig. 3. The correlation with the activation profile increases significantly with the smoothing kernel width for the two CR-based measures, while this effect does exist for GM but to a lesser extent. The amplitude of the bias, indeed, remains very low. An additional information inferred from this experiment for GM is related to the sensitivity to local minima observed when the alignment is applied to raw data. This observation led us to apply a 4-mm FWHM Gaussian smoothing to the data before using GM in all the following experiments. In return, no spatial smoothing is applied for CR and CRsym.

Finally, Fig. 4 is dedicated to a comparison of the six motion parameter estimations using the seven different methods. Several realignment parameters related to the least-square-based methods (LS-SPM and LS-AIR) demonstrate a high correlation

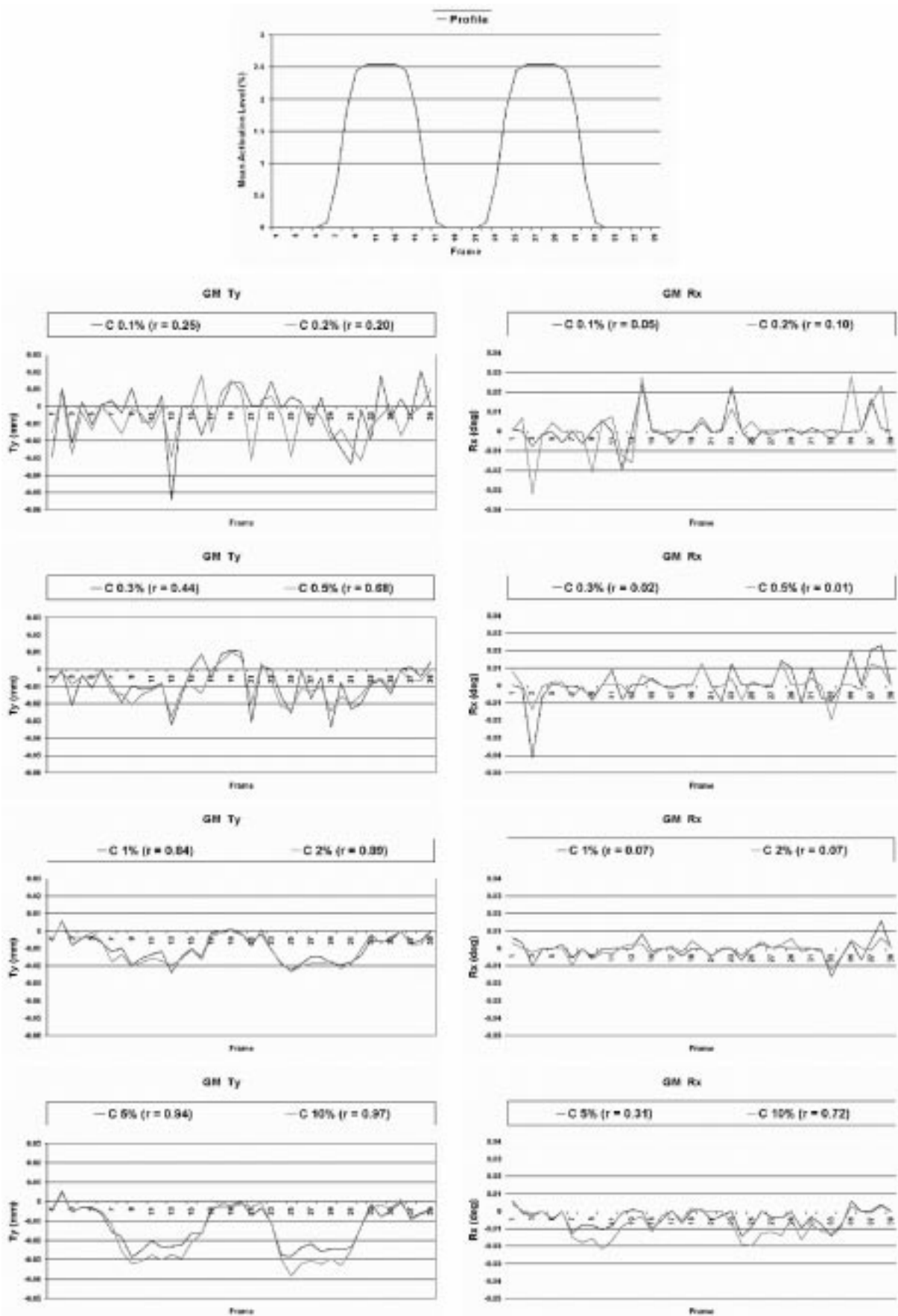


Fig. 2. Up: The activation profile used to create the motion free simulated time series including an activated area. Down: Study of  $C$  influence on GM estimator. A 4-mm FWHM smoothing Gaussian kernel is applied to the time series before motion parameter estimation to reduce local minima related problems (see Fig. 3).  $C$  is given as a percentage of mean brain value. The correlation with the activation profile is given for each chart.

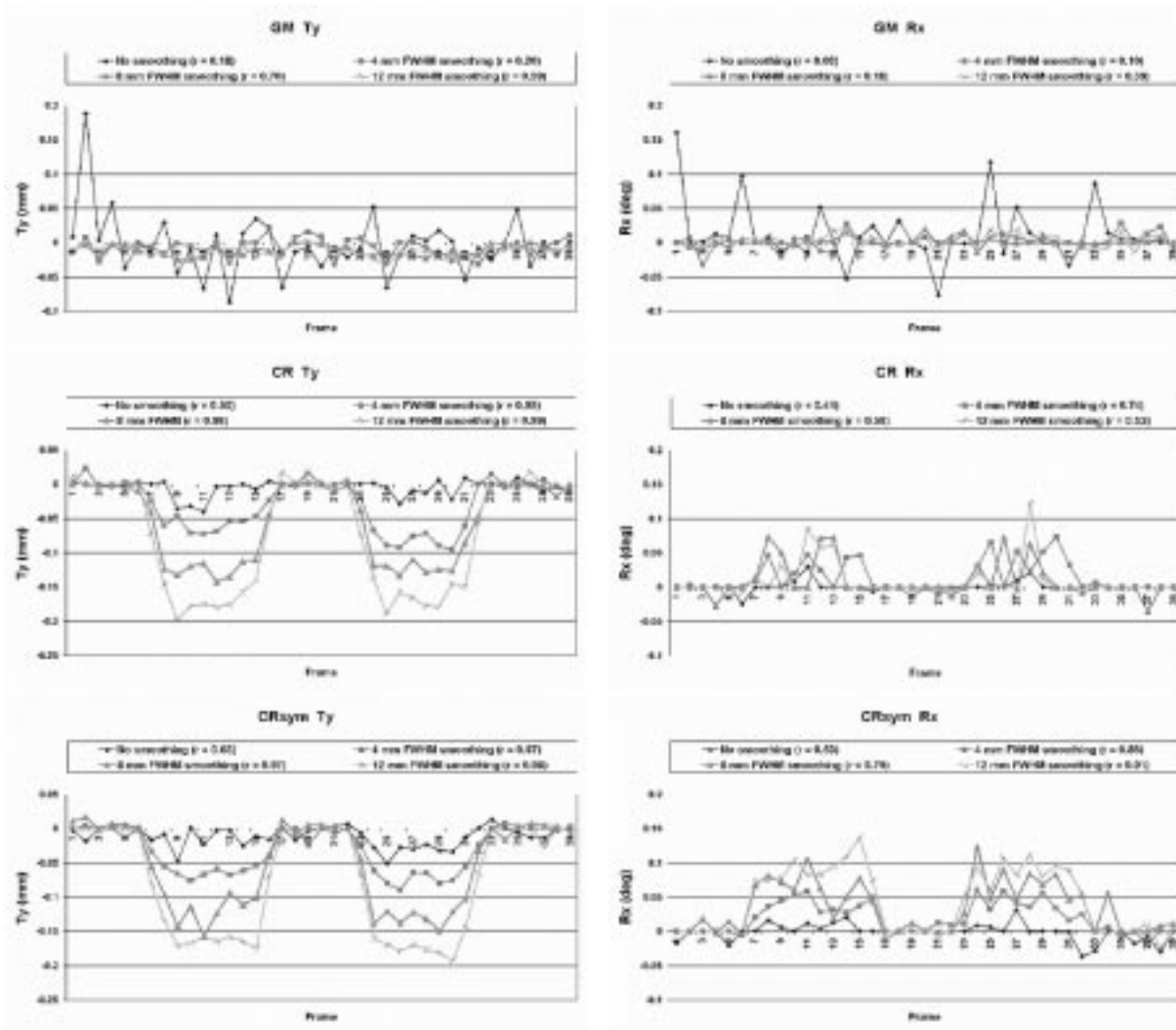


Fig. 3. Study of data spatial smoothing influence on GM (top), CR (middle) and CRsym (bottom) for two of the motion parameters. The correlation with the activation profile is given for each chart.

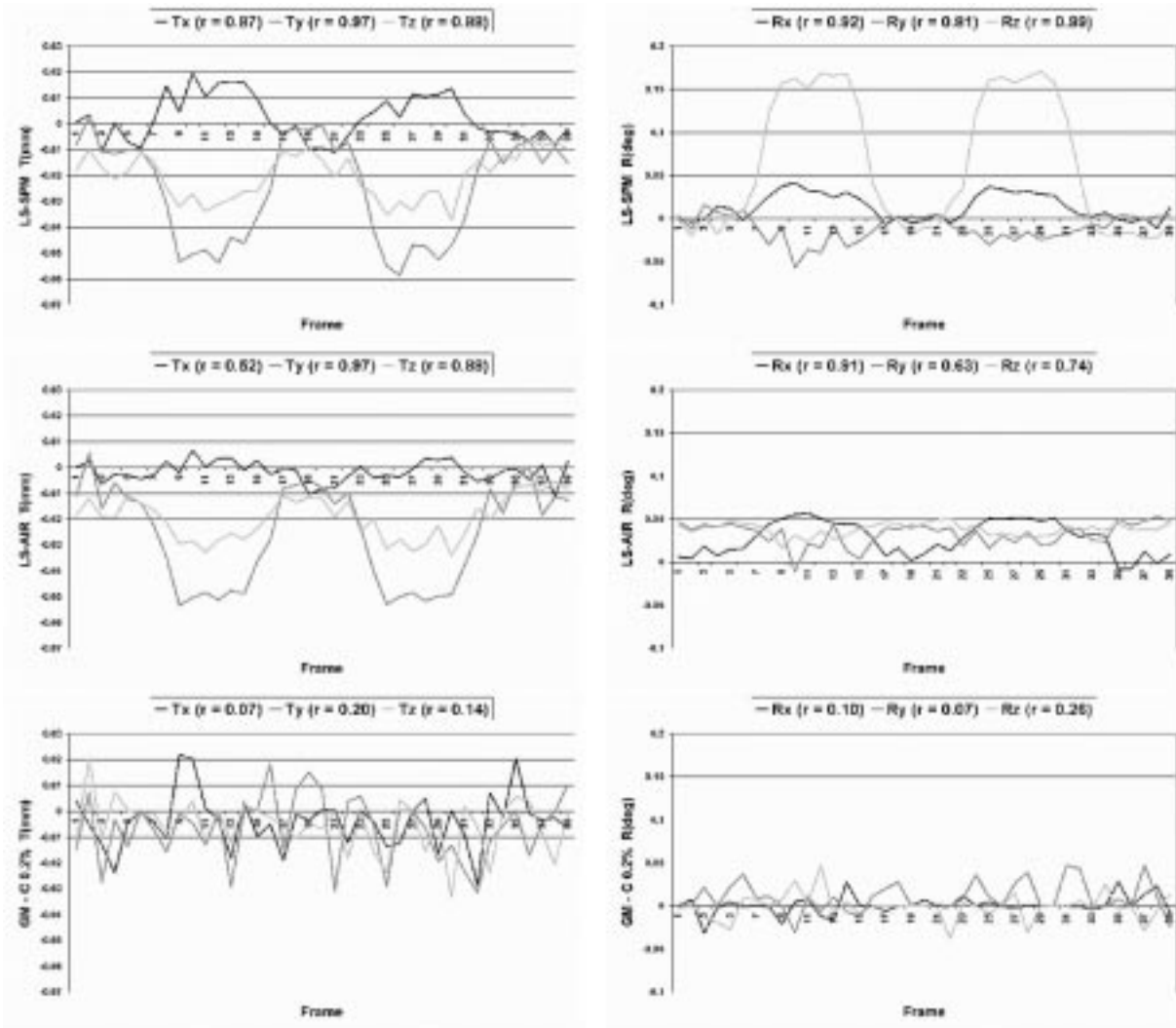
with the time course of the simulated activation (see Fig. 4(a): for LS-SPM, the highest correlation is obtained for the  $yaw$  parameter (0.99); for LS-AIR, the maximum correlation is obtained for the  $t_y$  parameter (0.97). The two LS-based methods give very different results for the  $rz$  parameter (LS-SPM being highly biased), which shows that the measure optimization scheme can highly influence the activation related artefact. The highest amplitude of the task-related parameter time course is 0.05 mm ( $t_y$ ) and 0.15 deg ( $yaw$ ) for LS-SPM, 0.05 mm ( $t_y$ ) and 0.04 deg ( $pitch$ ) for LS-AIR. An additional experiment with the same data has shown that this bias is largely decreased if no spatial smoothing is applied to the data, but remains significant although partly hidden by local minima [8].

Lower but significant correlations are observed for some parameters related to MI (0.67 for  $t_z$ ), CRsym (0.66 for  $t_y$ ), and CR (0.50 for  $t_y$ ). Bias amplitude remains important with CR-based measures but the charts are noisier than with LS because of local minima (no spatial smoothing was used). Bias amplitude is quite low with MI (0.01 mm for  $t_z$  and 0.02 deg for  $pitch$ ) and it has been shown elsewhere that no spurious activations were detected using SPM after realignment

[8]. Finally, correlation is quite low for GM (0.26 for  $yaw$ ) and RIU-AIR (0.10 for  $t_z$ ), but RIU-AIR seems especially plagued by local minima. It has been shown elsewhere that this problem was overcome by a larger spatial smoothing, leading unfortunately to a significant bias [8].

### B. Simulated Activations With Simulated Motion

The second experiment investigates the influence of simulated activations on registration method accuracy. A method robust to the presence of activations in the time series should keep the same level of accuracy whatever the activation features. The important point here is not the absolute accuracy of the method, which could depend on the tuning of some intrinsic parameters, but the potential accuracy modifications induced by signal changes in the activated areas. This experiment relies on a large number of simulated volumes, which allows us to study the influence of several parameters on a statistical basis. In order to get rid of the potential confound related to field of view variations after simulated motion, all volumes were stripped from their border voxels before realignment in order to reach a  $62 \times 62 \times 28$  geometry (a subvolume was used). To



(a)

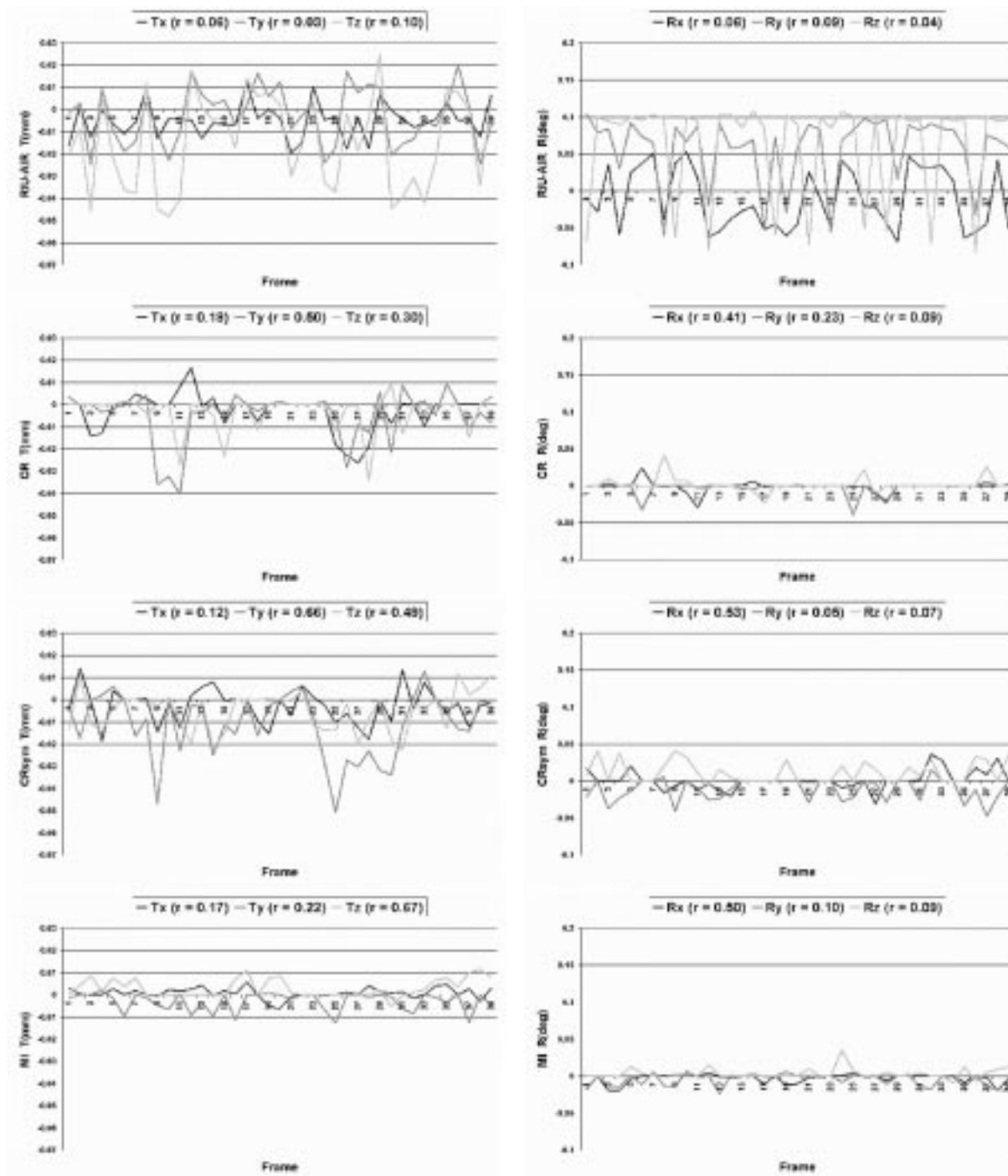
Fig. 4. The estimated motion parameters for (a) LS-SPM, LS-AIR, and GM. LS-based estimations are highly correlated with the simulated activation profile. The correlation with the activation profile is given for each chart.

eliminate simulated motion specificities relative to the reference volume axes as a potential confound, the simulated translations were applied systematically in the 20 directions of a regular dodecahedron and the simulated rotations were applied around the 20 different axes defined by the same dodecahedron. Hence, for a given translation or rotation amplitude, accuracy was assessed from means and standard deviations of translation ( $E_t$ ) and rotation ( $E_r$ ) errors relative to 20 different realignments. Simulated translation and rotation amplitudes have been chosen first small relative to voxel size, because our previous work has shown that with larger motions, the activation bias was hidden behind the method's decreased accuracy [8]. Hence, the first part of the following experiment is related to the realignment behavior when the subject is almost motionless. Nevertheless, we propose some results with larger amplitude simulated motions, in order to compare the variances of the different methods in a more difficult situation.

1) *Activation Level:* The influence of activation level was studied for 0.2 and 2.0 mm translations and  $0.2^\circ$  and  $2.0^\circ$  rotations. A1 pattern was added to reference image with 0.63%,

1.26%, 2.52%, 5.04%, and 10.08% mean signal increase and the results compared with the situation of no-activation. For 0.2-mm translation and  $0.2^\circ$  rotation [Fig. 5(a)], LS-SPM accuracy declines linearly relative to activation mean signal increase whatever the considered error ( $E_t$  and  $E_r$ ). A similar but smaller effect is observed for LS-AIR. The accuracy of the five other methods does not depend so dramatically on the signal increase amplitude in the activated area, although a slight increase may be noted for GM, CR, and CRsym when activation exceeds the noise level. For 2.0 mm translation and  $2.0^\circ$  rotation [Fig. 5(b)] it is clear that increased errors in motion estimate reduce or even completely hide the activation-induced bias effect, notably in all the metrics except MI, which remains unchanged and in LS-SPM, for which an induced-bias is still visible.

For the simulated translation of 2 mm, RIU-AIR translation errors are typically of the order of the translation value, which suggests significant local minima problems near the null translation. For commodity reasons, this chart is truncated to preserve visual discrimination for the other methods' values.



(b)

Fig. 4 (Continued.) The estimated motion parameters for (b) RIU-AIR, CR, CRsym, and MI. The correlation with the activation profile is given for each chart.

2) *Activation Size*: The influence of activation size was also studied for translations of 0.2 mm and rotations of 0.2° (Fig. 6). A1, A2, and A3 patterns were added separately to reference image (with 2.52% mean signal increase in activated regions) and the results displayed with the nonactivated situation for comparison. LS-SPM accuracy declines significantly when the activated area is enlarged, while this effect is smaller for LS-AIR methods and nonexistent for the other methods.

When a LS-based method is used, activation level has a more dramatic role on the accuracy decline than activation size, which

could be expected considering the squared cost of the outlier discrepancies. When activated region size is doubled, LS-SPM errors in rotation (the most modified ones) are typically increased by a 25%–50% factor. When activation level is doubled, however, LS-SPM errors in rotation may be doubled.

### C. Experiments With Actual Time Series

Finally, the seven registration methods were run on an actual time series made up of 180 3-D images acquired every 2 s (frames). GM was tested with two different values of

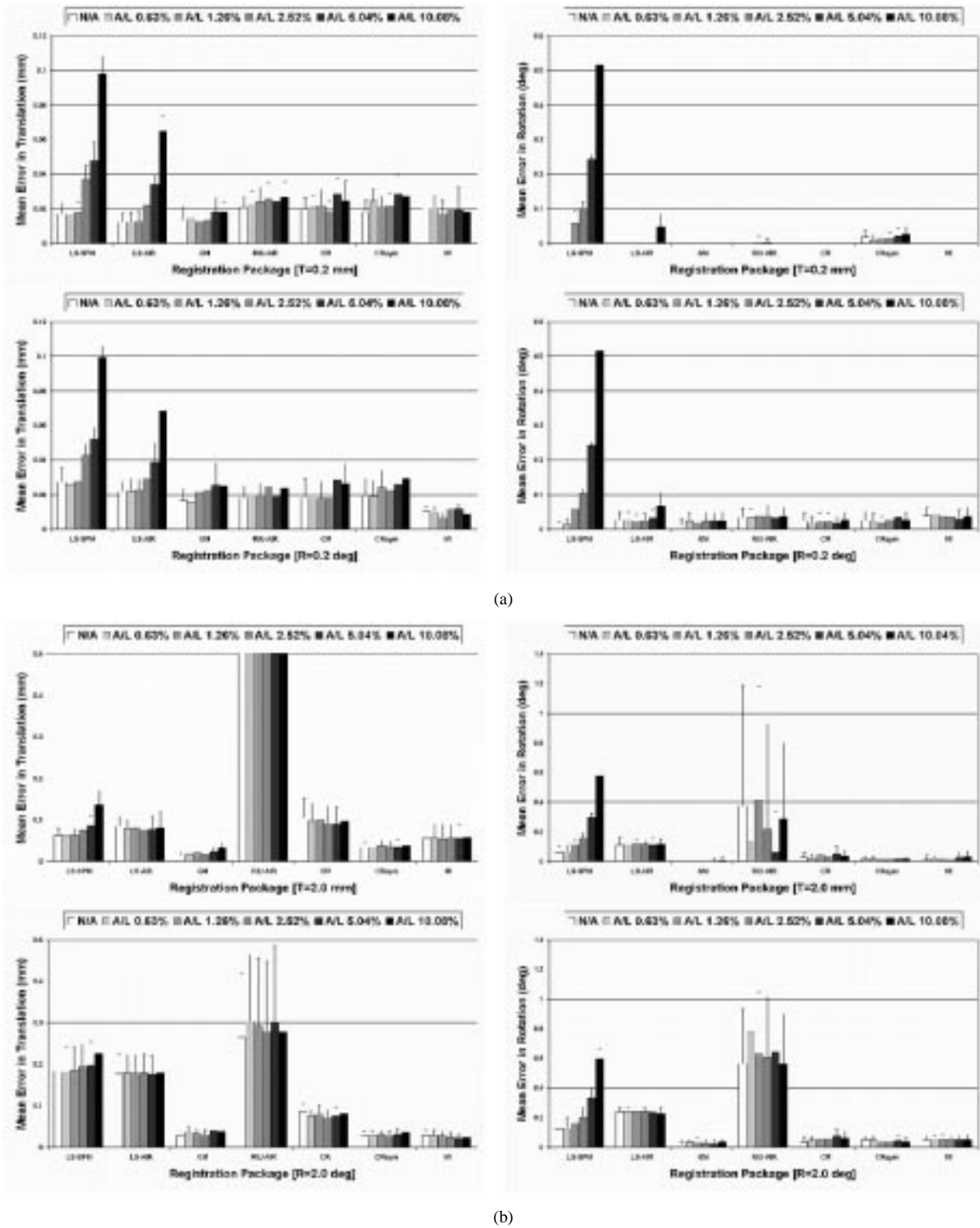


Fig. 5. Influence of activation level on registration accuracy. Accuracy of the different registration methods is evaluated by means of mean and standard deviation values of  $E_t$  (left) and  $E_r$  (right) for increasing activation intensities. 12.4% of the brain (occipital area, A1) is activated in the reference volume with mean signal increase (or activation level—A/L) ranging from 0.63% to 10.08% and results are displayed with no-activation situation (N/A) for comparison. Charts refer to: (a) simulated motions of 0.2 mm (top) and 0.2 deg (bottom); (b) simulated motions of 2.0 mm (top) and 2.0 deg (bottom).

$C$  (0.2% and 1% of mean brain value), because with the lowest value the method was especially prone to local minima. Furthermore, we initialized the optimization performed with these low scale factors with the result of an optimization performed with  $C = 100\%$  (which is equivalent to LS). More sophisticated adaptive approaches with a decreasing

scale factor during optimization are often used to deal with such robust metrics [1], [17].

The repeated stimulus period corresponds to 18 frames. Each period alternates two nine-frames-long presentations of two cognitively different visual stimuli. The six rigid-body registration parameters are displayed in Fig. 6 for the eight

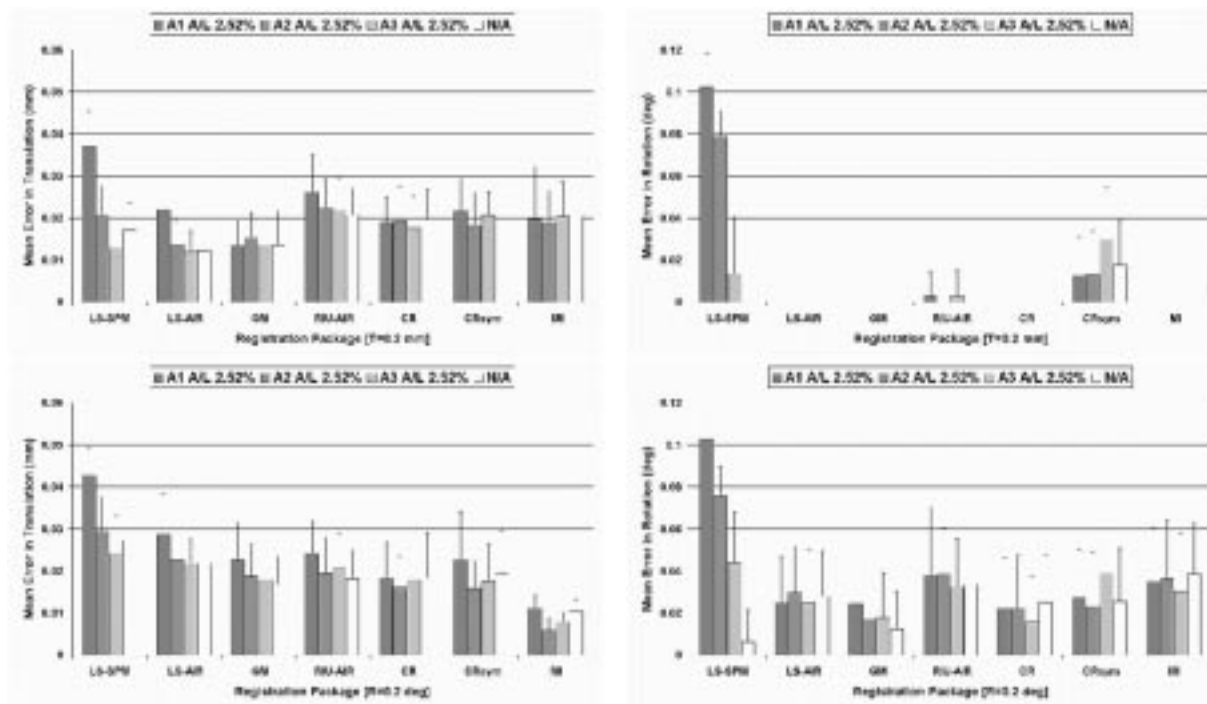


Fig. 6. Influence of activation size on registration accuracy: For each method, accuracy with decreasing activation size is displayed when 12.4%, 6.2%, and 3.2% of the brain (occipital area; A1, A2, and A3) are activated in the reference volume with mean activation level (A/L) of 2.52%. N/A refers to nonactivated situation. Charts produce means and standard deviations of  $E_t$  (left) and  $E_r$  (right) and refer to simulated motions of 0.2 mm (top) and 0.2 deg (bottom).

registration methods. The general trends of the six parameters' estimations are consistent across methods apart from the yaw parameter, which may be explained by some of the EPI related spatial distortions. It should be noted that according to the estimation results, the actual motion amplitude was rather small (less than  $0.15^\circ$  and 0.15 mm for all frames). Some of the charts clearly display stimulus correlated periodic variations (see Fig. 7). The more impressive periodic effect is observed on the pitch chart (Rx) for LS-SPM and LS-AIR, while this periodic trend is less clear for the other methods. The fact that this trend is not clearly observed for the majority of the methods discards, in our opinion, the existence of an actual quasiperiodic task correlated motion during the acquisition. Some less regular task correlated motions are nevertheless bound to exist to some extent, which has to be kept in mind during the result interpretation.

In order to assess correlation with the hemodynamic response, a moving average (one period width) was removed from each chart before computation of the cross correlation with the periodic stimulus convolved with the standard hemodynamic response of SPM99. For the pitch chart, which is the more affected in this study, the highest correlation is obtained for LS-SPM (0.79) and LS-AIR (0.64), while the other methods are also correlated: CRsym (0.54), CR (0.50), RIU-AIR (0.50), MI (0.46), GM-1 (0.45), and GM-0.2 (0.33). The amplitude of the putative bias, however, is small for the non LS methods. Hence, these correlations could stem either from a small bias, from a small actual motion, or, finally, from both.

To study the consequences of the putative bias on activation detection, the actual time series was realigned from each of the eight motion estimations using a cubic-spline interpolation.

SPM99 was used then to perform detection of activations. The following standard preprocessing was applied: spatial Gaussian smoothing (full-width at half maximum 5 mm), high-pass temporal filtering (period: 120 s) and low-pass temporal filtering by a Gaussian function with a 4-s width. The generalized linear model was used then to fit each voxel with a linear combination of two functions: the first one was derived by convolving a standard hemodynamic response function with the periodic stimulus, the second one was the time-derivative of the first one in order to model possible variations in activation onset. The voxels were reported as activated if the p-value exceeded a threshold of 0.05 corrected for multiple comparisons.

An illustration of the consequences of the stimulus-correlated motion estimates is shown for a few slices of the brain in Fig. 8. Considering the activation map obtained from the raw time series as a reference, several additional clusters of activated voxels are observed along some high-contrast brain edges after LS-SPM motion correction and to a smaller extent after LS-AIR correction. The largest of this additional cluster located in frontal lobe is also partly observed for all of the other methods. While the shape of this last cluster following the brain edge is very curious and corresponds to what may be expected for a spurious activation, the consistence of the eight results may indicate the presence of an actual activated area recovered by motion correction. The noisy activation map obtained after LS-SPM correction has been observed for numerous cognitive experiments performed with the 3T magnet in our institution.

#### IV. DISCUSSION

Our previous paper had shown that the difference of squares measure often used in the brain mapping community to per-

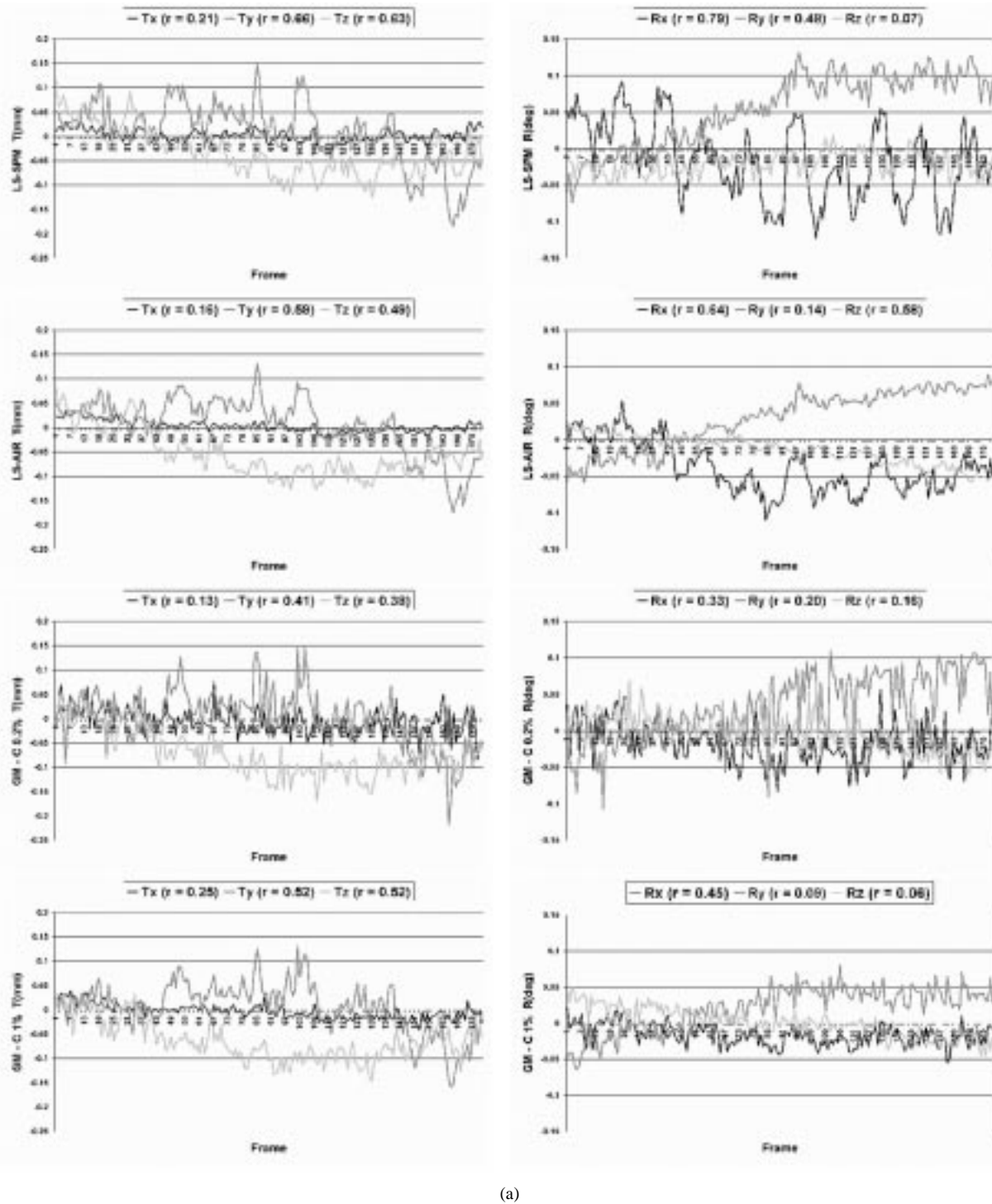
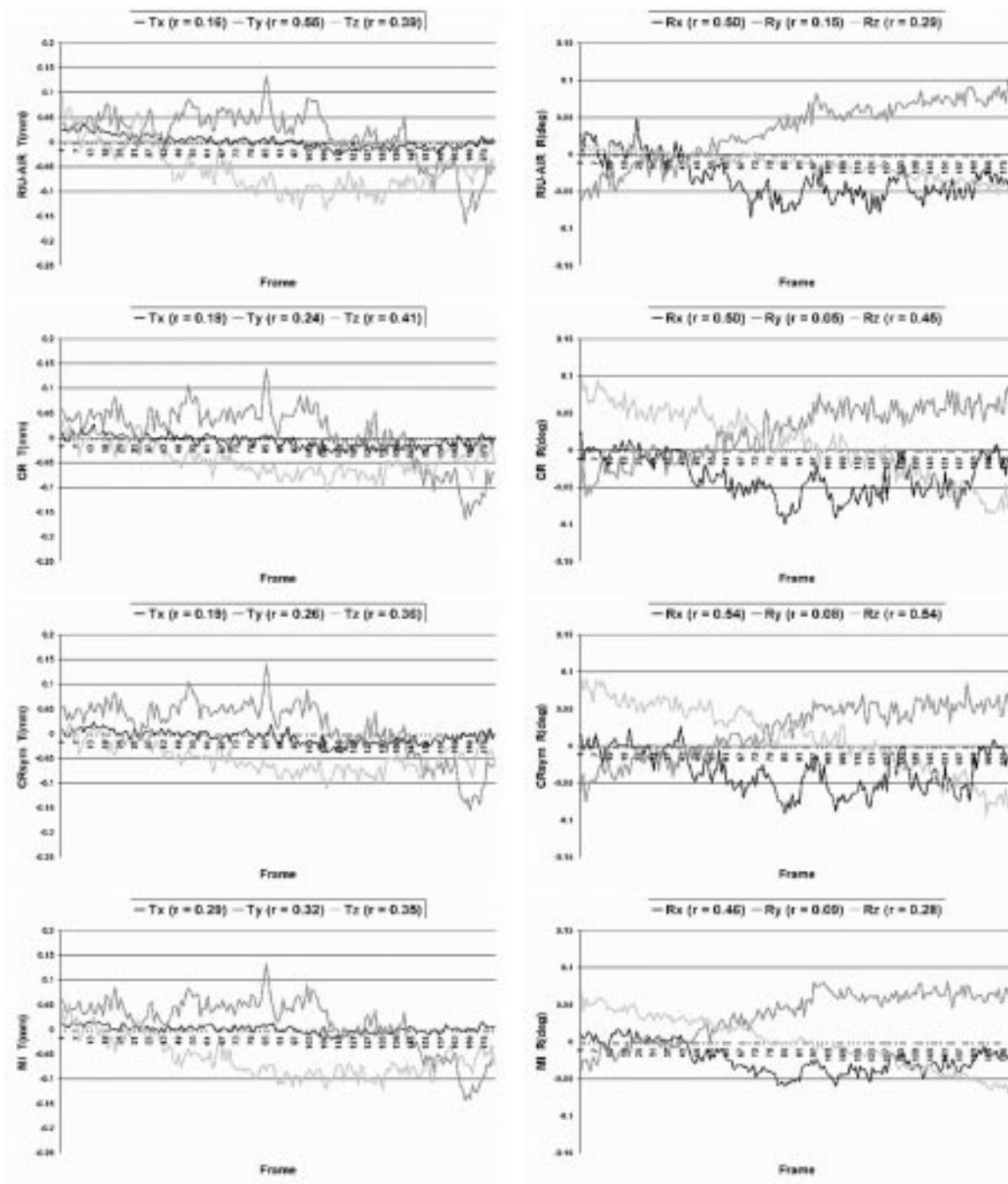


Fig. 7. Motion correction parameters for an actual time-series. (a) LS-SPM, LS-AIR, and GM with cutoff value  $C$  set to 0.2% and 1% of mean brain value.

form motion correction in fMRI time series could lead to some difficulties [8]. The standard motion correction packages (SPM and AIR), indeed, can yield task-correlated motion estimates resulting in spurious activations along some high-contrast edges (see Figs. 7 and 8). We had also highlighted the relationship between the bias amplitude and the amplitude of the signal change in activated areas (see Fig. 4), which allows us to forecast an increasing number of problems now that high-field magnets have appeared in a lot of institutions. Finally, we had studied an alternative measure to the sum of squared differences, namely MI,

which leads to motion estimates less prone to activation related bias. This new paper was aiming at evaluating whether MI was the best alternative.

It is interesting to note that most of our results could be predicted from the problem characteristics, namely the dependence that may be assumed between the intensities of two fMRI volumes of the same time series. Clearly, the assumptions underlying LS, RIU, and CR are contradicted by the signal changes in the activated areas, which explain the bias observed with LS and CR measures in some experiments.



(b)

Fig. 7 (Continued.) Motion correction parameters for an actual time-series. (b) RIU-AIR, CR, CRsym, and MI.

CR measure, however, is much more resistant to outliers than LS, which may result from the fact that each iso-set variance is biased in its own way or not biased at all, which leads to an average bias of lower amplitude. The results with RIU are more difficult to interpret because the bias may be hidden behind local minima problem, which has been shown in our previous paper [8].

Thanks to its larger number of DOF, MI never presents high-amplitude bias. In our experiments, this seem to provide a good protection against intensity inhomogeneities that may arise from

either bias field or partial volume effects, in the case of actual data, or resampling artifacts in the case of synthetic data. MI is likely to better accommodate for such inhomogeneities, as it amounts to a bin-by-bin intensity remapping. However, the choice of MI for this application may be questionable. This statistical measure, indeed, assumes very few constraints about the dependence between both image intensities. Unfortunately, too many DOF in the similarity measure or in the spatial transformation usually lead to registration methods plagued by local extrema [18], [33], [34].

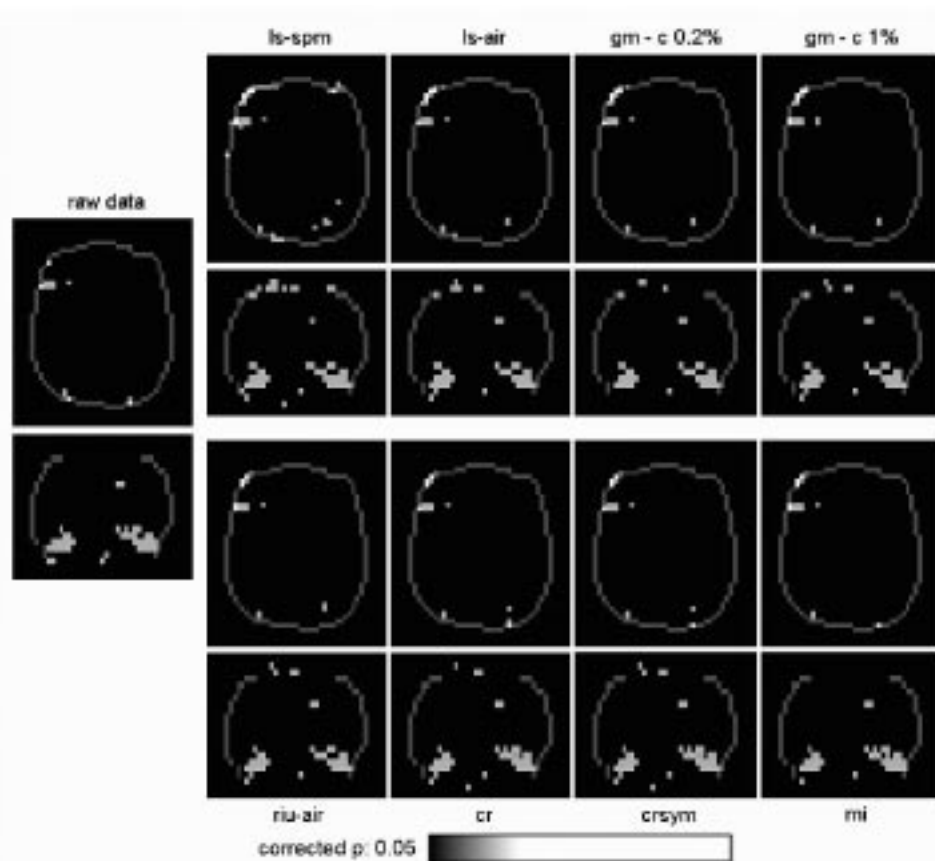


Fig. 8. A few slices of the activation maps obtained from SPM99 after realignment using the different methods.

Therefore, an important feature for the choice of a motion estimator is its variability. Our experiments with synthetic data suggest that GM and MI yield comparable standard deviations with respect to the mean estimation error (see Fig. 5), which could lead to the conclusion that the large number of DOF of MI is not a problem for this application. The experiment with 2-mm translations, however, shows a much lower variance of the translation error for the GM estimator than for MI. This effect may result from the interpolation artifact observed by Pluim in [34] and have then no impact on real data. Unfortunately, variability measurements were not provided in the case of real data since the true motion parameters were unknown. Assessing registration variability in the absence of ground truth is an especially difficult problem, even though some techniques to achieve this are currently emerging [32].

Anyway, some significant correlations with the task do occur with MI. This may be explained by the spatial smoothness of the hemodynamic response, which leads to a continuous range of activation amplitudes (note that no spatial smoothing was applied to data before MI-based registration). Hence, the activated area does not lead to some clear additional peaks in the joint histogram. It has to be noted that while a low amplitude bias may not lead to spurious activations, it prevents a trustworthy use of the motion parameters as regressors of noninterest during activation detection. Such regression, indeed, may remove some interesting activations. Therefore, we have investigated in this paper the behavior of some more constraining classic similarity measures.

Alternative measures relying on robust metrics can be easily devised for the three kinds of dependencies underlying LS, RIU and CR. In this paper, we have studied the behavior of the GM estimator, which is a robust similarity measure of the intensity conservation family. The same estimator could be used to derive a generalized correlation ratio [1], [35]. GM is only one possible popular choice among a lot of other robust estimators [36]. The key point inferred from our experiments is that the robust metrics approach seems to have the potential to overcome the activation induced bias. But some correlation remains during our experiment with actual data. This may be explained by the fact that, unlike MI, intensity inhomogeneities cause intensity conservation assumption to be violated, giving the idea that a robust metric like GM should be enhanced so as to include some spatially varying intensity correction. In our opinion, this could make GM to clearly outperform MI.

Some more work has to be done in order to choose the best robust metrics, which also requires, in our opinion, a better understanding of the features making a distinction between activation and noise. For instance, usual activation amplitudes observed in the visual system are in the range tested in this paper, while complex vascular effects can create even larger amplitudes up to 20% [37]. The interactions between EPI distortions, susceptibility artifacts and motion could also lead to some other kinds of outliers.

Our first experiment on the influence of the GM scale parameter is difficult to generalize to any activation study. The tuning of the cut power of the robust metrics, moreover, is

deeply related to the tuning of the spatial smoothing applied to the raw data before registration to reduce local minima problems. During the first experiments, the activation mean amplitude has been set to the same level as the standard deviation of the Gaussian noise added to mimic acquisition noise. Hence, the activated voxels cannot be distinguished by the amplitude of their associated residuals. After a spatial smoothing, however, the noise amplitude is decreased while the activation mean amplitude remains roughly the same because of its spatial coherence. Therefore, the activated voxels have higher residuals than the rest of the voxels, which induce a larger bias for the similarity measures based on the L2 metrics (LS and CR, Fig. 3). This interpretation means that while the bias induced by LS measure is largely reduced if no spatial smoothing is applied to the data, it is bound to come back with high-field magnets because of the improvement of the signal to noise ratio.

After a 4-mm-width spatial smoothing, nevertheless, the scale parameter of GM had to be decreased to a very low value in order to reach a low correlation with the activation profile (see Fig. 2). More than 70% of the residuals were larger than the “optimal” scale parameter that has been chosen for the following experiments. With actual data, this choice has resulted in additional local minima problems because too few voxels had some influence on the global measure. This last difficulty has been partly overcome with a two-stage strategy consisting in a first optimization with a high  $C$  GM almost equivalent to LS followed by a second optimization with a low  $C$  GM initialized by the result of the first one. Nevertheless, more sophisticated robust estimators performing an adaptive estimation of the cut power from the residual statistics could improve the results.

This also raises again the issue of motion estimates variability; it is known that when robustness against activation is done by systematically ignoring or down-weighting some of the data, this makes the similarity measure more prone to local minima. This is clearly visible in the experiment with actual data for the GM implementation with  $C$  set to 0.2 and 1% of mean brain value. Despite the same initialization with high  $C$ , increased cutoff power leads to smaller correlation with experimental paradigm, but motion estimates appear more “noisy.” On the other way, an increased number of DOF also contribute to robustness to outliers, but this generally leads to local minima problems. In our experiments with MI, however, this effect is not visible, which may be due to an increased ability of these methods to deal with intensity inhomogeneities, as mentioned above.

Throughout this paper, we have stuck to the idea that the solution to overcome the activation induced bias was relying in the choice of the similarity measure. A completely different strategy would consist of using the activation detection results to discard the activated voxels from the computation of the similarity measure. This strategy could afford the removal of some spurious activated voxels resulting from a first biased motion correction. In our opinion, however, the signal changes induced by the design of the more recent cognitive experiments, which usually include more complex stimulus timing than the standard block design, make such a strategy very difficult to follow. For instance, the simple notion of activated voxel is very difficult to define be-

cause a lot of different brain areas may be involved in the experiment with a different timing. Moreover, it would be very cumbersome to perform a new motion correction for each new improved statistical inference during data analysis. A better alternative would consist of filtering the motion parameter time series according to the cognitive experiment design. This third approach, however, cannot easily distinguish actual task-correlated motion from activation-induced bias.

## V. CONCLUSION

This paper has shown that the best solution to get rid of spurious effects induced by activated areas during motion correction seems to be the use of a robust metric to design the similarity measure driving the final motion parameter estimation. Regarding optimization issues, a simulated annealing strategy allowing a progressive decrease of the cut power (parameter  $C$ ) might be a good approach to avoid local minima problems. Also, the similarity measure could be enhanced so as to include a spatially varying intensity correction. We believe that this strategy could result in reducing the motion bias since intensity inhomogeneities between successive images would be better accommodated. The standard MI, however, remains an efficient simpler solution.

## ACKNOWLEDGMENT

The authors would like to thank J.-B. Poline, S. Berthoz, P. F. Van de Moortele, S. Dehaene, D. LeBihan, X. Pennec, and V. Frouin for the stimulating discussions about motion correction related problems.

## REFERENCES

- [1] A. Roche, “Recalage d’images médicales par inférence statistique,” Ph.D. dissertation, Université de Nice-Sophia Antipolis, Projet Epidauré, INRIA, Paris, France, 2001.
- [2] J. V. Hajnal, R. Myers, A. Oatridge, J. E. Schwieso, I. R. Young, and G. M. Bydder, “Artifacts due to stimulus correlated motion in functional imaging of the brain,” *Magn. Reson. Med.*, vol. 31, pp. 283–291, 1994.
- [3] S. Grootoink, C. Hutton, J. Ashburner, A. M. Howseman, O. Josephs, G. Rees, K. J. Friston, and R. Turner, “Characterization and correction of interpolation effects in the realignment of fMRI time series,” *NeuroImage*, vol. 11, pp. 49–57, Jan. 2000.
- [4] K. J. Friston, S. Williams, R. Howard, R. S. J. Frackowiak, and R. Turner, “Movement-related effects in fMRI time-series,” *Magn. Reson. Med.*, vol. 35, pp. 346–355, 1996.
- [5] M. D. Robson, J. C. Gatenby, A. W. Anderson, and J. C. Gore, “Practical considerations when correcting for movement-related effects present in fMRI time-series,” in *Proc. ISMRM 5th Annu. Meeting*, Vancouver, BC, Canada, 1997, p. 1681.
- [6] R. M. Birn, A. Jesmanowicz, R. W. Cox, and R. Shaker, “Correction of dynamic Bz-field artifacts in EPI,” in *Proc. ISMRM 5th Annu. Meeting*, 1997, p. 1913.
- [7] D. H. Wu, J. S. Lewin, and J. L. Duerk, “Inadequacy of motion correction algorithms in functional MRI: Role of susceptibility-induced artifacts,” *J. Magn. Res. Image*, vol. 7, pp. 365–370, 1997.
- [8] L. Freire and J.-F. Mangin, “Motion correction algorithms may create spurious brain activations in the absence of subject motion,” *NeuroImage*, vol. 14, pp. 709–722, Sept. 2001.
- [9] K. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. J. Frackowiak, “Spatial registration and normalization of images,” *Human Brain Mapping*, vol. 2, pp. 165–189, 1995.
- [10] R. P. Woods, S. R. Cherry, and J. C. Mazziotta, “Rapid automated algorithm for aligning and reslicing PET images,” *J. Comput. Assist. Tomogr.*, vol. 16, pp. 620–633, July/Aug. 1992.

- [11] R. P. Woods, S. T. Grafton, C. J. Holmes, S. R. Cherry, and J. C. Mazziotta, "Automated image registration: I. General methods and intrasubject, intramodality validation," *J. Comput. Assist. Tomogr.*, vol. 22, no. 1, pp. 139–152, 1998.
- [12] P. Jezzard and S. Clare, "Sources of distortion in functional MRI data," *Human Brain Mapping*, vol. 8, pp. 80–85, 1999.
- [13] J. L. R. Andersson, C. Hutton, J. Ashburner, R. Turner, and K. J. Friston, "Modeling geometric deformations in EPI time series," *NeuroImage*, vol. 13, pp. 903–919, May 2001.
- [14] J.-F. Mangin, V. Frouin, I. Bloch, B. Bendriem, and J. Lopez-Krahe, "Fast nonsupervised 3D registration of PET and MR images of the brain," *J. Cereb. Blood Flow Metab.*, vol. 14, no. 5, pp. 749–762, July 1994.
- [15] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, 1998.
- [16] D. L. G. Hill, P. G. Batchelor, M. Holden, and D. J. Hawkes, "Medical image registration," *Phys. Med. Biol.*, vol. 46, pp. R1–R45, 2001.
- [17] C. Nikou, F. Heitz, J.-P. Armspach, I.-J. Namer, and D. Grucker, "Registration of MR/MR and MR/SPECT brain images by fast stochastic optimization of robust voxel similarity measures," *NeuroImage*, vol. 8, pp. 30–43, July 1998.
- [18] A. Roche, G. Malandain, X. Pennec, and N. Ayache, "The correlation ratio as a new similarity measure for multimodal image registration," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer Verlag, 1998, vol. 1496, Proc. MICCAI'98, pp. 1115–1124.
- [19] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis, "Multi-modal volume registration by maximization of mutual information," *Med. Image Anal.*, vol. 1, no. 1, pp. 35–51, 1996.
- [20] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, pp. 187–198, Apr. 1997.
- [21] P. Viola and W. M. Wells, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.
- [22] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "Automated three-dimensional registration of magnetic resonance and positron emission tomography brain images by multiresolution optimization of voxel similarity measures," *Med. Phys.*, vol. 24, no. 1, pp. 25–35, January 1997.
- [23] C. R. Meyer, J. L. Boes, B. Kim, P. H. Bland, K. R. Zasadny, P. V. Kison, K. Koral, K. A. Frey, and R. L. Wahl, "Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations," *Med. Image Anal.*, vol. 1, no. 3, pp. 195–206, 1997.
- [24] M. Jenkinson and S. M. Smith, "A global optimization method for robust affine registration of brain images," *Med. Image Anal.*, vol. 5, no. 2, pp. 143–156, 2001.
- [25] A. Jiang, D. N. Kennedy, J. R. Baker, R. M. Weisskoff, R. B. H. Tootell, R. P. Woods, R. R. Benson, K. K. Kwong, T. J. Brady, B. R. Rosen, and J. W. Belliveau, "Motion detection and correction in functional MR imaging," *Human Brain Mapping*, vol. 3, pp. 224–235, 1995.
- [26] V. Frouin, E. Mességué, and J.-F. Mangin, "Assessment of two fMRI motion correction algorithms," *NeuroImage*, vol. 5, p. S458, May 1997.
- [27] J. West *et al.*, "Comparison and evaluation of retrospective intermodality brain image registration techniques," *J. Comput. Assist. Tomogr.*, vol. 21, no. 4, pp. 554–566, 1997.
- [28] M. Holden, D. L. G. Hill, E. R. E. Denton, J. M. Jarosz, T. C. S. Cox, T. Rohlfing, J. Goodey, and D. J. Hawkes, "Voxel similarity measures for 3-D serial MR brain image registration," *IEEE Trans. Med. Imag.*, vol. 19, pp. 94–102, Feb. 2000.
- [29] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Part I-theory," *IEEE Trans. Signal Processing*, vol. 41, pp. 821–833, Feb. 1993.
- [30] —, "B-spline signal processing: Part II-Efficient design and applications," *IEEE Trans. Signal Processing*, vol. 41, pp. 834–848, Feb. 1993.
- [31] R. P. Woods, J. C. Mazziotta, and S. R. Cherry, "MRI-PET registration with automated algorithm," *J. Comput. Assist. Tomogr.*, vol. 17, pp. 536–546, July/Aug. 1993.
- [32] X. Pennec, "Evaluation of the uncertainty in various registration problems," in *Methodology of Evaluation in Computational Medical Imaging*, K. Bowyer, M. Loew, H. Stiehl, and M. Viergever, Eds. Wadern, Germany: Schloss Dagstuhl Int. Conf. Res. Ctr. Comput. Sci., 2001, pp. 15–16.
- [33] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recogn.*, vol. 32, no. 1, pp. 71–86, 1999.
- [34] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Interpolation artefacts in mutual information-based image registration," *Comput. Vis. Image Understand*, vol. 77, no. 2, pp. 211–232, 2000.
- [35] A. Roche, X. Pennec, M. Rudolph, D. P. Auer, G. Malandain, S. Ourselin, L. M. Auer, and N. Ayache, "Generalized correlation ratio for rigid registration of 3D ultrasound with MR images," in *Lecture Notes in Computer Science*, vol. 1935, Proc. MICCAI'00. Berlin, Germany, 2000, pp. 567–577.
- [36] P. Meer, D. Mintz, D. Y. Kim, and A. Rosenfeld, "Robust regression methods in computer vision: A review," *Int. J. Comput. Vis.*, vol. 6, pp. 59–70, 1991.
- [37] R. Turner, P. Jezzard, H. Wen, K. K. Kwong, D. Le Bihan, T. Zeffiro, and R. S. Balaban, "Functional mapping of the human visual cortex at 4 and 1.5 Tesla using deoxygenation contrast EPI," *Magn. Reson. Med.*, vol. 29, pp. 277–279, Feb. 1993.